UNIVERSITY OF HERTFORDSHIRE

DOCTORAL THESIS

Structure and Information Parsimony:
Emergence of Symmetries from Generalised Information Bottlenecks

Author:
Hippolyte CHARVIN

Supervisors:
Daniel POLANI
Nicola CATENACCI VOLPI

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Adaptive Systems Research Group
School of Physics, Engineering & Computer Science

May 6, 2026

Abstract

The behaviour of embodied agents tends to unfold within a mesh of stringent constraints, each pulling in different directions. For instance, principled models of complexity-constrained behaviour suggest a fundamental tension between the enactement of the agent's purposeful behaviour and the "informational resources" at its disposal. The main motivation of this thesis is the hypothesis that this tension induces agents to leverage the *structure* of the interaction with their environment, at a "granularity" adapted to their informational resources — as relying on such structure limits the "informational expenditure" necessary to enact purposeful behaviour. From the agent's intrinsic perspective, this structure consists of regularities of the way its actions influence its sensory influx, i.e., of the *sensorimotor contingencies* (SMCs) that sensorimotor theories claim perception is based on — a concept that, despite the important progresses of the last decades, remains elusive in many respects.

Here, I develop novel mathematical and computational tools, at the intersection of information theory, group theory and dynamical systems, that provide a theoretical framework to explore the above hypothesis — so as to unlock new avenues for progress in adaptive behaviour, structure discovery and sensorimotor perception. In particular, previous research suggests to formalise SMCs through the *symmetries* defined by the dynamics of the sensorimotor interface — i.e., well-chosen *commutation relations* involving the way the agent's behaviour impacts its sensory influx. The above hypothesis thus becomes, at the formal level, that of a certain "duality" between a system's symmetries and the *informationally parsimonious* descriptions of it that such symmetries make possible. The main focus of this thesis is to establish and investigate this duality, in different forms — in a group-theoretic but also closed-loop and stochastic setting — and at different granularities — where the symmetries' granularity scales with the information parsimony that they make possible.

More precisely, the results cover three different themes. First, I look at symmetries of stochastic channels through the information-theoretic lens (in the finite case). Group-theoretic *invariances* of a stochastic channel are explicitly characterised in terms of information parsimony through the "classic" Information Bottleneck (IB) framework. Channel *equivariances*, however, cannot be captured by the classic IB, but require the introduction of a novel framework, which we call the Divergence Information Bottleneck (DIB). Here, information parsimony is traded-off against the preservation of the divergence of the data distribution from a given exponential family. For a well-chosen exponential family, the corresponding DIB formalises the intuition of an optimal compression preserving the "information carried by a given channel", and does characterise the channel's equivariances. These information-theoretic reformulations of "exact" channel symmetries then yield principled definitions of *soft* channel invariances and equivariances, where the "softness" of the symmetry is parametrised by the granularity of the corresponding coarse-graining.

Overall, this framework provides a novel building block for information theory-based symmetry discovery and symmetry-based coarse-graining. This formal progress could also be instrumental to understand the intrinsic structure of an agent's sensorimotor interface — i.e., what is known as *apparatus-related SMCs* — through the transformations of the agent's sensorimotor spaces that leave this interface unchanged.

While the latter symmetries describe structure at an exclusively sensorimotor level, previous work suggests that perception also relies on commutation relations between sensorimotor and *internal* dynamics. These results resonate with SMC theory's claim that the "*attunement*" of these internal dynamics makes them capture the invariants, but also the "*structure*

of changes" induced by a given, ongoing agent-environment interaction — i.e., what is known as *object-related SMCs*. I propose to explicitate these concepts by extending the *class-pose decomposition* framework. In this line of work, the aim is to decompose a group action into one coordinate (the "class") capturing the action's invariants, i.e., its orbits, and a second coordinate (the "pose") that is "strictly equivariant", in that it equivariantly tracks the changes induced by the group action without capturing any invariant.

This mathematical object is generalised in three directions: algebraic, dynamical and information-theoretic. The algebraic aspect starts from the observation that class-pose decomposition is only possible if all orbits are isomorphic — a highly non-generic assumption. To obtain a structure that can be built from any group action, we "reverse the arrows and break the bijectivity" in the commutation relations. This allows the transformations of the class-pose space to be *richer* than those of the original state-space. But we also require these commutation relations to be "maximally isomorphic", yielding what we call a *minimal joining* of the orbits: i.e., intuitively, the "simplest" group action that simultaneously "simulates" the original group action on each orbit. The dynamical aspect consists in moving from the setting of group actions — ultimately a poor model of embodied agents' own actions — to one that allows for a closed-loop "behaviour" made of non-invertible and stochastic actions: Markov Decision Processes (MDPs), here with fixed policy, no rewards and on standard Borel spaces. Classes become *ergodic components of the MDP*: i.e., intuitively, asymptotic attractors of the agent's behaviour. Poses are then defined by a minimal joining of these ergodic components, which can be seen as a dynamical, measure-theoretic and policy-dependent counterpart of least common multiples for integers. Eventually, these generalisations of classes and poses are reformulated (in the finite case) in terms of information parsimony. This yields, in particular, an information-theoretic characterisation of a group action's partition in orbits — an important step for formalising the links between symmetry and information.

These formal structures provide new tools to the operationalisation of sensorimotor perception. However, rather than yielding a pure explicitation of existing theories, they suggest new conceptual directions: e.g., that the attunement of brain dynamics is *induced by information parsimony constraints*, and yields *parsimonious fictions* which emerge from sensorimotor history but are not reducible to ongoing sensorimotor dynamics.

As most of the notions of structure above are defined for a continuous range of "granularities", parametrised by the corresponding informational trade-offs, it is crucial to understand the *relation* between such structures at different granularities. As a first step in this direction, I study the relations between coarser or finer bottlenecks for the classic IB method. More precisely, I investigate a property known as *successive refinement* (SR), which asks whether a coarser bottleneck can be obtained as a coarse-graining of a finer bottleneck. This property is important to determine whether for the soft invariances defined above, coarser invariances always include finer ones. It is also relevant to incremental learning, as it is equivalent to whether one can, *without incurring an additional informational cost*, design a finer bottleneck by first designing a coarser one, and then adding new information from the source.

SR is given (in the finite case and under mild assumptions) a geometric characterisation in terms of inclusion of convex hulls defined by bottleneck channels. This characterisation means, intuitively, that SR holds whenever the information captured by the coarser bottleneck is entirely "contained" in the information captured by the finer bottleneck. We then consider a *soft* notion of successive refinement, by quantifying the "lack" of it through a previously established notion of *unique information*. This allows to investigate, in synthetic numerical experiments, how the "amount to which" the SR property is satisfied depends on the respective granularities of the coarser and finer bottlenecks. The experiments suggest that, generically, the "lack" of successive refinability is relatively mild, and SR is the "closest" to hold for trade-off parameters poised close to bifurcation values.

Acknowledgements

First of all, I thank Daniel Polani, my main supervisor, who gave me the occasion to work on this topic and had a profound influence on both the substance and the style of my research. Perhaps most importantly, thanks for teaching me how to keep interpretation at the core of every step of the research process. Thanks to both Daniel and to Nicola Catenacci Volpi, my second supervisor, for the precise guidance, the honest and rich feedback, the deep confidence and the strong support all along.

I also want to warmly thank the whole Adaptive Systems Research Group (ASRG), which upholds a bold kind of research that has drastically broadened and sharpened my focus during these years — and, more importantly, has offered me an intellectual home. Thanks in particular to Christoph Salge, Faizan Rasheed, and Karen Archer. Actually, a massive thanks to Karen for such a stimulating presence, strength and embodiment of the meaning of mutual aid, including when first welcoming me to the UK. Thanks to Stavros Anagnou for our long discussions — scientific or otherwise; for his enthusiasm and inclusiveness, and for being a warm and socially proactive presence whenever he can. Beyond the ASRG at the University of Hertfordshire, I particularly thank Emil Dmitruk, Shabnam Kadir and Nathaniel Virgo for the scientific discussions that we shared.

Thanks to Nihat Ay for giving me the opportunity to present my work at the Institute for Data Science Foundations in Hamburg, and thanks to the whole group for warmly welcoming me — in particular Jesse van Oostrum, Adwait Datar, Frank Röder, Alexander Klemps, Pradeep Banerjee and Imke Christiane Bartscher.

Thanks to Fernando Rosas for our discussions at the workshop "Embracing Complexity: Principled and Practical Approaches to Emergence", and shortly after in London — on the relevance of computational mechanics to studying the structure of sensorimotor interfaces.

Beyond individuals, science usually is usually done within a given research community. Sometimes, when working on a topic that touches upon many different traditions while not being exactly encompassed by any, it actually does not feel quite this way. But sometimes it does: in this respect, my participations to the Guided Self-Organisation conferences were important moment of my PhD. I thank the organisers for giving me the opportunity to participate in it: in particular, the 2025 edition in Tübingen is surely the place where I had the highest concentration of stimulating discussions during my PhD.

Without indulging in a list of my intellectual debts to all the interesting people that I came across, I want to mention two particularly important episodes. First, thanks to Paulo Borges from the University of Lisbon, whose teaching of the philosophical content of various spiritual traditions set me on the path to try and understand reality through the lens of interdependence. This eventually led me to enactivism, and, through it, to the science that I practice today. Second, thanks to the Cardano group in Paris, who first showed me mathematical approaches to the study of living systems that I identified with. In particular, thanks to Alessandro Sarti and his Post-Structural Dynamics seminar at the École des Hautes Études en Sciences Sociales (EHESS), where the presentation of the Differential Heterogenesis framework strongly shaped my thinking and interests regarding the emergence of structure from dynamics.

Thanks to the Pazy Foundation, which funded conference travels and a work laptop for this thesis, under grant ID 195.

Eventually, a massive shout-out to the people that made my life what it has been in the past four years. Thanks, Unit G — a home, a place of friendship, and a place of care. Thanks to Sergiu for his energy, support and kindness in the last sprint when I needed it the most.

Merci à ma mère, merci à mon père, merci Ambroise — on reste ensemble, et ça me porte.

Contents

List of Abbreviations

BA	**B**lahut-**A**rimoto
CLP	**C**losed-**L**oop **P**erception
DIB	**D**ivergence **I**nformation **B**ottleneck
FEM	**F**ixational **E**ye **M**ovement
IB	**I**nformation **B**ottleneck
MDP	**M**arkov **D**ecision **P**rocess
RGC	**R**etinal **G**anglion **C**ell
SMC	**S**ensori**M**otor **C**ontingency
SR	**S**uccessive **R**efinement
UI	**U**nique **I**nformation
VS	**V**ertical **S**ystem

Chapter 0

How to navigate this thesis

The introductory Chapter 1 situates the work of the three core chapters (i.e., Chapters 2, 3 and 4) within the broader framework of adaptive behaviour, structure discovery and sensorimotor perception research. Chapters 2, 3 and 4 can still be read independently of each other and of Chapter 1 — except for some sections clarifying how they are relevant to each other, mostly in their introductions and conclusions. For the sections of Chapter 3 that root it in sensorimotor perception research, I recommend readers unfamiliar with the latter framework to first read Section 1.1.3. The concluding Chapter 5 assumes familiarity with all previous chapters.

Most mathematical proofs are gathered in the appendix. This allows for two levels of engagement with this thesis: while proofs are one of the core contributions, the main text can be read without diving into this level of detail. But readers interested in the proofs can easily navigate using the hyperlinks that I included: one from the main text to the appendix below each statement requiring a proof, and one from the appendix back to the main text at the beginning *and* end of each proof.

The main contributions are summarised in 2 pages in the Abstract, and in 10 pages in Section 1.3.

Chapter 1

Introduction

This thesis is driven, at the formal level, by the intuition that there is a fundamental duality between structure and information parsimony: in short, any structure in a given system affords the possibility of *simplifying* its description, and any simplification in the description of a system should only be made possible by the presence of some kind of structure in it.

The kinds of "structure" that we will consider here are rooted in group-theoretic notions of symmetry — mostly, different instances of invariance and equivariance. The unifying theme of this thesis is to reformulate, generalise and "soften" these "exact" group-theoretic symmetries with information-theoretic formalisations of the notion of information parsimony, mostly — but not exclusively — through the Information Bottleneck framework and novel extensions of it. This will often be achieved by characterising these structures through the *information-preserving compressions that they make possible*. These results mostly focus on systems with a finite number of elements. Other contributions generalise group-theoretic notions of invariance and equivariance to closed-loop stochastic actions — there, we will largely focus on a general measure-theoretic formalism that encompasses both discrete and continuous spaces.

This work is mostly mathematical, with numerical simulations supporting the exploration of several of the mathematical objects encountered. While these mathematics are valuable in their own right, they are motivated by, and aimed at, specific questions arising at the intersection of adaptive behaviour, structure discovery, and sensorimotor theories of perception.

Let me start by giving, in Section 1.1 below, an overview of these questions. In Section 1.2, I then analyse the convergences and tensions, in previous operationalisations of sensorimotor perception, that motivate the tools developed here. With this landscape laid down, I eventually present an overview of this thesis' contributions in Section 1.3 (see also the abstract for a shorter, two-pages summary of these results).

1.1 Motivations

1.1.1 Information parsimony and purposeful behaviour: a fundamental tension

This thesis is motivated by a long tradition of investigating the principles underlying adaptive behaviour — in biological and artificial agents — through the lens of information theory.[1] One of the main ideas of this line of work is that the behavioural, bodily and neural dynamics of embodied agents are shaped by — and, to a certain extent, *emerge from* — a number of informational constraints. Most relevant to us is the notion of *information parsimony*, which consists of the broad hypothesis that, all else being equal, embodied agents tend to parsimoniously consume the informational resources at their disposal — at the level of the sensorimotor interface with the environment, of internal dynamics, or a combination of both

[1] Here, the main motivation is the study of principles underlying *biological* agents' behaviour, but this includes the use of artificial systems to model it.

(Kline, 2025; Langer et al., 2024; Montúfar et al., 2015; Polani et al., 2001; Shalizi et al., 2001; Tishby et al., 2011; Tkačik et al., 2016; Tschantz et al., 2020).

What is meant by "informational resources" can vastly vary depending on the context: they can be understood as a proxy on energetic, metabolic, or computational constraints, or exist at a "purely informational" level. In the case of metabolic resources proper, an example illustrating the idea of parsimony is that of *ascidians* (commonly known as sea squirts). These organisms undergo a drastic metamorphosis along their life cycle: the swimming larva, which lacks a functional digestive system, scouts for a hospitable environment, where it cements itself on an appropriate surface. This trigers its metamorphosis into what is known as a *sessile filter feeder* — i.e., in this case, a kind of "vase-shaped" organism perpetually fixed to its rock, that feeds by filtering the sea water brought to it by currents. Crucially, this transformation involves not only the development of specific tissues and organs supporting "filter feeding", but also the *destruction* of others that were key to the larva's successful settlement: the fins are cast off, the tail is lost, the vesicle containing the light- and gravity-sensing organs retracts, and most neurons of the larval central nervous system, which primarily regulated swimming with the tail, are discarded (Cloney, 1982; Sasakura et al., 2012, 2018). This can be seen as the fact that the metamorphosis *parsimoniously re-allocates the metabolic resources* necessary to the organism's self-maintenance, in a way that adapts to the lack of self-locomotion of the post-larval form.[2]

However, the term "informational resources" often refers to concepts involving, in one way or another, Shannon information theory (Cover et al., 2009). This will also be our case in this thesis.[3] From this perspective, another example of information parsimony is given by retinal ganglion cells (which transmit retinal stimulation to downstream neurons through the optic nerve) in the salamander. In short, it has been shown that, when faced with a simple stimulus that has non-trivial temporal correlations, the activity of these retinal ganglion cells (RGCs) carries no more information about the stimulus' past than is necessary to achieve a given predictive power about the stimulus' future (Palmer et al., 2015). Similarly, in (biologically constrained models of) the fly, it has been shown that the activity of neurons from a specific area — the *Vertical System* (VS), relevant to evasive flight — carry essentially no more information about their upstream neurons' activity than is necessary to achieve a given predictive power about the fly's own movement during the evasive maneuver (Wang et al., 2021).[4] I.e., these neurons *parsimoniously* capture, within their dendritic inputs, the information predicting a variable directly relevant to the organism's survival — rather than only the information predicting their input's future, as was the case for the previous RGCs example.

Note that all three examples above involve more than just the metabolic/informational "resources" at the organism's or neurons' "disposal". Indeed, the notion of "parsimonious consumption" of the resources relies on a specific kind of *functionality*, i.e., *behavioural relevance* of the way these resources are used: the development of an organism adapted to sessile filter feeder life for the post-larval sea squirt, the prediction of the RGCs' future input for the salamander,[5] and the prediction of the evasive maneuvre for the fly. Crucially, the parsimony constraints are, so to say, *in conflict* with the enactmenent of this functionality. For instance, as the RGCs' activity can only carry information about its input's future through the one it

[2] But the widespread statement that sea squirts "eat their own brain" is now outdated: large parts of the larval central nervous system (CNS), mostly made of non-neuronal, glia-like cells, get repurposed within an adult CNS that regulates adult organs, e.g., the heart, gills, and digestive system (Sasakura et al., 2012, 2018).

[3] But this thesis does not adopt the "coding" perspective on neural activity that permeates the use of information theory in neuroscience (Brette, 2019) — see also Footnote 7 below.

[4] Here, prediction of the fly's evasive manoeuvre by VS activity is possible, in the first place, because VS activity is ultimately driven by the fly's past movement, which itself carries information about future movement.

[5] While the visual stimulus is somehow arbitrarily defined by the experimentalist (Brette, 2019), its prediction by RGCs might be induced, within the whole organism, by more intrinsic notions of behavioural relevancy.

carries about its input's past, the predictive power of this RGC activity cannot increase arbitrarily without, at some point, inducing an increase of the information about the input's past — i.e., a decrease of the information parsimony characterising RGC activity. More generally, these examples suggest that in embodied agents, there is a *tension* between stringent information parsimony constraints on the one hand, and, on the other hand, the unfolding of the agent's neural, sensory and motor dynamics driving purposeful behaviour — to the extent that the latter demand "informational expenditure". The driving force of the formal developments presented in this thesis is then the following intuition:

Main Intuition. It is *precisely* the tension between information parsimony and the enactment of purposeful behaviour that induces agents to leverage the *structure* of the interaction with their environment, as relying on such structure limits the "informational expenditure" necessary to enact purposeful behaviour.

Or, more figuratively: meaning emerges from a subtle balance of concern and laziness.

Or, more formally: this tension induces *trade-offs*, which can be seen as multi-objective optimisation problems, whose target functions model the constraints of information parsimony or behavioural relevancy, and whose solutions model the corresponding behaviourally relevant structures. Crucially, multi-objective optimisation problems are parametrised by the choice of the trade-off between the target functions, yielding a continuous family of corresponding solutions:[6] in other words, there would be a whole continuum of structures possibly induced by the tension between information parsimony and behavioural relevancy, depending on the "granularity" adapted to the use of the agent's informational resources.

Information theory, despite its historical focus on the storage and communication of digital data (Cover et al., 2009), has long ago stumbled upon formal questions of a much broader application range (Ay et al., 2017; Beer et al., 2015; Emmert-Streib et al., 2009; Jaynes, 1957; Lindgren, 2024; Shields, 1998; Touchette et al., 2004). These applications include trade-offs of the kind just mentioned. Indeed, while Shannon information was purposelly designed to abstract away any notion of "semantics" (Shannon, 1948) — which would include what I call "behavioural relevancy" — this initial framing was somehow subverted by more recent practices (Kolchinsky et al., 2018; Polani et al., 2001; Tishby et al., 2011; Tkačik et al., 2016). These variations of *rate-distortion* theory (Cover et al., 2009) propose to trade-off compression with the preservation of well-chosen notions of "relevancy". A watershed moment in this shift was the introduction of the *Information Bottleneck* (IB) method (Tishby et al., 2000) — note that the salamander (Palmer et al., 2015) and fly (Wang et al., 2021) examples above actually rely on this framework.[7]

In this thesis, we will both directly use this formal tool, and use it as the basis for our own variations and generalisations. In Section 1.1.2 below, I present this theoretical framework and its links with structure discovery. Designing precise and appropriate definitions of what is meant by "structure" will be at the core of this thesis. Here, we will root these notions

[6] Technically known as the *Pareto front* defined by the multi-objective optimisation problem.

[7] It is always delicate to adapt to a new scientific question mathematical objects that emerged from a different domain. The original metaphors from the old domain risk to introduce "Trojan horse" assumptions in the new one. And indeed, the use of information theory in neuroscience and adaptive behaviour is often characterised by "message" and "coding" metaphors that have sparked heated debates (Brette, 2019). Here, while we will heavily rely on the metaphor of "compression", I do *not* understand it as the "encoding" of a "message", but just as another word for the general notion of "coarse-graining" (though a potentially stochastic one). I will also use "channel" as a synonym of "stochastic map", but I would only reclaim the general "flow" metaphor underlying this term, rather than a specific reference to the physical transmission of strings of symbols. This should be understood similarly as in dynamical systems, trajectories are called "orbits": this legacy from celestial mechanics does not anyhow suggest that all dynamical trajectories should be "planet-like". I.e., here I follow the mathematical practice of using terminologies shaped by the history of a field, where the original metaphor carried by a term is understood as a specific case of a much more general formal structure. However, I will still avoid the terms "encoder" or "decoder", except when doing so would make technical discussions unnecessarily convoluted.

of structure in *group-theoretic symmetries*, which have become highly relevant to the study of perception, learning, and, more generally, adaptive behaviour in artificial and biological agents (Bertoni et al., 2021; Dorrell et al., 2022; Higgins et al., 2022; Keller et al., 2026; Keurti et al., 2023; van der Pol et al., 2020). Such symmetries have in particular been playing an important role in the formalisation of *sensorimotor theories of perceptions*. Together with the notion of information parsimony, the latter theories are the main conceptual motivation of this thesis, and I will present them in Section 1.1.3.

1.1.2 Information bottlenecks and structure

The IB method as originially defined in (Tishby et al., 2000) — which I will refer to as the *classic IB* — implements an optimal trade-off between the compression of a given "data" variable X (called the *source*) and the extraction of information that the latter holds about another variable Y (called the *relevancy*), yielding a third, "compressed" variable T (called the *bottleneck*). To fix the ideas, let us define the problem explicitly in the discrete case: let $\mathcal{X}$ and $\mathcal{Y}$ be finite alphabets, $\mu(X,Y)$ a given joint probability over the source and relevancy, and $\mathscr{K}(\mathcal{X},\mathcal{T})$ denote the set of conditional probabilities from $\mathcal{X}$ to a finite "bottleneck" alphabet $\mathcal{T}$. The corresponding IB problem is then, for a fixed a parameter $0 \leq \lambda \leq \Lambda := I(X;Y)$, defined as (Gilad-Bachrach et al., 2003; Tishby et al., 2000)

$$\mathrm{IB}(\lambda) := \underset{\substack{\kappa \in \mathscr{K}(\mathcal{X},\mathcal{T}) : \\ I_\kappa(T;Y) \geq \lambda}}{\arg\min} \; I_\kappa(X;T), \tag{1.1.1}$$

where the (Shannon) mutual informations $I_\kappa(X;T)$ and $I_\kappa(T;Y)$ are computed w.r.t. the distribution $q_\kappa(X,Y,T)$ defined by $\mu = \mu(X,Y)$, the compression channel κ and the Markov chain $T - X - Y$: i.e., for all $x, \in \mathcal{X}$, $y \in \mathcal{Y}$, $t \in \mathcal{T}$, we set

$$q_\kappa(x,y,t) := \mu(x,y)\kappa(t|x).$$

This condition means, intuitively, that the bottleneck T can capture information about the relevancy Y only through the information that the source X carries about Y. Solutions to (1.1.1) will always satisfy $I_\kappa(T;Y) = \lambda$.[8] Thus a compression channel κ solves (1.1.1) if, intuitively, it cannot compress X further without loosing some of the information $I_\kappa(T;Y) = \lambda$ that it captures about the relevancy variable Y; and the parameter λ controls the balance between compression of the source X and extraction of information about the relevancy Y.

For instance, the source variable can be an agent's past and the relevancy variable its future, which leads to the extraction of the most predictive features of the agent's past (Amir et al., 2015; Bialek et al., 2006; Clark et al., 2019; Creutzig et al., 2009; Sachdeva et al., 2021). More generally, the IB framework has been leveraged for unifying efficient and predictive coding principles in theoretical neuroscience — at the level of single neurons (Bialek et al., 2006; Buesing et al., 2010; Chalk et al., 2018; Klampfl et al., 2009) and neuronal populations (Buddha et al., 2013; Chalk et al., 2018; Kleinman et al., 2023; Palmer et al., 2015; Wang et al., 2021) — but also for studying sensor evolution (Klyubin et al., 2004; Nehaniv et al., 2002; van Dijk et al., 2012), the emergence of common concepts (Möller et al., 2023) and of spatial categories (Catenacci Volpi et al., 2020), the evolution of human language (Tucker et al., 2022; Zaslavsky et al., 2022), or for implementing informationally efficient control in artificial agents (Goyal et al., 2019; Lamb et al., 2022; Pacelli et al., 2019). This line of research brings increasing support to the hypothesis that, in particular for evolutionary reasons, biological agents are often poised close to optimality in the IB sense — and that artificial agents can reap significant advantages by being so as well. Moreover, since its inception, many variations

[8] See, e.g., Lemma 4 in Appendix A of (Charvin et al., 2023b).

and generalisations have greatly extended the classic IB. For instance, a kindred trade-off between the information parsimony of decision-making and reinforcement learning-like utility was formalised in (Tishby et al., 2011); a variant aimed at capturing interactive learning was proposed in (Still, 2009); and the IB was generalised to the multi-variate case in (Slonim et al., 2006), an instance of which is the distributed IB (Aguerri et al., 2017), used for analysing the structure of complex (Murphy et al., 2022b) and chaotic dynamical systems (Murphy, 2024; Murphy et al., 2022a).

The latter complex systems-related results exemplify how generalisations of the classic IB framework can be used to investigate not only the information optimality of a given system, but also the underlying *structure* induced by this informational optimality. This link to the structure of complex systems actually already holds for the classic IB method, through its connection to *computational mechanics* — a specific approach to the automatic discovery of patterns and structure in complex dynamical systems (Crutchfield, 2017). Indeed, if, similarly as in the first example above, we compress a (double-sided) stationary stochastic process' infinite past while preserving information about its infinite future, and favor as much as possible the latter over the former, then we obtain a construction known as the *ϵ-machine* (Shalizi et al., 2001) — see also (Grassberger, 1986). This object has been argued to capture the causal architecture of a given stochastic process (Shalizi, 2001), and is a central tool of computational mechanics — in particular, it is an important driver of work on the discovery of generalised symmetries in time-series (Rupe et al., 2022) or the emergence of coarse-grained processes from a lower level of description (Rosas et al., 2024). The ϵ-machine has also been generalised to stochastic processes with actions (known as *input-output* processes), yielding the *ϵ-transducer* (Barnett et al., 2015), which in particular allows for models of embodied agents based purely on their sensorimotor interface — i.e., without any reference to an "external world" defined independently from this sensorimotor interface (Rosas et al., 2025).

While ϵ-machines provide an example of IB problem where the compression-relevancy trade-off maximally favors the preservation of relevant information, exploring the full continuous range of trade-off parameters λ in the problem (1.1.1) exhibits a rich structure. Indeed, when λ decreases — thus increasingly favoring compression — the corresponding bottleneck solutions go through successive *bifurcations*, at which the intrinsic complexity of the bottleneck successively collapses (Agmon et al., 2021; Gedeon et al., 2012; Ngampruetikorn et al., 2021; Wu et al., 2020). More precisely, in the finite case, this consists in some distinct bottleneck symbols merging into a single symbol — thus inducing a reduction in what is known as *effective cardinality* (Zaslavsky et al., 2019) — while in the continuous case (or, at least for Gaussian variables), it is dimensions of the bottleneck solutions that successively collapse (Chechik et al., 2005). This shows that the IB framework can be seen as a specific information parsimony perspective on *dimension reduction* (Globerson et al., 2003; Martini et al., 2024) — and conversely, it suggests that methods involving dimension reduction might often carry formal links with well-chosen information-theoretic trade-offs.

The structure-extraction capabilities of the IB are also related to the IB theory of deep learning (Shwartz-Ziv et al., 2017). The latter claims in particular that, at least in the case of supervised learning, deep networks' strong generalisation capabilities stem from their successive layers implementing increasingly compressed information bottlenecks. This theory has been strongly challenged (Saxe et al., 2019), sparking new research in weaker versions of it (Lorenzen et al., 2022), the extent to which the classic IB is at least a normative framework for deep networks (Achille et al., 2018b; Elad et al., 2019; Kawaguchi et al., 2023; Lorenzen et al., 2022), and a recent variation based on *synergy* that addresses several of the original theory's limitations (Westphal et al., 2025). On the other hand, the generalisation capabilities of deep networks have been shown to rely on their *invariance extraction* properties (Achille et al., 2018a; Deng et al., 2022; Lyle et al., 2020): together with the IB theory of deep learning, this suggests that the IB induces the extraction of invariances. The IB framework has also

been leveraged for graph (supervised) structure learning (Sun et al., 2022) and self-supervised learning (Shwartz Ziv et al., 2024).

Beyond the IB framework itself, it has been shown that information parsimony principles — operationalised in a Bayesian setting through the optimisation of marginal likelihood — can indeed be leveraged for data-based discovery of invariances (Ouderaa et al., 2022; van der Wilk et al., 2018) and symmetries in Hamiltonian dynamics (van der Ouderaa et al., 2024).

These results from the machine learning literature provide a certain kind of "experimental" evidence suggesting the existence of strong, formal links between diverse notions of structure and corresponding notions of information parsimony. In particular, as "invariance" in deep neural networks is often formulated in the language of group theory, this suggests a connection between the IB framework and group-theoretic symmetries. However, to the best of my knowledge, the existing literature is surprisingly silent on explicit connections between group symmetries and the IB framework. Links between information and group theory have already been investigated in different contexts — e.g., equivalences between group inequalities and information inequalities (Yeung, 2008), applications of group algebras to coding theory (Guerreiro, 2016), or information theory on Lie groups (Chirikjian, 2012). But these contributions do not touch upon a potential information parsimony counterpart of group-theoretic symmetries — not to mention leveraging such information parsimony to discover these symmetries. On the other hand, (Möller et al., 2023) explicitly investigates, in synthetic experiments on a minimal model, the emergence of symmetries using a variation of the classic IB method, but it leaves many questions open, in particular concerning the mathematical objects that might underlie the numerical phenomena exhibited there. This thesis aims at bridging this gap:

Motivation 1. Investigate the links between the IB framework — broadly understood — and group-theoretic symmetries, and ground them in explicit mathematical results.

This is of course relevant to the adaptive behaviour motivations mentioned in Section 1.1.1. But the aim, here, is also to open the way to *information theory-based symmetry discovery*, as well as provide new methods for *symmetry-based coarse grainings*. As we will see, these "basic science" and "technological" motivations raise common rich and unresolved formal questions, which will be our focus in this thesis. To understand why, we first need to turn to sensorimotor theories of perception, and the relevance of group-theoretic symmetries to their formalisation.

1.1.3 Sensorimotor theories of perception

The problem, for embodied agents, of leveraging behaviourally relevant structure in the interaction with their environment, is directly relevant to sensorimotor theories of perception. By this, I am referring to a family of related approaches to studying perception in biological agents and building it in artificial ones, which investigate the claim that perception is based on *the way the agent's own actions affects its sensory input*, rather than on the passive processing of sensory input.[9] There is a vast body of literature related to this idea, from experimental (O'Regan et al., 2001; Rolfs et al., 2022), developmental (Jacquey et al., 2019; Piaget, 1964) and ecological (Gibson, 2014) psychology to neuroscience (Ahissar et al., 2016; Berthoz et al., 2000; Brette, 2019; Buzsáki et al., 2019; Pezzulo et al., 2024), developmental robotics (Godon et al., 2020; Hoffmann et al., 2017; Laflaquière et al., 2015b; Olsson et al., 2006), machine learning (Caselles-Dupré et al., 2021b; Keurti et al., 2023; Marchetti et al., 2023) or enactivist theory (Di Paolo et al., 2017; Varela et al., 1992). The approach that is the most relevant to this thesis is the theory of *sensorimotor contingencies* (SMCs).

[9] In this thesis, "sensory input" will always mean the *raw* sensory input impinging on the agent's sensory surface (e.g., photoreceptors, tactile receptors, viscerocceptive receptors, etc).

As this theory is quite subtle and can seem counter-intuitive in the contemporary research landscape, let us start with full quotes from its foundational paper (O'Regan et al., 2001), which sits at the intersection of vision research, experimental psychology and philosophy of mind. Our examples will also focus mostly on vision, even though SMCs are relevant to all sensory modalities.

Apparatus-related and object-related SMCs

The theory starts from the argument that from the agent's internal perspective, the stream of raw sensory inputs does not carry, on its own, any meaningful information:

> From the point of view of the brain, there is nothing that in itself differentiates nervous influx coming from retinal, haptic, proprioceptive, olfactory, and other senses, and there is nothing to discriminate motor neurons that are connected to extraocular muscles, skeletal muscles, or any other structures. Even if the size, the shape, the firing patterns, or the places where the neurons are localized in the cortex differ, this does not in itself confer them with any particular visual, olfactory, motor or other perceptual quality. (O'Regan et al., 2001)

From this internal perspective, the brain is akin to a "team of engineers operating a remote-controlled underwater vessel" whose "connections to and from the underwater cameras, sonar equipment, robot arms, actuators, and sensors" would have been scrambled by a "villainous aquatic monster": the engineers' only way to make sense of the vessel's sensory input is to "press various buttons and levers", and, essentially, see what happens.[10] I.e., the sensory influx only acquires meaning once it is probed by the agent's own actions:

> On the other hand, what *does* differentiate vision from, say, audition or touch, is the *structure of the rules* governing the sensory changes produced by various motor actions, that is, what we call the sensorimotor contingencies governing visual exploration. Because the *sensorimotor contingencies* within different sensory domains (vision, audition, smell, etc.) are subject to different (in)variance properties, the structure of the rules that govern perception in these different modalities will be different in each modality. (O'Regan et al., 2001)

For instance, vision is characterised by how "when the eyes rotate, the sensory stimulation on the retina shifts and distorts in a very particular way, determined by the size of the eye movement, the spherical shape of the retina, and the nature of the ocular optics" (O'Regan et al., 2001). An experimental phenomenon related to this claim is that of *sensory substitution devices* (Eagleman et al., 2023). E.g., vision can be partially "restored" in blind subjects thanks to a wearable device where the visual input captured by a head-mounted camera activates a small grid of electrodes placed on the tongue. After training, users manage to discern rich visual qualities such as distance, shape, direction of movement and size (Stronks et al., 2016), and report a subjective experience of vision rather than tactile stimulation. SMC theory contends that these sensory substitution devices do literally provide a visual experience. This experience is of course not exactly the same as usual, "eye-based" vision: for instance, it does not capture the sensorimotor contingencies induced by *saccades* (i.e., the fast, jump-like eye movements that occur when gaze changes direction). An explicit example of SMC related to these rapid eye movements has been uncovered by recent experimental work (Rolfs et al., 2025): it showed that the specific kinematic laws satisfied by saccades precisely shape — on the individual level — the limits of perceiving stimuli moving at high speed. This result sits within a broader line of work that investigates how the *incidental* sensory consequences of motor behaviour (e.g., eye movements) shape the visual apparatus (Rolfs et al., 2022).

[10] The quotes here refer to (O'Regan et al., 2001).

The "inter-modality" argument above can also be made for the attributes of objects in a given sensory modality. E.g., for vision: shape, texture or color cannot be basic features extracted by the brain from its sensory input only, as from an internal point of view, the nervous influx triggered by the sensory input does not discriminate, in itself, between these different features. Or, more precisely:

> It is tempting to think that seeing red is like seeing pink because the neural stimulation going on when we see something red is similar to that underlying our perception of pink: almost the same ratios of long, medium and short wavelength photoreceptors will be stimulated by red and pink. But note that though this seems reasonable, it does not suffice: there is no *a priori* reason why similar neural processes should generate similar percepts. (O'Regan et al., 2001)

Only once the agent explores the consequences of its own actions on this sensory influx, can there emerge a meaningful structure on which perception can be based. For instance, in the case of the shape of rigid objects:

> The idea we wish to suggest here is that the visual quality of shape is precisely the set of all potential distortions that the shape undergoes when it is moved relative to us, or when we move relative to it. Although this is an infinite set, the brain can abstract from this set a series of laws, and it is this set of laws which codes shape. (O'Regan et al., 2001)

Here, the "meaningful structure" underlying the perception of a rigid object's shape is thus a well-chosen set of *potential sensory changes*, with a key role of those changes induced by the agent's own actions. As we will see below, this movement-based perspective resonates with several sensorimotor notions of percepts proposed in the literature (Ahissar et al., 2016; Buzsáki et al., 2019; Gibson, 2014; Keller et al., 2026; Poincaré, 1952; Seth, 2014; Tsao et al., 2022). It will also be one of the key motivations of Chapter 3.

SMC theory thus distinguishes between two kinds of SMCs: on the one hand, the regularities induced specifically by the agent's own embodiment — e.g., the visual apparatus. Following (O'Regan et al., 2001), we will call them *apparatus-related SMCs*. On the oher hand, many regularities depend, beyond the agent's embodiment, on the specific content of the ongoing agent-environment interaction. Such regularities are instrumental to capture what is usually referred to as the attributes of objects (e.g., shape, color, spatial contiguity, etc): following (O'Regan et al., 2001) again, we will call them *object-related SMCs*. Of course, the distinction is in practice not clear-cut: the interaction with the environment clearly shapes the sensory apparatus on the time-scales of evolution, development and learning — e.g., it has been argued that the statistics of natural images enhances the visual apparatus' sensitivity to horizontal and vertical orientations (Appelle, 1972; Dragoi et al., 2001; Field, 1987; Girshick et al., 2011). But the distinction is still conceptually useful, and will be important throughout this thesis: the tools developed in Chapter 2 are relevant to apparatus-related SMCs, while those developed in Chapter 3 are more relevant to object-related SMCs.[11]

Skillful exercise of SMCs

While SMCs are learned during development and learning (Di Paolo et al., 2014; Jacquey et al., 2019), their perception does not consist in the activation of an internal representation,

[11] The term "object-related SMCs" does not capture adequately the facts that (*i*) these regularities depend on the agent's behaviour as much as on the content of the environment, (*ii*) the relevance of such regularities should ultimately be understood from the agent's intrinsic point of view, rather than from that of an experimentalist arbitrarily singling out "stimuli" of interest, and (*iii*) as argued below, perceptually relevant structure might consist, fundamentally, of dynamic spatiotemporal processes rather than static "objects". However, at this stage, we choose to stick with the terminology proposed in (O'Regan et al., 2001).

but in the ongoing, "*skillful exercise*" of appropriate "methods for probing the outside world" (O'Regan et al., 2001) that rely on the "*mastery*" of these SMCs. E.g., visual percepts are not merely patterns of retinal activation that map topographically to cortical activations: rather, visual perception only fully unfolds through *exploration* through eye, head, or whole body movements. However, while the broader field of *active vision* (Rolfs, 2015) investigates how vision relies on the exploration of visual scenes, SMC theory goes at a more fundamental level, and claims that what is ultimately being explored is the "*structure of changes*" induced by movements on the sensory influx.

Importantly, the notion of "skillful exercise" involves both learning and real-time behaviour: from this sensorimotor perspective, perceiving means *enacting* certain SMCs through actual movement on the time-scale of perception itself, but these SMCs can only be enacted because the agent (including its brain dynamics) has "*attuned*" to them along its sensorimotor history, thus making them "potentially available for recall" (O'Regan et al., 2001) in the here-and-now.[12] In particular, **SMC theory casts doubt on the relevance of the very notion of representation** for the study of perception, as it argues that at their most fundamental level, percepts do not "refer to" phenomena in the external world that can be defined independently of the agent, but are a specific kind of embodied *activity* — an activity that can be defined purely from the agent's intrinsic perspective, and involves the whole sensory-neural-motor loop.[13]

To make these ideas more concrete, let us turn to a recent experimental result regarding the detection of contact with an object by freely moving rats through their whiskers (Nelinger et al., 2025). This work shows that contact of the whiskers with an object relevant to a discrimination task is characterised by a specific spatiotemporal pattern in the whiskers' sensorimotor dynamics. In short, this pattern consists of the convergence to a one-dimensional "attractor"[14] in a state-space made of two coordinates: the time derivative of a variable controlled by the rat's movement (here, the whisker's angle), and the time derivative of a variable inducing a sensory influx (here, the whisker-base's curvature, which activates mechanoreceptors driving downstream neurons' activity). Importantly, along this attractor, a small change in angle produces a large change in curvature: i.e., the sensitivity of the mechanoreceptor to the rat's movement is heightened, making the attractor an efficient "method for probing" the objects that the rat is trying to discriminate between. Of course, the relation of this specific sensorimotor dynamics' pattern with simultaneously ongoing brain dynamics deserves further research, and it describes a limited aspect of the interaction between the rat and the object. But this example illustrates what it can mean for a percept (here, that of "whisker contact with a behaviourally relevant object") to be defined purely from an agent's internal point of view (here, an attractor of dynamics on a well-chosen sensorimotor space), and as an ongoing activity rather than the activation of an internal "representation". Note also that the fact that here the percept corresponds to a *low-dimensional* attractor suggests a link with information parsimony.

A common objection to these claims is that, while movements might be important for perception, they do not seem *necessary*, at least for simple forms of perception. As put in (Buzsáki et al., 2019), which proposes similar ideas:

[12] While "attunement" is not explicitly defined in (O'Regan et al., 2001), an explicitation in the language of dynamical systems is proposed in (Buhrmann et al., 2014). The formal tools developed in Chapter 3 are also aimed at clarifying this notion.

[13] To be compared with the currently widely accepted claims that "mental capacities involve computations acting on representations" and that "brains represent stimuli" — even though it is not always clear what scientists precisely mean by "representation" (Favela et al., 2023). Here, I mean some kind of internal brain activity that would "refer to" some kind of external phenomenon defined independently of the agent.

[14] (Nelinger et al., 2025) does not propose a mathematical model capturing the dynamics of their experimental data.

> Of course, you may counter my arguments by simple introspection and say that you can sit completely immobile and yet perfectly process the sensory flow. A touch on your hand or a bug flying in your visual field can be detected with no muscular effort whatsoever. ((Buzsáki et al., 2019), p.76.)

A first answer to this criticism is that "complete immobility" simply never happens under natural conditions (Musall et al., 2019), while even movements that do not reach awareness may actually matter quite a lot for perception. For instance, even between saccades, the eyes undergo incessant movements, known as *fixational eye movements* (FEM). If the effect of this motion on retinal activation is artificially cancelled, e.g., by displaying the visual stimulus on a contact lens-like device, vision fades drastically after a few seconds (Yarbus, 1967). Despite a loss of interest at the end of the 20th century, the study of FEM has a long history (Rolfs, 2009) and has in recent years seen a series of strong results. Indeed, it appears increasingly clear that these fine-grained ongoing movements, far from being perturbations to be compensated for by some internal processing in order to maintain a stable "spatial" activation in the primary visual cortex, might actually play a fundamental positive role in visual perception (Boi et al., 2017; Casile et al., 2019; Clark et al., 2022; Intoy et al., 2024; Wu et al., 2024) and visual learning (Arató et al., 2024). In particular, these results show that retinal activation is not a spatial stimulus, but an irreducibly *spatiotemporal* sensory flow (Rucci et al., 2018) that is integrated by the brain with oculomotor signals (Hafed et al., 2021).

Which role for ongoing brain dynamics in sensorimotor perception?

But the above argument — that complete immobility is never achieved in natural behaviour, and that this "residual" movement might be fundamental for perception — does not resolve the ambiguities of all the kinds of SMCs described in (O'Regan et al., 2001). For instance, in the case of shapes: it is clearly not the case that, say, a coin, must be explored *in real-time* under the "infinite set" of all possible perspectives on it to be perceived as disk-shaped. Rather, it seems that here, "mastery" of the SMC describing the coin's "shape" should mean that after learning, the corresponding sensorimotor percept can be perceived *without exhaustively enacting the SMC* through actual movement (Seth, 2014). Or, maybe it should rather mean that what is being "enacted" is *something more than a pattern of bodily movement and corresponding raw sensory influx* — something that somehow involves the unfolding of the potential changes of perspectives without physically realising them, or at least not all of them.

This example touches upon what is, in my understanding, one of the most ambiguous and contentious aspects of SMC theory. Let me first start by highlighting two points on which (O'Regan et al., 2001) is explicit: on the one hand, brain activity is *not sufficient* for perceptual experience. On the other hand, the "attunement" of brain activity to SMCs along development and learning is a crucial aspect of the "mastery" of these SMCs, thus making it a *necessary* condition for the "current exercise" of this mastery, i.e., for perception. What remains ambiguous here, however, is the role of brain activity *on the time-scale of perception itself*: if perception coincides with the "current exercise" of the "mastery of SMCs", what is the role of ongoing brain dynamics in this "exercise"? To the best of my knowledge, this ambiguity is characteristic of SMC-based accounts of perception. In this literature, mentions of brain activity have, historically, often focused on questioning the representationalist narrative that this activity is usually described with. From a sensorimotor perspective, this is indeed a much needed contribution in the contemporary research landscape. But it does not question, in itself, the potentially consitutive role of ongoing brain dynamics in perception: *it just argues that this role should be understood in non-representationalist terms*, and through the lens of the sensorimotor level — on the time-scale of perception (i.e., ongoing sensorimotor dynamics) but also development and learning (i.e., sensorimotor history).

One possible role of ongoing brain dynamics w.r.t. ongoing sensorimotor dynamics is simply to regulate the latter along the "current exercise" of SMCs. E.g., not only do fixational eye movements seem to play, as mentioned above, a fundamental role in visual perception: these movements also happen to be influenced by cognitive factors (Benedetto et al., 2023; Lin et al., 2023), finely tracked by the brain (Zhao et al., 2023) and indeed actively controlled during vision (Intoy et al., 2020; Poletti et al., 2015; Willeke et al., 2019), with a prominent role of sensory-motor integration in the superior colliculus (Hafed et al., 2021). In other words, the skillful exercise of SMCs seems, at least for vision, to crucially involve ongoing neural activity.

However, this still does not address the example of shapes mentioned above: rather, this example suggests that, if perception is to coincide with the current exercise of SMCs, then the "exercise" must also involve some kind of "fictional level", i.e., a level distinct from the sensorimotor interface itself. This could take the form of internal brain dynamics that have been shaped by the agent's sensorimotor history, but that on the time-scale of perception are at least partially decoupled from ongoing sensorimotor dynamics — where crucially, this decoupling would *not* make real-time brain dynamics less constitutive of the corresponding percept. It turns out that the *inside-out* approach (Buzsáki et al., 2019) — based on experimental neuroscience research — goes precisely in this direction.[15]

While similar in several respects to SMC theory, the core argument of this framework is that neuroscientists should adopt a fully *intrinsic* perspective on the study of brain dynamics: i.e., the brain should be studied from the point of view of the brain itself, as a self-sustained dynamical process that receives feedback from its own activity and is constantly reshaped by it.[16] More precisely, here neuronal activity is seen as similar to a "dictionary with pre-existing internal dynamics and syntactical rules but filled with initially nonsense neuronal words"[17]. These internal dynamics only acquire meaning through exploratory actions that "*ground*" them in the sensorimotor interface, thus "*calibrating*" them and eventually leading to a certain "*internalisation*" of the effect of actions on the sensory influx. (Buzsáki et al., 2019) argues that after development and learning, these internalised patterns might then be triggered with only limited involvement of the sensorimotor interface, or even with none at all. E.g., during human pregnancy, a specific class of "spindle-shaped" oscillatory patterns in the baby's somatosensory cortex induce uncoordinated kicks that are instrumental in the formation of body maps, as "the initially meaningless, action-induced feedback from sensors transduces the spatial layout of the body into temporal spiking relationships among neurons in the brain"[18]. But shortly after birth, these patterns become fully internalised, becoming what is known as thalamocortical "sleep spindles", which usually occur during (non-REM) sleep and induce no movement.

Interestingly for us, in the latter example, what we usually think of as the *spatial* structure of the body is, from the point of view of the brain, made of fundamentally *spatiotemporal* patterns of neuronal activation, that involve the brain's interface with both tactile sensors and movement initiation. This perspective aligns with the above results on fixational eye movements suggesting that visual perception is, at its core, based on the spatiotemporal evolution

[15] See also (Jost, 2016) for similar ideas from a non-linear dynamics and information theory perspective.

[16] Interestingly, this shift in neuroscience research resonates with an older one operated by 19th century geometers: i.e., the shift to an intrinsic perspective on the mathematical study of curves and surfaces. This allowed for the study of manifolds' geometric structure without any reference to an "ambient space" containing them, eventually leading to modern differential and Riemannian geometry. One may hope for a dynamical version of this geometric shift — where, e.g., the dynamical structure of "subprocesses" embedded in a broader, "ambient" stochastic process could be studied without reference to this "ambient" process. Contributions on coarse-graining and (causal) emergence in (stochastic) dynamics (Atay et al., 2017; Barnett et al., 2021; Pfante et al., 2014; Rosas et al., 2024; Rupe et al., 2024), as well as causal blankets (Rosas et al., 2020), could be relevant to this aim.

[17] (Buzsáki et al., 2019), p.32

[18] (Buzsáki et al., 2019), p. 78.

of both retinal and oculo-motor signals, rather than just spatial, "picture-like" retinal inputs. These two examples point to a broader picture that encompasses most of the ideas presented in this section: **the kind of structure underlying perception would be, fundamentally, *dynamical structure in the sensori-neural-motor loop***. The whole question becomes, then, to describe explicitly this dynamical structure, how exactly it integrates the sensorimotor and neural levels, and how dynamical structure on different time-scales — of perception itself, of learning, of development — interact with each other.

I have outlined some pieces of answer to these questions, mostly from SMC theory and more briefly from the inside-out approach. Other directions are offered by related frameworks. In particular, *Closed-Loop Perception* theory (Ahissar et al., 2016, 2025) frames percepts as *attractors* in the sensori-neural-motor loop (our above example of rat whisker contact (Nelinger et al., 2025) is part of this line of work); while enactivist approaches propose a similar dynamical focus on SMCs (Buhrmann et al., 2013) and integrate them within a broader sensorimotor *autonomy* framework (Barandiaran, 2017).

Importantly, these approaches are both conflicting and complementary. In particular, the inside-out approach's focus on the internalisation of sensorimotor dynamics seems to dismiss SMC theory's insistence on the central role of movement on the time-scale of perception itself. But on the other hand, it has the potential to clarify SMC's concept of *attunement*: during perception, brain dynamics might be coupled to ongoing sensorimotor dynamics, but also have, to a certain extent, their own self-sustained dynamics which *carry a rich sensorimotor history* (Virgo et al., 2022). This history may be instrumental in *shaping the ongoing, actual sensorimotor interaction*, but also in *amplifying this ongoing interaction at a "fictional" level*, i.e., an internal level that is at least partially decoupled from the sensorimotor interface.[19] Importantly, here, this decoupling would only hold on the time-scale of perception: on the time-scales of learning or development, the ongoing sensorimotor interaction would, on the contrary, shape the "fictional" level. Note that this point of view could interlock well with previously proposed models of sensorimotor habits as "self-sustaining patterns of sensorimotor behaviour" (Egbert et al., 2014, 2022).

Towards novel theoretical infrastructures

It is crucial to acknowledge that SMC theory maintains a certain amount of ambiguity in the concepts that it introduces, including what the term "sensorimotor contingencies" itself means exactly. The latter are described in (O'Regan et al., 2001) as a "structure of rules" or as "lawful regularities", and these concepts are refined and supported by experimental examples in the latter reference. Research in SMC and related theories has since made important progresses to refine these concepts further, investigate concrete examples in experiments, and build formalisms and artificial systems that operationalise these principles (we mentioned some above and will mention more below). But, still today, a central part of research in sensorimotor perception remains to *build the appropriate language*, the appropriate concepts to explicitly describe what kind of "structures" precisely are relevant, in the sensorimotor interface of embodied agents, to the emergence of perception. The challenge, here, is that we are dealing with spatiotemporal, dynamical structures which are *difficult to "probe" through experimentation without a dedicated theoretical framework*. In particular, such a framework should (*i*) describe *what* exactly are these structures, (*ii*) single them out according to their *behavioural relevancy* from the agent's intrinsic perspective, and (*iii*) offer *methods* to investigate them through the processing of well-chosen experimental data, or the simulation of adaptive behaviour. While philosophy, experimental psychology, neuroscience or robotic and

[19] Let me insist that focusing on an internal level does *not* in itself require any notion of representation: the internal dynamics do not need to refer to "external objects" arbitrarily defined by an experimentalist to have a rich and meaningful structure.

computational modeling are all irreplaceable tools in this endeavour, fully explicit mathematical formalisms can be instrumental in pinning down such behaviourally relevant structures, and developing a theoretical "infrastructure" to support their exploration in both simulations and experiments. This is where this thesis stands with respect to SMC theory:

Motivation 2. Develop novel mathematical frameworks that are relevant to the formalisation of sensorimotor theories of perception, with a primary focus on sensorimotor contingencies theory.

Let me emphasize, however, that *this thesis does not propose models or simulations of concrete embodied agents*: rather, the focus is on some of the formal questions raised by sensorimotor theories. Indeed, the last decades of modeling work in sensorimotor perception research have shown the relevance of a rich landscape of pre-existing mathematical tools. This offers anchor points for the development of novel, specifically tailored formalisms — at the intersection of group theory, dynamical systems and information theory — that would unlock new avenues for progress on SMCs. Conversely, it appears increasingly clear that the formal questions raised by sensorimotor theories have such a depth that they can give birth to fresh mathematics with a much broader application range, while being interesting in their own right.

This is, at least, what I hope the next section will convince you of: let us now have a look at previous operationalisations of sensorimotor perception, and how they relate to other contemporary approaches to perception and learning based on group-theoretic symmetries. This will clarify the link between the ideas outlined in the present section and this thesis' theoretical invesigations.

1.2 Operational approaches to sensorimotor perception: review & conceptual analysis

In this section, I review previous work aimed at operationalising sensorimotor theories of perception,[20] or from a different background but directly relevant to this aim. Along the way, I provide a conceptual analysis of this formal and computational research landscape, through the lens of sensorimotor theories. This analysis highlights the convergences but also tensions that will motivate the formal tools developed in this thesis. For more details on the group-related notions discussed here, including those of "invariance" and "equivariance", see Appendix A.

1.2.1 Geometric, probabilistic & informational approaches

Some sensorimotor aspects of perception were already formulated in a mathematical language, more than a century ago, by H. Poincaré (Poincaré, 1952). He claimed that our notion of space can only emerge from our ability of *moving* in space: more precisely, of performing movements that compensate specific changes of the environment.[21] E.g., eye movements can correct for the displacement of a moving object in the visual field, thus maintaining the retinal stimulation that it induces on the fovea; while if a liquid changes color due to a chemical reaction, no body movement will revert the liquid's color. Poincaré proposed to formalise this as the existence of a *group* made of invertible transformations of the sensory input that can be obtained equally well as a change of the environment independent of the agent's action, or as an agent's own action: i.e., the group of, precisely, *spatial* transformations. The learning

[20] Several contributions to this agenda will however not be mentioned, as I focus on the ones most relevant to this thesis. See also (Pak, 2025) for a recent review.

[21] See Chapter IV in (Poincaré, 1952).

of this group would then lead to the emergence of the sense of space. These ideas were a direct inspiration to one of the first formal contributions to SMC theory (Philipona et al., 2003) which, interestingly for us, consisted in capturing this group as a *low-dimensional subspace* of the agent's high-dimensional sensorimotor space — while no mention of Claude Shannon is to be found in this work, it is worth noting the conceptual similarity with information parsimony. This sensorimotor approach to the discovery of 3D space, including its topological and metric properties, was further pursued (without group theory) in (Laflaquière, 2020; Laflaquière et al., 2015b, 2019; Terekhov et al., 2016), with extensions to an agent's tactile space in (Marcel et al., 2017, 2022).

Another geometrical perspective on sensorimotor perception explicitly links it to movement generation, exploring the hypothesis that the brain implements different kinds of geometries depending on the movements that it participates in, and the corresponding body spaces that these movements involve (Bennequin et al., 2009, 2017, 2025; Langlois et al., 2024). This approach has increasingly called on, and been intertwined with the active development of, tools from category theory — more specifically, the theory of *toposes* and *stacks* (Belfiore et al., 2022).

SMCs have also been studied from a *predictive processing* (Clark, 2016) perspective. For instance, (Seth, 2014) proposes to interpret the "potential" dimension of SMCs (see Section 1.1.3) through *counterfactually rich* hierarchical generative models, i.e., ones that predict "not only the likely causes of current sensory inputs, but also the likely causes of those sensory inputs predicted to occur given a large repertoire of possible (but not necessarily executed) actions". While, to the best of my knowledge, this idea has not yet been implemented in *hierarchical* models, several contribution used simpler predictive models of sensorimotor transitions (namely, conditional probabilities between sensory or sensorimotor states) as the basic mathematical object from which to model object discovery (Laflaquière et al., 2015a), study the structure a simulated visual field (Laflaquière et al., 2018), or induce a topology on a visual sensory space (Goasguen et al., 2023). In a similar direction, (Godon et al., 2020) uses a group-theoretic language to study the compositional structure of motor actions through that of sensory predictions.

Information-theoretic tools were used to study the emergence of sensor structure (Olsson et al., 2006), body structure (Díaz Ledezma et al., 2023) and multi-modal SMCs (Ledezma et al., 2025) in naive agents based only on their sensorimotor data — though in these contributions, information theory is not explicitly used to formalise information parsimony. On the other hand, less directly relevant to SMC theory but closer to information parsimony: (De Llanza Varona et al., 2024) investigates action-centric representations with tools very similar to the *variational IB* (Alemi et al., 2017); while within the *active inference* framework (Pezzulo et al., 2024), similar questions have been investigated in (Tschantz et al., 2020), and structure learning in (Friston et al., 2024).

However, previous work explicitly linking, in an embodied agency setting, group-theoretic symmetries to information parsimony, is to the best of my knowledge very scarce. One exception is (Möller et al., 2023) which investigates, in a minimal model, how information parsimony among a group of agents can induce the emergence of symmetries in the concepts shared by this group — where information parsimony is operationalised with a variation of the IB method. In particular, the latter work investigates what it calls *intrinsic symmetries*, i.e., permutations of the agents' spaces of sensors and actuators that preserve, in some specific sense, the "quality" of a common concept. Interestingly, these intrinsic symmetries resonate strongly with the scrambling of an agent's sensorimotor space described informally in (O'Regan et al., 2001), through the metaphor of the "villainous aquatic monster" (see Section 1.1.3): indeed, intrinsic symmetries are explicitly thought of as "transplanting [the agent's] brain into a rewiring of [its] sensorimotor embodiment" (Möller et al., 2023). While the multi-agent setting from (Möller et al., 2023) has its own aims, not necessarily related to SMC theory,

this suggests an interesting direction: capturing certain aspects of SMCs with compressions of an agent's sensorimotor space that would be *defined* by permutations of an agent's sensor and action spaces preserving a well-chosen, behaviourally relevant quantity.

1.2.2 Enactivist & dynamical approaches

Another line of work refines the partially ambiguous concepts from (O'Regan et al., 2001) by combining philosophical work rooted in the *enactivist* tradition (Di Paolo et al., 2017; Varela et al., 1992) with simple but tractable *dynamical* models. In particular, it has been proposed to differentiate between four kinds of sensorimotor relations:

> We can distinguish sensorimotor relations in senses that vary with the degree of agent-centredness. One possible relation describes how sensory input changes with induced motor activity in an *open-loop* fashion. This depends on the embodiment of the agent and the environment only, not on what the agent is actually doing. Another relation looks at co-variations that obtain once the *loop is closed* by taking into account the agent's internal activity and responsiveness to sensory changes. The next sensorimotor relation is more specific and looks at the coordination patterns that contribute to the performance of a *task*. And finally, another sensorimotor relation indicates how such coordination patterns may be organized *normatively* so as to distinguish levels of skilfulness, efficiency, stability, etc. [(Buhrmann et al., 2013), italics are mine.]

(Buhrmann et al., 2013) calls these four levels of sensorimotor relations, resp.: sensorimotor environment, sensorimotor habitat, sensorimotor coordination and sensorimotor strategy. The first three are then exemplied by solving a simple categorical perception task with coupled differential equations modeling the agent-environment interaction. A similar methodology has been used to operationalise the concepts of "attunement" and "mastery of SMCs" in a non-representationalist direction (Buhrmann et al., 2014), explore the links of SMCs with the emergence of sensorimotor habits (Egbert et al., 2014, 2022), and argue for the fundamental role of closed-loop coupling with the environment in internal (e.g., brain) dynamics (Aguilera et al., 2013).

This approach shows how the rich philosophical and methodological framework of enactivism can help gradually disambiguate SMC theory, and, along the way, integrate the *structural* perspective of the latter with the *norm-based* perspective of the former (Barandiaran, 2017). Such an effort is particularly relevant to this thesis: the formal contributions presented here can be seen as a preliminary step in this direction. More precisely, on the one hand, one of the motivations for exploring a duality between structure and information parsimony is to show that the structural aspects of sensorimotor perception (e.g., symmetries of the sensorimotor environment or habitat) actually *already* adress a certain norm — if one accepts to regard information parsimony as an implicit norm regulating the agent's coupling with its environment. On the other hand, here we will not try to formalise any notion of non parsimony-related norm.

Enactivist approaches, though, are not the only ones to operationalise the dynamical aspects of SMCs. For instance, *Closed-Loop Perception* theory (Ahissar et al., 2016, 2025; Nelinger et al., 2025) proposes that percepts correspond to attractors in sensori-neural-motor dynamics (see Section 1.1.3). In a similar direction, *NeuroEcological Nexus Theory* proposes to formalises affordances (Gibson, 2014) — a notion similar in many ways to that of object-related SMC — as the coupling of dynamics on low-dimensional manifolds at resp. the neural, bodily and environmental level (Favela, 2024). Note that the latter theoretical framework is

based on non-linear dimensionality reduction, which suggests possible links with information parsimony (see Section 1.1.2).[22]

1.2.3 Representation learning & SMCs: a paradoxical convergence

Meanwhile, the group-theoretic framework suggested by Poincaré (Poincaré, 1952) and which initiated the formalisation of SMC theory (Philipona et al., 2003) has been absorbed, to a significant extent, by a growing line of work in the machine learning field of *representation learning* — even though, ironically, SMC theory is strongly critical of the concept of representation (see Section 1.1.3). Indeed, in the past decade, the representation learning community has turned its attention to representations that are structured by group symmetries. This led to the emergence of a sprawling area of contemporary machine learning, which I have no ambition to exhaustively overview here — see, e.g., (Higgins et al., 2022) for a review. Rather, let us focus on some specific examples that are particularly relevant to either apparatus-related SMCs or object-related SMCs (see Section 1.1.3 for a description of this distinction), despite the conceptual tension between SMC theory and the representational narrative that motivates these contributions.

Representation learning, reinforcement learning & apparatus-related SMCs

To understand the relevance of representation learning research to apparatus-related SMCs, let me first take a short detour to formulate the latter in a more mathematical language. Here, I propose to interpret this kind of SMCs as "abstract symmetries of the embodied agent's sensorimotor interface": i.e., transformations of the agent's "sensorimotor spaces" which would leave invariant its "sensorimotor interface". Importantly, the transformations defining the symmetries are here not understood as the actions of the embodied agent itself, but are "abstract" mathematical operations on both the agent's "sensory" space and its "action" space. I.e., these abstract transformations are similar to the "intrinsic symmetries" from (Möller et al., 2023), mentioned in Section 1.2.1, or to the "scrambling" of the underwater vessel's controls and sensors by a "vilainous aquatic monster" in the thought experiment from (O'Regan et al., 2001), mentioned in Section 1.1.3 — with the difference that the "monster" would here actually be quite "gentle", as it chooses precisely those transformations that leave invariant the sensorimotor interface. Moreover, by "sensorimotor interface", I mean the way the agent's actions affect its own sensory influx — which can be seen as the agent's "interface" with the external world.

While more general formalisations are possible,[23] let us focus on a simplified, fully-observed setting using a *Markov Decision Process* (MDP). Here, sensor values correspond to the MDP's states $s \in \mathcal{S}$, and the agent's actions to the MDP's actions $a \in \mathcal{A}$. The "sensorimotor interface" is then interpreted as the MDP's *transition channel*,[24] i.e., the conditional probability $\rho(s'|s,a)$ of any next sensor state $s' \in \mathcal{S}$ given any current sensor state $s \in \mathcal{S}$ and any agent's action $a \in \mathcal{A}$. "Symmetries" of this sensorimotor interface ρ can then be interpreted, e.g., as pairs of transformations $\phi : \mathcal{S} \to \mathcal{S}$ and $\psi : \mathcal{S} \times \mathcal{A} \to \mathcal{A}$ such that the

[22] The affordances of an animal's environment are, in short, "what [the environment] offers the animal, what it provides or furnishes, either for good or ill" (Gibson, 2014), and have been argued to be a fundamental building block of perception. Due to our current focus on SMC theory, I will not discuss in further details this central concept of ecological theories of perception, which has permeated many other fields over decades. See, e.g., (Chong et al., 2020; Jamone et al., 2018) for reviews.

[23] E.g., the sensorimotor interface could be the sensory and actuation channels of sensorimotor loop models (Ay et al., 2014; Tishby et al., 2011), or the *ϵ-transducer* defined by the conditional probability of the whole sensory process given the whole action process (Barnett et al., 2015; Marzen, 2025; Rosas et al., 2025).

[24] More conventionally called transition function.

function

$$(\phi, \psi) : \mathcal{S} \times \mathcal{A} \to \mathcal{S} \times \mathcal{A}$$
$$(s, a) \mapsto (\phi(s), \psi(s, a))$$

is invertible, and[25]

$$\rho \circ (\phi, \psi) = \phi \circ \rho, \tag{1.2.1}$$

which can be represented visually as the commutation of the diagram[26]

$$\begin{array}{ccc} \mathcal{S} \times \mathcal{A} & \xrightarrow{\rho} & \mathcal{S} \\ {\scriptstyle (\phi,\psi)}\downarrow & & \downarrow{\scriptstyle \phi} \\ \mathcal{S} \times \mathcal{A} & \xrightarrow{\rho} & \mathcal{S} \end{array} \tag{1.2.2}$$

It can be easily verified that the latter condition is equivalent to[27]

$$\forall s, s' \in \mathcal{S},\ \forall a \in \mathcal{A}, \quad \rho(s'|s, a) = \rho(\phi(s')|\phi(s), \psi(s, a)),$$

i.e., we are here considering the group $\mathcal{G}$ made of pairs (ϕ, ψ) that, indeed, transform the "sensorimotor space" $\mathcal{S} \times \mathcal{A}$ in such a way that the "sensorimotor interface" ρ is left unchanged. The idea that I am proposing, then, is that the structure of the distinct sensory modalities (i.e., their corresponding apparatus-related SMCs) would be reflected in the structure of this abstract symmetry group $\mathcal{G}$ of the sensorimotor interface ρ.

It turns out that the kind of symmetries described in equation (1.2.1) has been investigated in the reinforcement learning literature (van der Pol et al., 2020; Wang et al., 2022a), with a generalisation to partially observed environments in (Nguyen et al., 2023), and a variation explicitly aimed at robots' morphological symmetries in (Apraez et al., 2025). For instance, it has been shown in (van der Pol et al., 2020) that if a group of pairs (ϕ, ψ) satisfying (1.2.1) also leaves the MDP's reward function invariant, then one can design a corresponding *quotient MDP*, in which one can perform learning and planning and then *lift* the resulting policy to the original MDP. This result is relevant to our aims because, from an information parsimony perspective, the operation of quotienting the MDP w.r.t. the symmetries in (1.2.1) can be seen as a "compression preserving the sensorimotor interface". This suggests that *optimal* compressions "preserving the sensorimotor interface" could be very helpful to understand SMCs — as they would, in a sense, capture their "platonic core". This intuition is the main sensorimotor motivation for the work presented in Chapter 2 — even though, there, the symmetries that we will consider will have a simpler, more generic form.

Let me, however, clarify the use of the reference to Plato in this intuition. The philosopher postulates a world of "Ideas", also called "Forms", that would be ontologically primary abstractions that are aspatial, atemporal and *a fortiori* agent-independent, of which our imperfect, worldly objects would be mere shadows (Plato, 1943, 1952). The "platonic core" that I referred to above is thus, actually, the *exact converse* of Plato's "Forms": while the optimal compression would provide a certain kind of abstraction, it would be one that *emerges from the agent's sensorimotor interface* under appropriate informational trade-offs. I.e., here it is

[25] In (1.2.1), the symbol $\circ$ denotes composition of channels, and ϕ, (ϕ, ψ) are seen as deterministic channels.

[26] For simplicity, we use the same notation as would be used for deterministic functions. But formally, in (1.2.2), vertices should be probability simplices and edges push-forwards of channels (see Definition 3.2.2).

[27] See, e.g., Lemma 15 in (Charvin et al., 2023b).

the agent's sensorimotor behaviour — and the constraints regulating it — that are ontologically primary. The abstract "Forms" then emerge from the "shadows" that constitute this sensorimotor experience, under information parsimony constraints that induce the agent to *make sense* of this experience (see the Main Intuition in Section 1.1.1). In particular, such "Forms", rather than shaping our wordly objects, are *shaped by* the coarse-graining of spatially extended, dynamical and agent-dependent patterns (see Section 1.1.3).

Representation learning & object-related SMCs

As mentioned above, the symmetries considered in equation (1.2.1) are "abstract", in the sense that the corresponding group transformations do not model an actual action by an actual agent. This "abstract" point of view on symmetries, inspired by the use of symmetries in physics, is shared by many contributions to representation learning research: e.g., (Higgins et al., 2018) argues that for internal representations to "disentangle" distinct dimensions of the natural world's features, the symmetries of the former should mirror those of the latter. However, crucially, the symmetries structuring these disentangled representations were interpreted, in (Caselles-Dupré et al., 2019), as generated by the *interaction* of an agent with its environment: i.e., the action of the "symmetry" group is here seen as a model of the *agent's own action*. This interpretation has since then been fueling a number of contributions to the field of representation learning (Dean et al., 2025; Higgins et al., 2022; Keller et al., 2026; Keurti et al., 2024; Pérez Rey et al., 2023; Quessard et al., 2020), sometimes with explicit references to sensorimotor theories of perception (Keurti et al., 2023) or even straightforward continuations (Caselles-Dupré et al., 2021a,b), with the modern tools of machine learning, of the line of work on sensory compensability initiated in (Philipona et al., 2003).

Along these lines, particularly important to us will be the *class-pose decomposition* framework (Marchetti et al., 2023; Oizumi et al., 2025; Pérez Rey et al., 2023; Winter et al., 2022), which investigates the decomposition of a space into two coordinates adapted to a given group action. More precisely, consider the action ρ of a group $\mathcal{G}$ on a state-space $\mathcal{X}$, and denote by $\kappa : \mathcal{X} \to \mathcal{C}$ its projection on orbits, where $\mathcal{C}$ denotes the space of orbits (see Definition A.0.3) and is here called the "class" space. Consider another space $\mathcal{P}$, called the "pose" space, and for any action ξ of the group $\mathcal{G}$ on $\mathcal{P}$, denote by $\mathtt{Id}_{\mathcal{C}} \otimes \xi$ the action of $\mathcal{G}$ on the full class-pose space $\mathcal{C} \times \mathcal{P}$ that leaves $\mathcal{C}$ invariant and applies ξ on $\mathcal{P}$: i.e., $(\mathtt{Id}_{\mathcal{C}} \otimes \xi)_g(c, p) := (c, \xi_g(p))$ for all $g \in \mathcal{G}$, $c \in \mathcal{C}, p \in \mathcal{P}$. The class-pose decomposition literature then usually aims at learning the projection on orbits κ, together with a map θ from $\mathcal{X}$ to the pose space $\mathcal{P}$, and an action ξ of $\mathcal{G}$ on $\mathcal{P}$, such that (κ, θ) defines an isomorphism between ρ and $\mathtt{Id}_{\mathcal{C}} \otimes \xi$: i.e., such that (κ, θ) is bijective and for all $g \in \mathcal{G}$, the following diagram commutes:

$$\begin{array}{ccc} \mathcal{X} & \xrightarrow{\rho_g} & \mathcal{X} \\ {\scriptstyle (\kappa,\theta)}\downarrow & & \downarrow{\scriptstyle (\kappa,\theta)} \\ \mathcal{C}\times\mathcal{P} & \xrightarrow{(\mathtt{Id}_{\mathcal{C}}\otimes\xi)_g} & \mathcal{C}\times\mathcal{P} \end{array} \tag{1.2.3}$$

where we defined $(\kappa, \theta)(x) := (\kappa(x), \theta(x))$ for all $x \in \mathcal{X}$. If this is possible, we thus obtain a decomposition of the action ρ into an invariant "class" coordinate $\mathcal{C}$ and an equivariant "pose" coordinate $\mathcal{P}$. Let us take a paradigmatic example of this line of work (we will see in Chapter 3 that it turns out to be problematic in full generality, even though this is not necessarily a problem in the data-sets used in previous work). Assume that the group action ρ consists of the rigid transformations (i.e., compositions of rotations and translations) of rigid 3D objects (formalised, e.g., as closed surfaces in $\mathbb{R}^3$). In this case, the "class", i.e., orbit of a given surface, is *the family of all surfaces into which it can be transformed through rigid transformations*. This family of surfaces is here seen as a formalisation of the "shape" of

the given surface. On the other hand, the "pose" coordinate should correspond to, literally, the pose in space of surfaces, i.e., their position and orientation — seen as a single, abstract coordinate that simultaneously fixes, *for any given class/shape*, a surface in it with a specific position and orientation. Interestingly, this example aligns closely with the example of shapes mentioned in the foundational paper of SMC theory (O'Regan et al., 2001) (see fourth quote in Section 1.1.3). More broadly, the decomposition into invariant and equivariant coordinates resonates strongly with the notion of capturing the "(in)variance properties" of SMCs (see second quote in Section 1.1.3). This suggests that the mathematical object underlying the class-pose decomposition framework is directly relevant to the formalisation of object-related SMCs (see Section 1.1.3 again). Note that this relevance to SMCs holds despite the class-pose decomposition literature being steeped into a strongly representationalist narrative: i.e., in this literature, successfully learned class and pose variables are usually referred to as "representations" of the learning system. This contrasts with SMC theory's stance[28] that the study of perception can and should dispense with the notion of representation, as perception is seen as a fundamentally ongoing activity — which might rely on a specific "attunement" of brain dynamics acquired along learning and development, but where these brain dynamics are not seen as "representing", i.e., in particular, "referring to" external phenomena defined independently from the agent (see Section 1.1.3). We will come back below on this conceptual tension.

A related result (Tsao et al., 2022), groundbreaking for sensorimotor perception research, has successfully formalised and computationally implemented the idea that visual objects themselves can be defined purely from what is known by ecological psychologists as the *ambient optic array* (Gibson, 2014). This term refers to the set of all light rays in a given environment, in all directions and from all possible observation points. More precisely, (Tsao et al., 2022) shows that this structure of light rays is enough to identify contiguous surfaces in 3D space, and track them while they are moving and occluding one another. Importantly for us, contiguous surfaces are here captured as classes of an equivalence relation, on the set of light rays, defined by the action of a *pseudogroup* formalising the intuition of change of perspective.[29] These equivalence classes, in short, identify a given contiguous surface to the family of all light ray cones that "look towards" this surface — and only this surface. This mathematical object is very similar to the "class" coordinate from the class-pose decomposition framework.[30]

Towards a confluence of algebraic, dynamical & informational approaches?

To clarify both the convergences and tensions between SMC theory and the representation learning research reviewed in the previous section, let me insist on the following fact: the "representations" of the latter line of work are now *defined by the way they are transformed by a group action*, where the latter is often interpreted as the agent's own actions. In a sense, this "action-structured" perspective on representations opens a breach in the notion of representation itself. Indded, it is worth pointing out here that many contemporary textbooks in ergodic theory define a *dynamical system as a group action*, where the acting group is most commonly made of time translations, but can be very different as well — see, e.g., (Glasner, 2003; Kerr et al., 2016). This suggests that if the group action models the agent's action on its

[28] Or, at least, that of the enactivist strand of SMC theory — the criticisms of (O'Regan et al., 2001) are mainly aimed at *picture-like* representations (see Sections 1.1.3 and 1.2.2).

[29] Pseudogroups are a variation of group actions adapted to *local* transformations on a topological space.

[30] The "ambient optic array"is understood by ecological psychologists as a structure *in the environment*: it can be probed by agents equipped with a visual apparatus, but is not in itself part of the sensorimotor interface (Gibson, 2014). But the result from (Tsao et al., 2022) suggests that the ambient optic array might induce a structure, *in real-world agents' visual sensorimotor interface*, that would also be sufficient to identify and track contiguous surfaces.

sensory surface (an admittedly drastically simple model), the formalism describing "group-structured representations" can be understood as describing the dynamical structure induced by an agent's actions on its sensory influx. In particular, the agent's internal activity would not need to "encode" external "objects" arbitrarily defined by an experimentalist (Brette, 2019) or the designer of a machine learning benchmark. Rather, defining an "object" or "class" as a group action's orbit would mean defining it from a purely intrinsic perpsective: i.e., as an invariant of the agent's sensorimotor dynamics.

It turns out that a recent line of work on "flow equivariance" (Keller et al., 2026) goes, to certain extent, in this direction — even directly citing one of the main initiators of the concept of affordances (Gibson, 2014), which is similar to that of object-related SMC:

> If we, like J. J. Gibson, consider an ecological approach to vision, the natural transformations of sensory inputs over time are caused by movement, e.g. by self-motion or by external motion of the world. An important aspect of the world, irrespective of our senses, is that it does not tend to change very quickly (a founding principle of Slow Feature Analysis). For example, if a child is playing in the garden now, it is very likely she is still playing in the garden a second later, despite the fact the input to my senses might have changed dramatically. We can therefore interpret fast external changes, and also changes caused by our body movements, as approximate symmetries of the true world state, and we may want to build representations which are structured with respect to these generalized symmetry transformations. (Keller et al., 2026)

I.e., the group action on the sensory space is here interpreted as the outcome of physical movement (of the environment or the agent) that leaves invariant underlying features of interest — in this case, whatever is left invariant by fast enough movement, as it is considered a reliable proxy on "true world states".[31] The group action on the latent, i.e., "internal" space is then defined, in (Keller et al., 2026), as the Lie group action induced by the time evolution of *spatiotemporal* dynamics inspired from the study of *traveling waves* in the brain (Muller et al., 2012, 2018; Sato et al., 2012). This kind of dynamical structure is argued to be characteristic of spatially extended dynamical systems with non-negligible time delays that increase with spatial distance. In the brain, such time delays are due, in particular, to a pressure towards short-range connections, as well as axonal and dendritic conductance delays (Chklovskii et al., 2002). Interestingly for us, both these factors are argued in (Keller et al., 2026) to be, at least in part, the consequence of metabolic efficiency requirements. I.e., the structure of traveling wave-like dynamics would be *induced by (metabolic) parsimony constraints*.

Now, adopting a *predictive processing* (Clark, 2016) point of view suggests that these internal spatiotemporal dynamics should be predictive of the neural system's, movement-induced sensory influx. The "flow equivariance" put forward in (Keller et al., 2026) is then the commutation relation between these sensory and internal dynamics that is required for the latter to be predictive of the former. More precisely, in the language of ergodicists (Glasner, 2003), "flow equivariance" means that the group action defined by internal dynamics should be a *factor* of that defined by sensory dynamics (see Appendix A for a formal definition).

Crucially, the "representations" considered here are thus not picture-like, but consist in spatiotemporal patterns that, when flow equivariant, allow the neural system to "track" its input, which in natural conditions is itself made of spatiotemporal dynamics (see Section 1.1.3). In this sense, this line of work provides significant computational innovations to dynamical, and specifically *movement-based* alternatives to the dominant "filter" paradigm in machine

[31] A similar "Slow Feature Analysis" assumption was also made in (Laflaquière et al., 2015a), which models object discovery in naive agents from a sensorimotor perspective.

learning, which typically frames perception as the passive processing of static sensory inputs.[32] As they are based on biologically plausible dynamics, these artificial systems could in return help building models for corresponding alternatives in neuroscience and adaptive behaviour.[33] In particular, traveling waves-based models could operationalise the self-sustained, internal dynamics that are central to the inside-out account of the brain (Buzsáki et al., 2019), where, for well-chosen objectives, the training of the flow equivariant neural system could be interpreted as the *calibration* of these internal dynamics through sensorimotor interaction with the environment (see Section 1.1.3). Moreover, the flow equivariance property seems to be relevant to a predictive processing interpretation (Seth, 2014) of SMC theory's notion of "attunement" (O'Regan et al., 2001) — see Section 1.1.3 again.

However, the flow equivariance framework retains fundamental limitations from the point of view of operationalising sensorimotor theories of perceptions, especially in their enactivist interpretations (Barandiaran, 2017; Buhrmann et al., 2013; Degenaar et al., 2017):

- Here, the neural system's spatiotemporal patterns are thought of as implementing the symmetries of "true world states", thus allowing the neural system to "encode" stimuli whose meaning is defined, somehow arbitrarily, by the experimentalist (Brette, 2019) or the designer of the machine learning benchmark: e.g., (sequences of) digit(s) (Keller et al., 2023a,b), geometric shapes (Jacobs et al., 2025; Liboni et al., 2025), moving objects (Lillemark et al., 2025), etc. The veridical prediction of these pre-defined stimuli might in some cases be an accurate proxy on more intrinsic notions of behavioural relevancy. But this approach still remains silent on how perceptual structure emerges from the embodied agent's "own problems" (Egbert et al., 2022), including the tension between behavioural relevance and information parsimony constraints (see the Main Intuition at the beginning of Section 1.1).[34]

- While it suggests a fundamentally dynamical perspective on cognition, this line of work has not, to this day and to the best of my knowledge, studied the impact of *closed-loop* behavioural dynamics on perceptual dynamics (Ahissar et al., 2016; Buhrmann et al., 2013).

- By focusing on the symmetries of "true world states", it understands brain dynamics as *internalising the structure of the external environment*, while SMC theory understands them as *attuning to the structure of interaction of the agent with its environment*, in a way that supports the "skillful mastery" of this interaction (see Section 1.1.3).

These conceptual tensions with sensorimotor approaches to perception are not specific to flow equivariance, but are characteristic of the representationalist framework that imbues (unsurprisingly) the whole representation learning community. While the latter's notion of representation evolved along the last decade of work on symmetry-structured representations, it is thus still a long way from being aligned with a non-representational, SMC-based notion of percept. But the work overviewed in this Section 1.2 does show a paradoxical and partial

[32] Recurrent neural networks (RNNs) implementing traveling wave-like dynamics addressed classification (Keller et al., 2023a), short-term prediction (Benigno et al., 2023), image segmentation (Jacobs et al., 2025; Liboni et al., 2025) or working memory (Keller et al., 2023b). Novel RNN architectures explicitly incorporating flow equivariance yielded prediction of moving stimuli (Keller, 2025), and partially observed dynamical world modeling (Lillemark et al., 2025).

[33] Note that these task-trainable models are closely related to long-running traditions in computational and mathematical neuroscience (Keller et al., 2026; Muller et al., 2018), including neural field theory (Coombes et al., 2014) and neurogeometry (Baspinar et al., 2021; Mazzetti et al., 2026; Petitot, 2017; Sarti et al., 2008).

[34] As mentioned above, (Keller et al., 2026) stresses that in the brain, traveling waves are induced by connectivity constraints which themselves might result, at least in part, from metabolic efficiency constraints. But the connectivity constraints are hard-wired in the computational models, which does not provide a formalisation of how perceptual structure would be induced by parsimony constraints themselves.

alignement between the mathematical objects considered by these two traditions, which is significant enough to suggest that we are witnessing a **convergence of conflicting scientific narratives on increasingly similar mathematical objects**. Crucially, these formal similarities do *not* take away the conceptual tension, which can be fruitful for pursuing further the formalisation of SMC theory. Indeed, on the one hand, the representation learning community provides numerous "experimental" numerical results that exemplify how some specific invariance- and equivariance-based formal structures can be relevant to operationalise perception. On the other hand, the enactive and dynamical aspects of SMC theory highlight both the limitations of the representational narrative, and how these limitations are reflected in the choice of the mathematical objects operationalising this narrative: group symmetries, i.e., specific commutation relations defined by group actions, among all the possible commutation relations defined by all possible kinds of transformations. This indicates directions in which to generalise the group-theoretic objects, so as to better capture the principles of sensorimotor perception.

Even though, again, our focus will here be mostly formal, with often no explicit reference to embodied agents' sensorimotor interface, a large section of this thesis (i.e., Chapters 2 and 3) sits within the long-term perspective of integrating algebraic and dynamical approaches to sensorimotor perception. To this emerging confluence, however, we aim to divert a third stream: that of information-theoretic trade-offs involving information parsimony — and through it information-theoretic models of embodied agents' closed-loop behaviour (Ay, 2015; Langer et al., 2024; Salge et al., 2014; Tishby et al., 2011). This effort has three motivations:

- Identifying a given structure in the sensorimotor loop (dynamical or otherwise) is not sufficient to explain *why* this structure would be relevant from the agent's intrinsic perspective. I.e., if we take seriously the claim that perception should emerge from the agent's "own problems" (Egbert et al., 2014), then we should pin down the constraints that would *induce* a given dynamical structure to be integrated into the perceptual process. As mentioned in Section 1.2.2, a first step in this direction can be to identify formal dualities between structure and information parsimony.

- From the latter perspective, the perceptually relevant "structures" correspond to solutions of multi-objective optimisation problems whose target functions model the corresponding "constraints" that are traded-off against each other (see Section 1.1.1). Identifying these optimisation problems is thus also a first step in the design of methods for data-based discovery of these mathematical structures. Such methods are necessary if we want to use this theoretical framework to actually "probe" reality with simulations or experiments.

- As demonstrated by decades of research in a constantly evolving field, information theory provides powerful tools to analyse the structure and dynamics of complex systems (Ay et al., 2022; Coudène, 2016; Lindgren, 2024; Mediano et al., 2022a,b; Rosas et al., 2024). In particular, Information Bottleneck-related formalisms are a promising tool to capture *soft* structure in the sensorimotor loop, i.e., structure that only becomes apparent under appropriate *coarse-grainings* — and at appropriate "granularities" (see Section 1.1.2).

The last point above motivates the method that we will use the most in thesis, as explained in the next section.

1.3 This thesis' contributions

1.3.1 Our general method

In a different context than ours, some neuroscientists have argued that the study of natural behaviour — which challenges both the usual controlled laboratory settings and the theories that the latter rely on to interpret experimental data — requires a *gradual* approach of building the "new" by simultaneously relying on the "old" and selectively undoing it:

> While studies of constrained scenarios have provided an important conceptual foundation for experimental and theoretical approaches for studying the brain, it is likely that many of the resulting concepts will fail to generalize to explain the richness of natural interactive behavior. Along the way we will need to abandon some deeply held assumptions, perhaps even some foundational concepts. However, at the same time, we must not simply discard all that we have learned from classical constrained laboratory studies. Here, we advocate [...] identifying key theoretical assumptions associated with more constrained conditions, modifying established experimental paradigms in a way that gradually relaxes them, and then explicitly testing whether our current theory holds up in the new situation or whether it must be revised. (Cisek et al., 2024)

The method that we will follow in this thesis can be seen as a purely formal counterpart of this stepwise strategy. In this analogy, the "constrained laboratory studies" are replaced by previous group-theoretic approaches to perception and learning in biological and artificial agents, and their modification "in a way that gradually relaxes them" by the generalisation of these group-theoretic tools in an information-theoretic and/or dynamical language — while the aim is, on the long term, to design formalisms that can capture the same "richness of natural interactive behavior". Let me unpack these ideas.

Even though the study of group-theoretic symmetries in neural and embodied systems has already yielded deep results and might still deliver many more (see Section 1.2), it appears increasingly clear that this theoretical framework is too rigid for the kind of "structure" encountered in adaptive behaviour — including those relevant to sensorimotor perception. In particular, if the structure of a group action is used to model the agent's own actions, then it ultimately needs drastic generalisations: e.g., we should be able to deal with closed-loop, non-invertible and stochastic actions, and where there is no "identity action" available to the agent. Moreover, even though group actions can be a powerful point of view for the study of dynamical systems (see Section 1.2.3), in many cases the unfolding along time of the behaviour's dynamics should not be abstracted away — i.e., the *arrow of time* should be explicitly part of the formalism. Limitations arise as well if the group structure rather describes abstract symmetries: e.g., whatever structure is to be found in the sensorimotor interface of embodied agents, it is unlikely to correspond to *exact* symmetries — the best we can hope is for the group-theoretic formalism to *approximately* capture this structure.

However, at core, the use of group theory fosters the development of an algebraic perspective on the study of adaptive behaviour, which given the work overviewed in Section 1.2, seems to be a promising approach. In particular, this previous work has already established the central role of *invariance* and *equivariance* in learning systems in general, and it suggests that these notions are also crucial to sensorimotor perception. In this thesis, as far as the group-theoretic formalism is concerned, we thus focus almost exclusively on (different versions of) invariance and equivariance.

Our approach is, more precisely, to use these notions of symmetry as, so to say, a "conceptual scaffold": i.e., we selectively discard parts of the formalism and selectively preserve

others, thus leveraging the clarity offered by group theory while obtaining more flexible mathematical objects that have the potential to capture a broader range of the "richness of natural interactive behaviour" (Cisek et al., 2024). In contrast with recent generalisations of group-theoretic symmetries, in particular of equivariances (Ashman et al., 2024; Romero et al., 2022; Song et al., 2023; Wang et al., 2022b), ours build upon *characterising symmetry in terms of information parsimony* — where "information parsimony" is here formalised using the IB framework, and novel variations or generalisations of it. The central mathematical object for this characterisation is the *partition in orbits* (see Definition A.0.3 in Appendix A). Indeed, given a group action on some state-space, it can be easily verified that the corresponding partition in orbits is the finest partition of the state-space into invariant subsets (see Proposition A.0.4 in Appendix A for a formal statement). This suggests that the projection on orbits is, in some sense, an *optimal compression of the state-space under the constraint of preserving the information left invariant by the group action*. **This intuition is, at the formal level, one of the main driving forces of this thesis**. It underlies the work presented in Chapter 2, and will be explicitly formalised (in the finite case) in Chapter 3.

More generally, the characterisations of symmetry that we aim for are not only in terms of information parsimony, but more precisely in terms of *trade-offs* of information parsimony with other well-chosen information quantities. Crucially, once a group symmetry is characterised by such an information-theoretic trade-off, it can then naturally be "softened" by varying the value of the corresponding trade-off parameter(s). The "softness" of the resulting symmetry is thus, in particular, parametrised by the amount of compression that it induces (where "softer" symmetries correspond to more compression).

In certain cases, the information-theoretic characterisation of group-theoretic concepts will first require a generalisation of the latter in the language of ergodic theory and Markov Decsion Processes (see Chapter 3). Whether or not this is required, however, the procedure that I described above is the binding feature of the approach developed in this thesis.

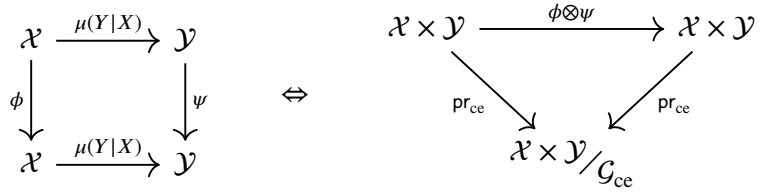
1.3.2 Overview of results

Chapter 2: Information Parsimony and Symmetries of Stochastic Channels

Here, we start from the intuition that the classic IB method extracts invariances (see Section 1.1.2). We show that this intuition can indeed be formalised (in the finite case) in the language of group theory: for maximal trade-off parameter λ, the compression channels that solve the IB problem (1.1.1) coincide, essentially, with the projection on orbits of the group of *invariances* of the channel $\mu(Y|X)$ from source X to relevancy Y. These invariances are defined as transformations ϕ of the source space $\mathcal{X}$ such that composing $\mu(Y|X)$ at the input by ϕ does not change it, i.e., in symbols: $\mu(Y|X)\circ\phi = \mu(Y|X)$. As the projection on orbits characterises the invariance group of $\mu(Y|X)$, this provides an information-theoretic characterisation of channel invariances. While given the previous literature on the IB, deep learning and invariances (see Section 1.1.2), this result is not surprising, it opens a new way to investigating generalisations of this phenomenon to more general symmetries.

In particular, we then turn our attention to the group of *equivariances* $\mathcal{G}_{\text{ce}}$ of a channel $\mu(Y|X)$, defined as pairs of transformations (ϕ, ψ), of resp. the channel's input space $\mathcal{X}$ and output space $\mathcal{Y}$, such that composing $\mu(Y|X)$ by ϕ at the input is the same thing as composing it by ψ at the output: i.e., $\mu(Y|X)\circ\phi = \psi\circ\mu(Y|X)$. Similarly as for channel invariances, the projection on orbits pr_{ce} from $\mathcal{X}\times\mathcal{Y}$ to the space of orbits $\mathcal{X}\times\mathcal{Y}/\mathcal{G}_{\text{ce}}$ here characterises the group

of channel equivariances, through the following equivalence of commutative diagrams:



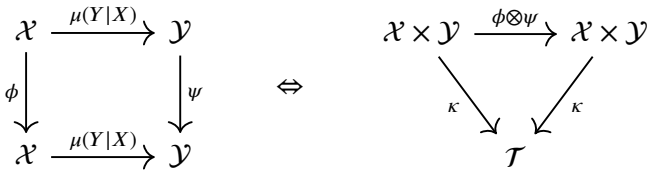
(1.3.1)

which means that the commutation of the left-hand-side diagram — the defining property of equivariances — is equivalent to the commutation of the right-hand-side diagram, where $\phi \otimes \psi$ is defined by $(\phi \otimes \psi)(x, y) := (\phi(x), \psi(y))$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Our aim is then to design an IB-like problem such that the solutions for the maximal trade-off parameter λ coincide with the projection on orbits, or at least mimic its characterisation (1.3.1) of equivariances. In contrast with the case of invariances, though, the projection on orbits here processes *jointly* the channel's input space $\mathcal{X}$ and its output space $\mathcal{Y}$. The compression channel κ of our new IB-like problem should thus also take the whole product space $\mathcal{X} \times \mathcal{Y}$ as input, and the question becomes: which information-theoretic quantity should be preserved by such a κ for the solutions at the maximal trade-off parameter λ to characterise the equivariances of $\mu(Y|X)$? We show (still in the finite case) that the relevant quantity is the Kullback-Leibler divergence

$$D(\kappa \cdot \mu || \kappa \cdot \mathcal{E}_{\mathrm{ce}}) := \inf_{\nu \in \mathcal{E}_{\mathrm{ce}}} D\big(\kappa \cdot \mu || \kappa \cdot \nu\big) \tag{1.3.2}$$

between the projection $\kappa \cdot \mu$ of the data distribution $\mu := \mu(X, Y)$ on the latent space, and the corresponding projection $\kappa \cdot \mathcal{E}_{\mathrm{ce}}$ of a well-chosen exponential family $\mathcal{E}_{\mathrm{ce}}$ (technically, these "projections" are defined as *push-forwards*, see the introduction of Chapter 2). By trading-off the preservation of the quantity above with the compression of (X, Y), we obtain a novel variation of the IB problem formalising the intuition of an *optimal compression preserving the information carried by the channel* $\mu(Y|X)$ *about the system* (X, Y). Crucially, this object yields an information-theoretic characterisation of equivariances: given a full support distribution on $\mathcal{X} \times \mathcal{Y}$, and a channel κ implenting an optimal compression of (X, Y) under the constraint of preserving $D(\kappa \cdot \mu || \kappa \cdot \mathcal{E}_{\mathrm{ce}})$ as much as possible, a pair (ϕ, ψ) is an equivariance of $\mu(Y|X)$ if and only if we have $\kappa \circ (\phi \otimes \psi) = \kappa$. This characterisation consists of the same equivalence of commutative diagrams as in (1.3.1), but with the quotient space $\mathcal{X} \times \mathcal{Y} / \mathcal{G}_{\mathrm{ce}}$ replaced by the bottleneck space $\mathcal{T}$ and the projection on orbits $\mathsf{pr}_{\mathrm{ce}}$ replaced by the bottleneck channel κ:



(1.3.3)

These information parsimony reformulations of channel invariances and equivariances yield natural definitions of resp. *soft invariances* and *soft equivariances*, whose "softness" is parametrised by the compression-information preservation trade-off: i.e., softer symmetries correspond to more compression and less preserved mutual information $I(T; Y)$, resp. less preserved divergence $D(\kappa \cdot \mu || \kappa \cdot \mathcal{E}_{\mathrm{ce}})$.

Importantly, our novel variation of the IB framework capturing equivariances can be generalised to *any* exponential family $\mathcal{E}$ on a finite alphabet $\mathcal{A}$, yielding what we call the *Divergence IB* problem, defined as

$$\mathrm{DIB}(\lambda) := \underset{\substack{\kappa \in \mathscr{K}_{\mathsf{shape}} \\ D(\kappa\cdot\mu||\kappa\cdot\mathcal{E})\geq\lambda}}{\arg\min} \; I_{\kappa}(A;T). \tag{1.3.4}$$

Here, $D(\kappa \cdot \mu||\kappa \cdot \mathcal{E})$ is defined similarly as in (1.3.2), and $\mathscr{K}_{\mathsf{shape}}$ is a set of compression channels from $\mathcal{A}$ to a bottleneck space $\mathcal{T}$, with potentially additional constraints on the channels' shape. In particular, if $\mathcal{E}$ is a *hierarchical model* (Ay et al., 2017), the corresponding solutions to (1.3.4) can be seen as optimal compressions preserving a specific set of stochastic interdependencies in $\mathcal{A}$ (Ay, 2015). For $\mathscr{K}_{\mathsf{shape}}$ coinciding with the set of all channels from $\mathcal{A}$ to $\mathcal{T}$, we derive a Blahut-Arimoto algorithm to approximate the solutions to (1.3.4), and use it to explore numerically, in a simple synthetic experiment, the soft equivariances defined above.

Overall, our novel Divergence IB is a flexible framework that opens a promising new avenue for the study, and automatic discovery, of "soft symmetries" through the lens of information parsimony. In particular, it could be applied to equivariances in Markov Decision Processes, e.g., of the kind considered in equation (1.2.1), or more general models of the sensorimotor loop (Rosas et al., 2025; Tishby et al., 2011). This could yield the discovery of abstract symmetries in embodied agents' sensorimotor interface, thus making the framework directly relevant to sensorimotor perception (as explained at length in Section 1.2).

However, achieving these aims would first require solving fundamental limitations: first, despite characterising equivariances through (1.3.3), compression channels κ for maximal trade-off parameter λ do not coincide with the projection on orbits $\mathsf{pr}_{\mathsf{ce}}$: they actually define a coarser partition. Second, our framework yields the "discovery" only of the compression channel κ, not of equivariances (ϕ, ψ) themselves. To address these issues, we identify the need to design a broader information-theoretic constrained optimisation problem, where we would optimise over both a compression channel κ *and* a channel ρ parametrising transformations of the data space (see Sections 2.5.2 and 2.5.3 of Chapter 2). The main stumbling block then becomes to characterise information-theoretically, and generalise to a stochastic setting, what it means for a compression channel κ to be the "projection on orbits" of a given family of transformations ρ. This technical question is addressed in Chapter 3, even though the conceptual focus of the latter chapter is quite different.

Publication Information. *Chapter 2 is an extended version of (Charvin et al., 2025). A preliminary version of this work was published in (Charvin et al., 2023b).*

Chapter 3: Minimal class-pose parametrisation in Markov Decision Processes

This chapter is the most directly relevant to sensorimotor perception, even though it is also the most mathematical, relying to a large extent on a measure-theoretic setting that constrasts with the rest of this thesis. Here, we turn to the class-pose decomposition framework, which indicates an interesting direction for operationalising the notion of object-related SMCs from (O'Regan et al., 2001) (see Section 1.2.3). Our aim is to make the framework more relevant to adaptive behaviour modeling, in particular sensorimotor theories of perception, by extending it in three directions: algebraic, dynamical, and information-theoretic.

To describe the algebraic aspect, let us first reformulate the object defined in Section 1.2.3. Denote the orbits under the group action ρ by $(\mathcal{X}^c)_{c\in C}$ and the corresponding restrictions of ρ by $(\rho^c)_{c\in C}$. Then a class-pose decomposition is equivalent to a group action ξ on the pose

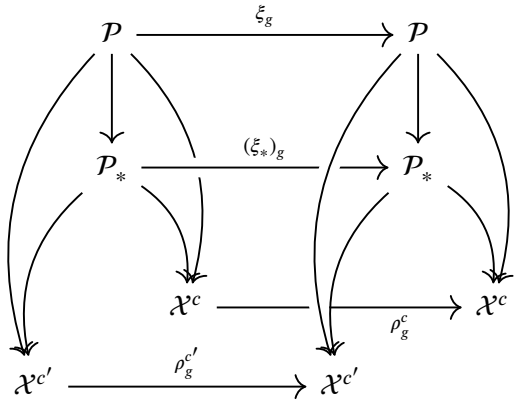
space $\mathcal{P}$ that is isomorphic to each restriction ρ^c: i.e., for all $c \in \mathcal{C}$, we must have a bijective $\theta^c : \mathcal{X}^c \to \mathcal{P}$ such that for all $g \in \mathcal{G}$, the following diagram commutes:

$$\begin{array}{ccc} \mathcal{X}^c & \xrightarrow{\rho_g^c} & \mathcal{X}^c \\ {\scriptstyle\theta^c}\downarrow & & \downarrow{\scriptstyle\theta^c} \\ \mathcal{P} & \xrightarrow{\xi_g} & \mathcal{P} \end{array} \tag{1.3.5}$$

We start from the observation that this is only possible under the assumption that all the restrictions to orbits ρ^c are isomorphic — i.e., informally, that the group action is exactly the same on every orbit. This requirement happens to rule out the supposedly paradigmatic example of rigid transformations of rigid objects, and even drastically simple examples in the finite case. We therefore propose a similar, but more general structure that does not require this restrictive assumption. Namely, we propose to "reverse the direction of the arrows and break the bijectivity" in the diagram (1.3.5), by rather searching for a *factor* from $\mathcal{P}$ to each $\mathcal{X}^c$: i.e., a group action ξ on $\mathcal{P}$ such that for all $c \in \mathcal{C}$, there exists a surjective (but not necessarily injective) "projection" map $\phi^c : \mathcal{P} \to \mathcal{X}^c$ such that for all $g \in \mathcal{G}$, the following diagram commutes:

$$\begin{array}{ccc} \mathcal{P} & \xrightarrow{\xi_g} & \mathcal{P} \\ {\scriptstyle\phi^c}\downarrow & & \downarrow{\scriptstyle\phi^c} \\ \mathcal{X}^c & \xrightarrow{\rho_g^c} & \mathcal{X}^c \end{array} \tag{1.3.6}$$

An action ξ on $\mathcal{P}$ satisfying diagram (1.3.6) for all $g \in \mathcal{G}$ and all $c \in \mathcal{C}$ is called a *(set-theoretic) joining* of the orbits — a terminology inspired from ergodic theory's measure-theoretic notion of joining (de la Rue, 2006; Glasner, 2003). However, crucially, we require the joining to satisfy a certain notion of *minimality*: i.e., in short, we require it to be "as isomorphic as possible". For instance, a *minimal joining* of group actions ρ^c on $\mathcal{X}^c$ and $\rho^{c'}$ on $\mathcal{X}^{c'}$ is a joining ξ_* on a pose space $\mathcal{P}_*$ such that for any other joining ξ on a pose space $\mathcal{P}$, the following diagram commutes for all $g \in \mathcal{G}$:



where for simplicity, I ommited the labels of the "projection" maps. *Minimal joinings* of all the group action's orbits then yield what we call *minimal class-pose parametrisations* of the group action. This framework generalises arithmetics' notion of *least common multiple*: here, numbers are replaced by group actions and their (least) common multiples by their (minimal)

joinings. At the conceptual level, this new mathematical object captures the "simplest" description of how *any* class (i.e., orbit) is transformed by the group action. In particular, in a certain sense, this "simplicity" constraint induces poses to capture the "common structure of changes" accross all classes, thus providing one specific formal interpretation of SMC theory's notion of "structure of changes" (see Section 1.1.3).

However, as mentioned in Section 1.3.1, the framework of group actions is a limitation in itself when it comes to modeling the agent's own actions. To make the minimal class-pose parametrisation framework outlined above more relevant to embodied agents, it is thus crucial to generalise it to more flexible mathematical objects — in particular, ones that account for non-invertible and stochastic actions, as well as the time-evolution of the agent's potentially closed-loop behaviour.[35] Moreover, we would like, as in Chapter 2, to understand which kind of information-theoretic trade-offs might characterise this novel invariant-equivariant structure.

To address these requirements, we replace group actions by a pair made of a *policy* π that stochastically determines the action given a state, and a *transition channel* ρ that stochastically determines the next state if a given action is applied from a given state. We call such a pair a *Markov Decision Process (MDP)*, even though, in contrast with MDPs usually considered in the literature, here there is no reward function and the policy is fixed — this is because we are interested in understanding the structure of the MDP's dynamics rather than the learning of optimal policies. In particular, focusing on fixed policies is crucial to capture the specific structures that *emerge through the agent's (potentially closed-loop) behaviour*. Moreover, we only assume the MDP's state-space $\mathcal{X}$ and action space $\mathcal{G}$ to be *standard Borel spaces* — a general assumption that encompasses, e.g., finite, countably infinite, or Euclidan spaces, as well as differentiable manifolds and many infinite-dimensional spaces.

To generalise the group-theoretic version of minimal class-pose parametrisation, we then proceed in two steps: we first generalise classes, before building upon the latter to generalise poses. Classes, which were previously defined as a group action's partition in orbits, are generalised into a *decomposition into ergodic components* of a standard Borel MDP. Here, we use a previously obtained decomposition into ergodic components of standard Borel Markov chains (Worm et al., 2011), in the following way. The average $\bar{\rho}$ of the MDP's transition channel ρ over actions sampled through the policy π defines a Markov chain on the state-space $\mathcal{X}$. Applying (a fine-tuned version of) the result from (Worm et al., 2011), we obtain a decomposition into ergodic components $(\mathcal{X}^c)_{c\in C}$ of a subset of $\mathcal{X}$ that has full probability under any stationary probability w.r.t. $\bar{\rho}$. The result from (Worm et al., 2011) also ensures that each ergodic component $\mathcal{X}^c$ is equiped with a unique ergodic probability ϵ^c. Our contribution is then the following:

- By modifying ρ on state-action pairs that are never visited using the policy π, we obtain a family of transition channels $(\rho^c)_{c\in C}$, each with state-space the corresponding ergodic component $\mathcal{X}^c$, such that, denoting by $(\pi^c)_{c\in C}$ the corresponding restrictions of the policy, the following holds: for all stationary initial distribution $q_0 \in \Delta_{\mathcal{X}}$, the resulting process of states and actions defined by π and ρ is an average of those obtained on each ergodic component $\mathcal{X}^c$ by starting with the ergodic distribution ϵ^c and using the policy π^c and transition channel ρ^c. In other words, we obtain an *ergodic decomposition theorem for standard Borel MDPs with fixed policy*, which provides a decomposition of the MDP's process distributions, but also a corresponding family of MDP tuples $(\epsilon^c, \pi^c, \rho^c)_{c\in C}$, where each MDP pair (π^c, ρ^c) is equipped with the corresponding ergodic initial distribution ϵ^c. Crucially, this decomposition depends on

[35] Generalisations to partial observability would also be crucial, but we do not address this aspect here.

the policy π, which makes it a promising tool to capture the invariants of closed-loop behaviour in MDPs.

- We specialise our results to measurable group actions of standard Borel groups that have a stationary probability (which includes the group actions usually considered in the class-pose decomposition literature). In particular, we show that, for an MDP naturally defined by the group action and corresponding stationary probability on the group, *the partition in orbits of the group action coincides with the MDP's decomposition into ergodic components*. While this might not be the case for groups with no stationary probability, we argue that ergodic components provide, in general, a more natural concept of class than orbits, as they capture the *asymptotic properties* of the MDP's dynamics, which are key to understanding its dynamical invariants.

Our MDP version of the pose coordinate then relies on an adaptation of the group-theoretic notion of factor above to one for stationary MDPs, i.e., tuples made of a stationary initial distribution, a policy and a transition channel. With this new notion of factor, those of joining and minimal joining naturally adapt to stationary MDPs on measurable spaces — even though proving their basic properties requires a substantial amount of work. To the best of our knowledge, minimal joinings have not been previously considered in the literature, whether in a stationary MDP, group action or dynamical systems setting. We then define the minimal class-pose parametrisation of a standard Borel MDP (π, ρ) as a *minimal joining of the stationary MDPs* $(\epsilon^c, \pi^c, \rho^c)_{c \in C}$ from the ergodic decomposition above. We prove that in the finite case, minimal joinings of stationary MDPs always exist, and thus minimal class-pose parametrisations as well — unlike class-pose decompositions. Eventually, we prove that — for a countable number of ergodic components, but otherwise arbitrary standard Borel MDPs — this measure-theoretic notion of minimal class-pose parametrisations does generalise group-based class-pose decomposition. Along the way, we flag several technical difficulties arising from our measure-theoretic setting, especially for an uncountable number of ergodic components.

We then turn to the information-theoretic characterisation of these generalised notions of class and pose, in the finite case. As far as classes are concerned, we prove that the decomposition into ergodic components of a finite Markov chain provides the essentially unique solution, for maximal trade-off parameter, to an instance of the Divergence IB framework from Chapter 2. Here, what is being traded-off with compression is, in short, the *mutual information between the initial state and the time-average of the resulting trajectory*. We specialise the result to the case where the MDP implements a group action, which yields an *information-theoretic characterisation of projections on orbits*, thus addressing one of the main limitations identified at the end of Chapter 2 (see above). More generally, these results are an important step for formalising the duality between symmetry and information parsimony, as it provides a trade-off that binds transformations of a space to compressions preserving the features left invariant by these transformations (to a degree controlled by the trade-off parameter).

Our information-theoretic characterisation of the pose coordinate consists in proving that in the finite case, minimal joinings of stationary MDPs coincide with *minimum entropy joinings*, or, equivalently, *maximum multi-information joinings*. These characterisations formalise the intuitions, resp., that minimal joinings are the "most parsimonious" joinings, and that they are the ones that "maximise the interdependency" across the MDPs that are being joined. However, in contrast with the class coordinate, we do not propose, at this stage, a full information-theoretic trade-off whose solutions would capture minimal joinings.

Relevance to sensorimotor perception Conceptually, our formalism opens new directions for sensorimotor approaches to the study of perception.

First, it suggests that the group-structured "representations" from the representation learning community can be stripped from their representational narrative. I.e., the mathematical objects underlying this line of work could be integrated to a broader algebraic analysis of embodied agents' behavioural and neural dynamics, based on well-chosen commutation relations that these dynamics would satisfy (see Section 1.2.3). For instance, classes and poses are, in our reformulation, inherently *dynamical* objects — as they rely on the asymptotic properties of an MDP. Importantly, these dynamics do not refer to objects in the "external world" that would be defined independently from the agent, but only to the agent's sensorimotor interface — where here, the effect of an agent's actions on its sensory influx is formalised as the transition channel of an MDP, but this simplified setting is only aimed as a first step towards more realistic models. Moreover, let us stress that ergodic components are sets of points that — in a measure-theoretic sense — have the same attractor.[36] Each "class" can thus be seen as the asymptotic attractor of a given behaviour (i.e., here, an MDP ergodic component), rather than a "representation" outputed by some internal processing. This perspective strongly aligns with Closed-Loop Perception (CLP) theory's formalisation of sensorimotor percepts as attractors in the sensori-neural-motor loop (Ahissar et al., 2016) — see Section 1.1.3.

But on the other hand, these MDP ergodic components are also a generalisation of the orbits of a group action, which we saw were an interesting starting point for formalising object-related SMCs (see Section 1.2.3). Our framework thus highlights a potential integration of CLP theory's "percepts-as-attractors" into SMC theory's object-based SMCs, through the notion of MDP ergodic components — or suitable notions of ergodic component to be defined, in future work, in more general models of the sensorimotor loop (Rosas et al., 2025; Tishby et al., 2011).

Moreover, the fact that we "reversed the arrows and broke the bijectivity" in the factor relations means that the dynamics of the pose space can in certain respects be "richer" than (i.e., non-isomorphic to) the state-space dynamics, which strongly resonates with the self-sustained aspect of brain dynamics that the inside-out framework (Buzsáki et al., 2019) insists on (see Section 1.1.3). The fact that the pose dynamics are a *minimal* joining of the ergodic components can then be seen as operationalising the notions that (*i*) brain dynamics are *calibrated* by the sensorimotor interaction with the environment (Buzsáki et al., 2019), or (*ii*) brain dynamics *attune* to this sensorimotor interaction, thus allowing the agent to master the *structure of change* that underlies it (O'Regan et al., 2001). This specific formalisation, however, also introduces a conceptual innovation: the calibration/attunement would be *induced by information parsimony constraints*, as the latter would yield a tendency of different episodes of the agent's sensorimotor history to leave their trace, *as much as possible, on the same brain dynamics*. Crucially, such dynamics would thus yield some kind of *parsimonious fiction* — where the "fictional" dimension holds to the extent that the minimal joining of sensorimotor dynamics is "richer" than the latter. In particular, from this perspective, not only do ongoing brain dynamics not "represent true world states", but they don't even "refer to" sensorimotor activity strictly speaking — or to any "coarse-graining" of it. Rather, the "attunement" of brain dynamics through sensorimotor history induces them to *expand the realm of what is being "enacted" during perception*, beyond the patterns of bodily movement and corresponding raw sensory changes enacted by ongoing sensorimotor dynamics. It must be acknowledged, however, that our mathematical objects do not yet adequately describe the *coupling* between neuronal and sensorimotor dynamics on the time-scale of perception (Aguilera et al., 2013).

Eventually, the links that we draw with information parsimony open the way for connecting our results to information-theoretic models of embodied adaptive behaviour (Ay, 2015; Langer et al., 2024; Salge et al., 2014; Tishby et al., 2011).

[36] Ergodic components are defined as points whose *Césaro means* have the same limit (see Definition 3.3.1, Chapter 3): i.e., intuitively, whose trajectories are asymptotically concentrated on the same subset.

At this stage, though, we do not yet consider concrete models of embodied agents. This would require algorithms to discover minimal class-pose parametrisations (in their measure-theoretic MDP version). An information theory-based approach seems natural for that purpose, and our information-theoretic characterisations of classes and poses are first steps in this direction.

Publication Information. *The content from Chapter 3 has not been published yet.*

Chapter 4: Exact and Soft Successive Refinement of the Information Bottleneck

In Chapters 2 and 3, we capture several notions of "exact" symmetry as solutions to the edge cases of information-theoretic trade-offs, which yields to the definitions of *families* of corresponding "soft" symmetries, whose "softness" is parametrised by the trade-off parameter λ. The underlying intuition is that, as λ decreases, the system under study is increasingly coarse-grained, leading to the extraction of symmetries of a correspondingly coarser "granularity" λ. This intuition suggests that soft symmetries corresponding to a given granularity λ_1 should at least include those for finer granularity $\lambda_2 > \lambda_1$. It is thus crucial to investigate whether this property indeed holds in general, and if not to which extent and what it depends on.

This question is directly related to the broader one of whether for two compression channels κ_1 and κ_2 that solve the same IB-like trade-off for resp. parameters $\lambda_1 < \lambda_2$, the channel κ_1 is itself a coarse-graining of κ_2: i.e., of whether there exists a channel γ such that

$$\kappa_1 = \gamma \circ \kappa_2 \tag{1.3.7}$$

This is the question which we explore in Chapter 4, in the case of the classic IB. It turns out to be formally equivalent to that of the *successive refinability* of the IB problem, which reverses the order of the trade-off parameters: can the finer bottleneck channel κ_2 be obtained by combining κ_1 with another channel capturing new information about the source X, but without discarding any of the information captured by κ_1? This second formulation is relevant to understanding the informational limits of incremental learning. In both cases, the question can be understood, intuitively, as *whether the information that the coarser bottleneck κ_1 captures about the source X is entirely "contained" in the information that the finer bottleneck κ_2 captures about X*. Our aim is here to contribute to a better understanding of the successive refinability of the IB, but also to *quantify* the lack of successive refinement when it does not hold.

We start by proving that in the Gaussian and deterministic case, the IB is indeed successively refinable. These proofs are relatively straightforward applications of previous results in the IB literature. We then formalise the above intuition of successive refinability corresponding to an "inclusion" of information contents, by deriving (under mild assumptions) a characterisation of successive refinement in terms of inclusions of the *convex hulls* defined by the Bayesian inverses of the bottlenecks' compression channels. More precisely, if $q(X, T_1)$ and $q(X, T_2)$ are the joint distributions on source and bottleneck for resp. trade-off parameters $\lambda_1 < \lambda_2$, then successive refinement w.r.t. parameters (λ_1, λ_2) is here equivalent, essentially, to the condition

$$\mathsf{Hull}\{q(X|t_1), t_1 \in \mathcal{T}_1\} \subseteq \mathsf{Hull}\{q(X|t_2), t_2 \in \mathcal{T}_2\}, \tag{1.3.8}$$

where, for a set $E \subseteq \mathbb{R}^n$, we denote by $\mathsf{Hull}(E)$ the convex hull of E, i.e., the set of points obtained as convex combinations of points in E. This condition thus capture the intuition of "inclusion" of "information content" above if the "information content" a bottleneck T about the source X is understood as the convex hull $\mathsf{Hull}\{q(X|t), t \in \mathcal{T}\}$. This condition is visually represented in Figure 1.1.

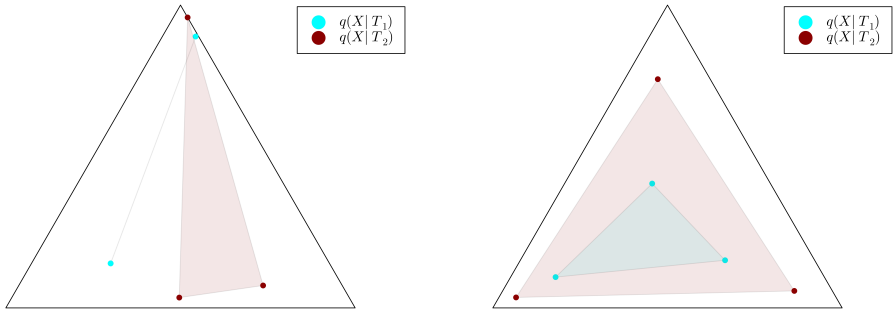


FIGURE 1.1: Illustration of the convex hull condition (satisfied on the right but not on the left). Black triangle: probability simplex $\Delta_{\mathcal{X}}$ with $|\Delta_{\mathcal{X}}| = 3$, where each vertex represent a Dirac probability δ_x on a symbol x, and the interior of the triangle represents arbitrary probability distributions on $\mathcal{X}$, seen as convex combinations of these Dirac probabilities. Red triangle: $\mathsf{Hull}\{q(X|t_2), t_2 \in \mathcal{T}_2\}$, with T_2 the finer bottleneck. Cyan triangle/segment: similarly for the coarser bottleneck T_1.

We use this characterisation, together with previous work studying the IB problem as the problem of computing the lower convex envelope of a well-chosen function, to prove that successive refinement (SR) of the IB always holds for binary source and relevancy variables. This geometric characterisation also yields a simple linear program that can be used to assess SR in the discrete case, which we use to numerically investigate this property in synthetic examples.

We then go beyond "exact" successive refinement by quantifying the "amount to which" successive refinement holds. Here, we start from a previously established characterisation of SR as the existence of a joint distribution distribution $q(X, T_1, T_2)$ that both extends each source-bottleneck distribution $q(X, T_1)$ and $q(X, T_2)$, and satisfies the Markov chain $X - T_2 - T_1$. As the latter Markov chain is equivalent to the condition $I(X; T_1|T_2) = 0$ of vanishing conditional mutual information, we propose to quantify the "lack" of successive refinement through the quantity

$$UI(X : T_1 \setminus T_2) := \min_{q \in \Delta_{q_1,q_2}} I_q(X; T_1|T_2),$$

where Δ_{q_1,q_2} denotes the space of joint distributions $q(X, T_1, T_2)$ extending both $q_1 := q(X, T_1)$ and $q_2 := q(X, T_2)$. This quantity has been previously studied under the name of *unique information*[37] (UI) (Bertschinger et al., 2013), and here yields a notion of *soft successive refinement* — in contrast with soft symmetries and ergodic components from resp. Chapter 2 and 3, this new notion of "softness" does not refer to a new structure, but to the quantification of a property, here (the lack of) successive refinement. We use unique information to numerically explore, on simple synthetic examples with low cardinalities of the source and relevancy variables, how this "lack" of SR varies as a function of λ_1 and λ_2. We then compare the evolution of the "information contents" $\mathsf{Hull}\{q(X|t_1), t_1 \in \mathcal{T}_1\}$ and $\mathsf{Hull}\{q(X|t_2), t_2 \in \mathcal{T}_2\}$ on the source simplex $\Delta_{\mathcal{X}}$ with the evolution of the unique information. Moreover, we observe that the unique information is always relatively small — even though not entirely negligible — with often sharp drops just after one of the trade-off parameters crosses a bifurcation value. While further work is necessary to test the extent to which these features generalise to more realistic scenarios, these results suggest that generically, the "lack" of successive refinability

[37] While this quantity has initially been introduced as part of the *partial information decomposition* literature, here we do not aim to contribute to this line of work.

of the IB is relatively mild, and that successive refinement is the "closest" to hold for trade-off parameters poised close to bifurcation values.

Eventually, we point out that previous results relating to what is known as the *Blackwell order* (Bertschinger et al., 2014) directly yield an interpretation of successive refinement in terms of decision problems. Crucially for the relevance of this framework to the study of soft invariances, it is this last characterisation that proves that successive refinement is indeed equivalent to equation (1.3.7) above.

Publication Information. *Chapter 4 is a condensed version of (Charvin et al., 2023a). A preliminary version of this work was published in (Charvin et al., 2022).*

Chapter 2

Information Parsimony and Symmetries of Stochastic Channels

2.1 Introduction

Group symmetries have become highly relevant to the study of adaptive behaviour and intelligent systems, from equivariant brain dynamics (Bertoni et al., 2021; Bressloff et al., 2001; Jirsa et al., 2022) to equivariant neural networks (Gerken et al., 2023) or AI-assisted scientific discovery (Liu et al., 2023; Teichner et al., 2023), with growing convergences between neuroscience and machine learning research (Higgins et al., 2022; Keller et al., 2026). Closer to the motivations of this thesis, the study of symmetries involving both sensory and motor spaces has long been part of or adjacent to *sensorimotor perception* research (O'Regan et al., 2001) — see Section 1.2 for a detailed analysis of these intersecting lines of work. In particular, the abstract symmetries of embodied agents' sensorimotor interface seem to be a promising mathematical object for the formalisation of what (O'Regan et al., 2001) refers to as *apparatus-related sensorimotor contingencies* (see the subsection "Representation learning, reinforcement learning & apparatus-related SMCs" in Section 1.2.3).

This relevance of group-theoretic tools is often understood as a consequence of the pervasiveness of symmetries in the natural world (Higgins et al., 2018, 2022) or in the interaction of embodied agents with their environment (O'Regan et al., 2001). However, the existence of such symmetries does not explain why they would be relevant from the *intrinsic* point of view of an embodied, in particular biological agent. In other words, one needs to ask what constraints would *induce* agents to leverage these symmetry structures.

One possible answer is that the presence of symmetries in a system allows for a *simpler* description of it — which is highly valuable under stringent informational constraints, as it offers opportunities for informationally "cheap" interactions (Langer et al., 2021; Montúfar et al., 2015). In other words, a system's symmetries afford the possibility of *informationally parsimonious* descriptions of it (van der Ouderaa et al., 2024) — see Sections 1.1.1 and 1.1.2 for a more detailed discussion of the hypothesis of a "duality" between structure and information parsimony.

In particular, the *projection on orbits* of a symmetry group's action can be seen as an "information-preserving compression", as it preserves the information about anything invariant under the group action (see Proposition A.0.4 for a formal statement). This suggests that projections on orbits might be solutions to well-chosen *rate-distortion* problems (Cover et al., 2009; Zaidi et al., 2020), hence opening the way to the integration of group symmetries into an information-theoretic framework. If successful, such an integration could formalise the link between symmetry and information parsimony, but also (*i*) yield natural ways to "soften" group symmetries into flexible concepts more relevant to real-world data — which often lacks exact symmetries despite exhibiting a strong "structure" — and (*ii*) enable symmetry discovery through the optimisation of information-theoretic quantities.

As a first step in this direction, we introduce a novel rate-distortion-inspired framework whose solutions capture, or at least mimick, the projection on orbits of certain group symmetries. We call it Divergence Information Bottleneck (DIB), as it generalises the Information Bottleneck (IB) method (Tishby et al., 2000). Here, the compression of the data's distribution is traded-off with the perservation its divergence from a given hierarchical model, potentially under constraints on the shape of compression channels. As divergences from hierarchical models are geometric measures of complexity (Ay et al., 2011), this setting produces *complexity-preserving* compressions. With appropriate choices of hierarchical models and shape constraints, we obtain compression channels which, for full divergence preservation, characterise various group-theoretic symmetries. Allowing only *partial* divergence preservation then leads to principled definitions of *soft symmetries*, whose "softness" is parametrised by the compression-divergence preservation trade-off: i.e., softer symmetries correspond to more compression and less preserved divergence.

The IB method has previously been argued to extract invariances (Achille et al., 2018a). In Section 2.2, we prove that it indeed characterises group-theoretic channel invariances. This motivates the DIB framework, which we specialise to characterise the equivariances of a channel $\mu(Y|X)$ and the invariances of a distribution $\mu(A)$ (Section 2.3). We then present simple synthetic numerical experiments on soft equivariances (Section 2.4), where we study channels satisfying a series of nested equivariances that have been perturbed to various degrees. We show that our framework recovers the perturbed equivariances, at successive bifurcation points of the trade-off parameter corresponding to increasingly compressed resolutions. In Sections 2.5.1 and 2.5.2, we summarise our results and discuss their limitations. Eventually, in Section 2.5.3, we outline a way forward to address the latter, in particular regarding how to discover not only the compression made possible by the presence of probabilistic symmetries, but the corresponding symmetry transformations themselves.

Let us stress that, again, while these results are of a formal and computational nature, they are also motivated by and aimed at sensiromotor theories of perception. Indeed, while the equivariances considered here are more generic than the symmetries of embodied agent's sensorimotor interfaces, the former provide a preliminary step that could be specialised to the latter in future work (see our discussion on apparatus-related sensorimotor contingencies in Section 1.2.3).

Assumptions and notations All alphabets are finite, except bottleneck alphabets $\mathcal{T} := \mathbb{N}$.[1] The probability simplex defined by an alphabet $\mathcal{A}$ is denoted by $\Delta_{\mathcal{A}}$. The *support* of a distribution $\mu \in \Delta_{\mathcal{A}}$ is

$$\operatorname{supp}(\mu) := \{a \in \mathcal{A} \ : \ \mu(a) > 0\}. \tag{2.1.1}$$

We say that μ is *full-support* if $\operatorname{supp}(\mu) = \mathcal{A}$. The set of conditional probabilities, also called *channels*, from $\mathcal{A}$ to $\mathcal{B}$, resp. to $\mathcal{A}$ itself, is denoted by $\mathscr{K}(\mathcal{A}, \mathcal{B})$, resp. $\mathscr{K}(\mathcal{A})$. Whenever this creates no confusion, a function $f : \mathcal{A} \to \mathcal{B}$ is seen as the corresponding deterministic channel γ_f defined, for all $a \in \mathcal{A}$, $b \in \mathcal{B}$, by $\gamma_f(b|a) := \delta_{f(a)=b}$. The *hook-up* of a distribution $\mu \in \Delta_{\mathcal{A}}$ with a channel $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$ is the joint distribution $\mu\gamma \in \Delta_{\mathcal{A}\times\mathcal{B}}$ defined by $\mu\gamma(a, b) := \mu(a)\gamma(a|b)$. The *push-forward* of a distribution $\mu \in \Delta_{\mathcal{A}}$ through a channel $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$ is the distribution $\gamma \cdot \mu \in \Delta_{\mathcal{B}}$ defined, for all $b \in \mathcal{B}$, by

$$(\gamma \cdot \mu)(b) := \sum_{a\in\mathcal{A}} \gamma(b|a)\mu(a).$$

[1] Our results should adapt easily to the case of finite $|\mathcal{T}|$, but by choosing $\mathcal{T} := \mathbb{N}$ we here ignore the question of the minimal cardinality of $\mathcal{T}$ for which our results still hold.

By extension, $f \cdot a := f(a)$ for an element a and a function f. The symbol $\circ$ denotes channel composition. The set of bijections of $\mathcal{A}$ is $\mathrm{Bij}(\mathcal{A})$, the uniform distribution $\upsilon_{\mathcal{A}}$, the identity map $\mathtt{Id}_{\mathcal{A}}$, and $\mathcal{S}^{\complement} := \mathcal{A} \setminus \mathcal{S}$ for $\mathcal{S} \subseteq \mathcal{A}$. For $\mu_1 \in \Delta_{\mathcal{A}}, \mu_2 \in \Delta_{\mathcal{B}}$, their *tensor product* is defined through $(\mu_1 \otimes \mu_2)(a, b) := \mu_1(a)\mu_2(b)$. Similarly $(\phi \otimes \psi)(b, b'|a, a') := \phi(b|a)\psi(b'|a')$ for $\phi \in \mathscr{K}(\mathcal{A}, \mathcal{B})$, $\psi \in \mathscr{K}(\mathcal{A}', \mathcal{B}')$. Eventually, we refer to Chapter 2 in (Ay et al., 2017) for an explicit definition of an exponential family on a probability simplex $\Delta_{\mathcal{A}}$.

2.2 Information Bottleneck and Group Invariances

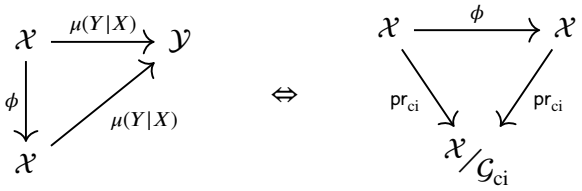
Recall that the IB implements a trade-off between compressing a variable, say X, and preserving the information that X carries about a second variable Y. More precisely, let $\mu := \mu(X, Y) \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ with $\mu(X)$ full-support. The corresponding IB problem is then (Gilad-Bachrach et al., 2003)

$$\mathrm{IB}(\lambda) := \underset{\substack{\kappa \in \mathscr{K}(\mathcal{X},\mathcal{T}) : \\ I_\kappa(T;Y) \geq \lambda}}{\arg\min} \; I_\kappa(X;T), \tag{2.2.1}$$

where mutual informations are computed w.r.t. $q_\kappa(X, Y, T) := \mu\kappa$ and $0 \leq \lambda \leq \Lambda := I(X;Y)$. The IB method can be seen as a softening of the notion of *minimal sufficient statistic* (see, e.g., Definition 3.3.16), and, crucially for us, has been suggested to extract invariances (Achille et al., 2018a). However, surprisingly, we are not aware of an explicit result showing that the solutions to the IB problem (2.2.1) do extract invariances in the group-theoretic sense. In this section, we show that this is indeed the case.

Definition 2.2.1. An *invariance* of the channel $\mu(Y|X)$ is a bijection $\phi \in \mathrm{Bij}(\mathcal{X})$ such that $\mu(Y|X)\circ\phi = \mu(Y|X)$. The *channel invariance group* $\mathcal{G}_{\mathrm{ci}}$ of $\mu(Y|X)$ is the set of invariances of $\mu(Y|X)$ equiped with channel composition. The corresponding projection on orbits is written $\mathsf{pr}_{\mathrm{ci}} : \mathcal{X} \to \mathcal{X}/\mathcal{G}_{\mathrm{ci}}$.

Our starting point is the observation that this projection on orbits characterises the channel invariance group. Indeed, it can be easily verified that $\phi \in \mathcal{G}_{\mathrm{ci}} \Leftrightarrow \mathsf{pr}_{\mathrm{ci}}\circ\phi = \mathsf{pr}_{\mathrm{ci}}$. This can be expressed with the following equivalence of commutative diagrams:



(2.2.2)

which means that the commutativity of the left-hand-side diagram — i.e., the defining property of invariances — is equivalent to that of the right-hand-side one.

We will show that the solutions $\kappa \in \mathrm{IB}(\Lambda)$ essentially coincide with $\mathsf{pr}_{\mathrm{ci}}$, thus yielding an information-theoretic characterisation of $\mathcal{G}_{\mathrm{ci}}$ through such κ. We will need the equivalence relation

$$x \sim_{\mathcal{X}} x' \quad \Leftrightarrow \quad \mu(Y|x) = \mu(Y|x'), \tag{2.2.3}$$

with corresponding partition $(\mathcal{X}^c)_{c\in C}$ and projection $\mathsf{pr}_{\mathcal{X}} : \mathcal{X} \to C$ on equivalence classes; along with the following notion.

Definition 2.2.2. The set of *congruent channels* (Ay et al., 2017) from $\mathcal{A}$ to $\mathcal{B}$, denoted by $\mathscr{K}_{\mathsf{cong}}(\mathcal{A}, \mathcal{B})$, is that of channels ι such that there exists a function $f : \mathcal{B} \to \mathcal{A}$ with $f\circ\iota = \mathtt{Id}_{\mathcal{A}}$.

In particular, composing an compression channel κ with a congruent channel ι can be seen as a trivial operation, in the sense that it does not induce further compression, as the output of κ can be unambigously recovered from that of $\iota \circ \kappa$.

Theorem 2.2.3. *For $\Lambda := I(X;Y)$ and all $\phi \in \mathrm{Bij}(\mathcal{X})$, the following holds:*

(i) $\mathrm{IB}(\Lambda) = \left\{\iota \circ \mathsf{pr}_{\mathcal{X}} \ : \ \iota \in \mathscr{K}_{\mathsf{cong}}(\bar{X}, \mathcal{T})\right\}$.

(ii) For all $0 \leq \lambda \leq \Lambda$ and $\kappa \in \mathrm{IB}(\lambda)$, there exists a channel $\gamma \in \mathscr{K}(\mathcal{C}, \mathcal{T})$ such that $\kappa = \gamma \circ \mathsf{pr}_{\mathcal{X}}$.

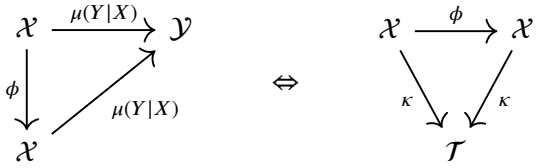
(iii) Let $\kappa \in \mathrm{IB}(\Lambda)$. Then $\phi \in \mathcal{G}_{\mathrm{ci}}$ if and only if $\kappa \circ \phi = \kappa$.

(iv) If $\phi \in \mathcal{G}_{\mathrm{ci}}$, then $\kappa \circ \phi = \kappa$ also holds for all $0 \leq \lambda \leq \Lambda$ and $\kappa \in \mathrm{IB}(\lambda)$.

(v) The projection $\mathsf{pr}_{\mathcal{X}}$ defined by $\sim_{\mathcal{X}}$ coincides with the projection on orbits $\mathsf{pr}_{\mathrm{ci}}$.

Proof. See Appendices B.1 and B.2. □

Crucially, point *(iii)* means that invariances are thoses bijections ϕ such that the effect of transforming $\mathcal{X}$ with ϕ is "quotiented out" by $\kappa \in \mathrm{IB}(\Lambda)$, making them indistinguishable from the identity. This information-theoretic characterisation of group invariances from point *(iii)* can be expressed through the following equivalence of commutative diagrams:



(2.2.4)

I.e., we replaced the quotient space $\mathcal{X}/\mathcal{G}_{\mathrm{ci}}$ from Diagram (2.2.2) by a bottleneck space $\mathcal{T}$, and the corresponding projection $\mathsf{pr}_{\mathrm{ci}}$ by the bottleneck projection $\kappa \in \mathrm{IB}(\Lambda)$. Points *(i)* explains why this is possible: these compression channels κ implement precisely (up to trivial post-processing by congruent channels) the quotient of $\mathcal{X}$ by the equivalence relation $\sim_{\mathcal{X}}$ that equates elements of $\mathcal{X}$ providing the same information about Y (see (2.2.3)).

Point *(ii)* shows that the factorisation of bottleneck channels $\kappa \in \mathrm{IB}(\lambda)$ by the projection pr actually holds for all parameter λ — note however that for $\lambda < \Lambda$, the channel γ in point *(ii)* might not be congruent. Point *(iv)* is a direct consequence of point *(ii)*, and shows that the "quotienting out" of invariances by bottlenecks also occurs for all values of the trade-off parameter λ, even though it is only a full characterisation for $\lambda = I(X;Y)$. Point *(v)*, combined with point *(i)*, means that the projection $\mathsf{pr}_{\mathrm{ci}}$, defined purely in group-theoretic terms, is characterised as the solution to the zero-distortion case of a generalised rate-distortion problem, here the IB (Zaidi et al., 2020). Note that, as a consequence of point *(i)*, the solutions for $\lambda := \Lambda$ only depend on the channel $\mu(Y|X)$ from source to relevancy, and not on the distribution $\mu(X)$ on the source. However, this source distribution becomes important once we consider the cases of *partial* information preservation $0 \leq \lambda < \Lambda$. Let us also mention that point *(i)* is redundant with existing results: it can be seen as the fact that $\mathrm{IB}(\Lambda)$ consists of minimal sufficient statistics of X w.r.t. Y, proven in (Shamir et al., 2010). But our new proof also yields point *(iv)* and generalises to that of Theorem 2.3.1 below.

Moreover, this information parsimony perspective on group invariances suggests a principled way to "soften" them:

Definition 2.2.4. Let $0 \leq \lambda \leq \Lambda$. A *$\lambda$-invariance* is a channel $\phi \in \mathscr{K}(\mathcal{X})$ such that there exists $\kappa \in \mathrm{IB}(\lambda)$ with $\kappa \circ \phi = \kappa$.

In other words, a soft invariance is defined through the very same equation that characterises exact invariances, but where the fully information-preserving compression κ is now only a *partially* information-preserving compression. Moreover, we allow ϕ to be non-invertible and stochastic. Here, the smaller λ, the more coarse-grained is the bottleneck $\kappa \in \mathrm{IB}(\lambda)$, and thus, intuitively, the corresponding λ-invariances captures structure in more "schematised" version of the source X. Note, however, that at this stage, it is not clear how to explicitly recover the λ-invariances ϕ from a bottleneck $\kappa \in \mathrm{IB}(\lambda)$ (see Section 2.5.2).

While previous work has already largely explored the idea that the classic IB method extracts invariances (see Chapter 1, Section 1.1.2), our explicit group-theoretic formalisation suggests deeper connections between group-theoretic symmetries and trade-offs between information parsimony and other well-chosen information-theoretic quantities. It also provides guidance for investigating such generalisations — i.e., for reformulating and softening probabilistic symmetries with the language of information theory. The following sections provide first steps in this direction.

2.3 Divergence Information Bottleneck and Group Symmetries

In this section, we present a novel generalisation of the IB framework, which we then apply to reformulate channel equivariances and distribution invariances in an information-theoretic language.

2.3.1 General framework

Fix a full-support distribution $\mu = \mu(A) \in \Delta_{\mathcal{A}}$, an exponential family $\mathcal{E} \subseteq \Delta_{\mathcal{A}}$, and a subset of compression channels $\mathscr{K}_{\mathsf{shape}} \subseteq \mathscr{K}(\mathcal{A}, \mathcal{T})$. We then define the *Divergence Information Bottleneck* (DIB) as

$$\mathrm{DIB}(\lambda) := \underset{\substack{\kappa \in \mathscr{K}_{\mathsf{shape}} \\ D(\kappa\cdot\mu||\kappa\cdot\mathcal{E}) \geq \lambda}}{\arg\min} \; I_\kappa(A;T), \tag{2.3.1}$$

where $0 \leq \lambda \leq \Lambda := D(\mu||\mathcal{E})$, and

$$D(\mu||\mathcal{E}) := \inf_{\nu\in\mathcal{E}} D\big(\mu||\nu\big) = D\big(\mu||\tilde{\mu}\big), \tag{2.3.2}$$

$$D(\kappa\cdot\mu||\kappa\cdot\mathcal{E}) := \inf_{\nu\in\mathcal{E}} D\big(\kappa\cdot\mu||\kappa\cdot\nu\big) = D\big(\kappa\cdot\mu||\kappa\cdot\tilde{\mu}\big). \tag{2.3.3}$$

Here $\tilde{\mu} \in \mathcal{E}$ is the unique distribution which achieves the minimum in (2.3.2), and happens to also achieves the minimum in (2.3.3) (see Appendix B.3.1). While $D(\mu||\mathcal{E})$ is the divergence of the distribution μ from the exponential family $\mathcal{E}$, here $D(\kappa\cdot\mu||\kappa\cdot\mathcal{E})$ measures the "divergence of μ from $\mathcal{E}$ *after compression by the channel* κ". Solutions to (2.3.1) can thus be seen as optimal compressions of A under the constraint of (partially or fully) preserving the divergence of μ from the exponential family $\mathcal{E}$. The choice of $\mathscr{K}_{\mathsf{shape}}$ allows to potentially enforce constraints on the shape of compression channels κ (e.g., if $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ is a Cartesian product, we could require that $\kappa := \kappa_1 \otimes \kappa_2$, i.e., that κ processes each coordinate separately).

Intuitively, $D(\mu||\mathcal{E})$ measures the presence of a specific structure in μ, formalised as the divergence from the family $\mathcal{E}$ of distributions which do not have such structure. E.g., for $\mathcal{A} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{E} := \Delta_{\mathcal{X}} \otimes \Delta_{\mathcal{Y}}$, we have $D(\mu||\mathcal{E}) = I(X;Y)$: the corresponding DIB (with e.g. $\mathscr{K}_{\mathsf{shape}} = \mathscr{K}(\mathcal{X}\times\mathcal{Y}, \mathcal{T})$) is a *mutual information-preserving* joint compression of X and Y (Charvin et al., 2023b). More generally, the divergence from a *hierarchical model*[2]

[2] See, e.g., (Ay et al., 2017) for an explicit definition of hierarchical models — which we will not rely on here.

$\mathcal{E}$ measures the complexity of a system's given set of interdependencies (Ay et al., 2011). The DIB problem is tailored for this setting, where solutions to (2.3.1) are thus *complexity-preserving optimal compressions*. However, here we consider a general exponential family $\mathcal{E}$ because this is all that is required to prove Theorem 2.3.1 below.

Let us define, on $\mathcal{A}$, the equivalence relation

$$a \sim a' \quad \Leftrightarrow \quad \mu(a)\tilde{\mu}(a') = \mu(a')\tilde{\mu}(a), \tag{2.3.4}$$

with corresponding partition $(\mathcal{A}^c)_{c\in C}$ and projection $\mathsf{pr} : \mathcal{A} \to C$ on equivalence classes. Then:

Theorem 2.3.1. *If $\mathscr{K}_{\mathsf{shape}} = \mathscr{K}(\mathcal{A}, \mathcal{T})$ and μ is ful-support, then:*

(i) $\mathrm{DIB}(\Lambda) = \big\{\iota\circ\mathsf{pr} \ : \ \iota \in \mathscr{K}_{\mathsf{cong}}(C, \mathcal{T})\big\}$.

(ii) For all $0 \le \lambda \le \Lambda$ and $\kappa \in \mathrm{DIB}(\lambda)$, there exists a channel $\gamma \in \mathscr{K}(C, \mathcal{T})$ such that $\kappa = \gamma\circ\mathsf{pr}$.

Proof. See Appendices B.1 and B.3.2. □

Point (i) shows that for full support $\mu(A)$ and no constraints on the shape of compression channels, the fully divergence-preserving solutions $\kappa \in \mathrm{DIB}(\Lambda)$ coincide, up to trivial transformations, with the clustering of $\mathcal{A}$ defined by the relation (2.3.4). Point (ii) shows that, similarly as for the classic IB, the factorisation of a bottleneck channel κ by the projection pr also holds for arbitrary trade-off parameter λ (compare with point (iii) in Theorem 2.2.3).

2.3.2 Application to equivariances

Consider now $\mathcal{A} = \mathcal{X}\times\mathcal{Y}$ equipped with a full support distribution $\mu = \mu(X, Y)$; in particular, the conditional distribution $\mu(Y|X)$ is uniquely defined.

Definition 2.3.2. An *equivariance* of the channel $\mu(Y|X)$ is a pair $(\phi, \psi) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$ such that $\mu(Y|X)\circ\phi = \psi\circ\mu(Y|X)$. The corresponding *group of channel equivariances* is the set of equivariances of $\mu(Y|X)$ equiped with the group law $(\phi, \psi)\cdot(\phi', \psi') := (\phi\circ\phi', \psi\circ\psi')$. The corresponding space of orbits is denoted by ${}^{\mathcal{X}\times\mathcal{Y}}\!/_{\mathcal{G}_{\mathrm{ce}}}$, and the corresponding projection on orbits by $\mathsf{pr}_{\mathrm{ce}} : \mathcal{X}\times\mathcal{Y} \to {}^{\mathcal{X}\times\mathcal{Y}}\!/_{\mathcal{G}_{\mathrm{ce}}}$.

Crucially, as was the case for invariances (see Section 2.2), the projection on orbits $\mathsf{pr}_{\mathrm{ce}}$ characterises the group $\mathcal{G}_{\mathrm{ce}}$:

Lemma 2.3.3. *Let $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$. The following are equivalent:*

(i) (ϕ, ψ) is an equivariance of $\mu(Y|X)$,

(ii) For all $(x, y) \in \mathcal{X}\times\mathcal{Y}$, we have $\mu(y|x) = \mu(\tau\cdot y|\sigma\cdot x)$,

(iii) $\mathsf{pr}_{\mathrm{ce}}\circ(\phi\otimes\psi) = \mathsf{pr}_{\mathrm{ce}}$.

Proof. See Appendix B.3.3. □

The characterisation $(i) \Leftrightarrow (iii)$ of equivariances can be expressed with the following equivalence of commutative diagrams:

$$\begin{array}{ccc} \mathcal{X} & \xrightarrow{\mu(Y|X)} & \mathcal{Y} \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\psi} \\ \mathcal{X} & \xrightarrow{\mu(Y|X)} & \mathcal{Y} \end{array} \quad \Leftrightarrow \quad \begin{array}{ccc} \mathcal{X}\times\mathcal{Y} & \xrightarrow{\phi\otimes\psi} & \mathcal{X}\times\mathcal{Y} \\ & {\scriptstyle\mathsf{pr}_{\mathrm{ce}}}\searrow \quad \swarrow{\scriptstyle\mathsf{pr}_{\mathrm{ce}}} & \\ & \mathcal{X}\times\mathcal{Y}/_{\mathcal{G}_{\mathrm{ce}}} & \end{array} \tag{2.3.5}$$

Our aim is thus to proceed similarly as for channel invariances and the classic IB mehod: we want to obtain a generalised IB problem such that the solutions for maximal trade-off parameter λ coincide with the projection on orbits pr_{ce}, or at least mimic its characterisation (2.3.5) of equivariances.

Note that here pr_{ce} does not compress $\mathcal{X}$ and $\mathcal{Y}$ separately but *jointly* (Charvin et al., 2023b): i.e., the orbits are not necessarily of the form $\mathcal{X}^c \times \mathcal{Y}^c$ for some $\mathcal{X}^c \subseteq \mathcal{X}$ and $\mathcal{Y}^c \subseteq \mathcal{Y}$. This can be seen, e.g., by considering $\mathcal{X} := \mathcal{Y} := \{x_1, x_2\}$, $p \in [0,1]$ and the binary symmetric channel defined by $\mu(x_j|x_i) = 1-p$ for $i = j$ and $\mu(x_j|x_i) = p$ for $i \neq j$. Here the equivariance group is $\{\mathsf{Id}, \phi \otimes \phi\}$, where Id is the identity on $\mathcal{X} \times \mathcal{X}$, while ϕ permutes x_1 and x_2. The corresponding orbits are thus $\{(x_1,x_1),(x_2,x_2)\}$ and $\{(x_2,x_1),(x_1,x_2)\}$: i.e., they are "diagonal". This example suggests that we should impose no constraint on the shape of compression channels: $\mathscr{K}_{\mathsf{shape}} := \mathscr{K}(\mathcal{X} \times \mathcal{Y}, \mathcal{T})$.

Moreover, from Lemma 2.3.3, it can be verified that for all $(x,y),(x',y') \in \mathcal{X} \times \mathcal{Y}$, the equality $\mathsf{pr}_{\text{ce}}(x,y) = \mathsf{pr}_{\text{ce}}(x',y')$ implies $\mu(y|x) = \mu(y'|x')$. Based on this observation, we search for an exponential family $\mathcal{E}$ such that the relation $\sim$ from equation (2.3.4) is equivalent to

$$(x,y) \sim (x',y') \quad \Leftrightarrow \quad \mu(y|x) = \mu(y'|x'). \tag{2.3.6}$$

This is achieved by choosing

$$\mathcal{E} := \mathcal{E}_{\text{ce}} := \{\nu_{\mathcal{X}} \otimes \upsilon_{\mathcal{Y}},\ \nu_{\mathcal{X}} \in \Delta_{\mathcal{X}}\},$$

where we recall that $\upsilon_{\mathcal{Y}}$ is the uniform distribution on $\mathcal{Y}$. Indeed, it can be easily verified that the projection of μ on the exponential family $\mathcal{E}_{\text{ce}}$ is then $\tilde{\mu} := \mu_{\mathcal{X}} \otimes \upsilon_{\mathcal{Y}}$, where $\mu_{\mathcal{X}}$ is the marginal of μ on $\mathcal{X}$ and $\upsilon_{\mathcal{Y}}$ is the uniform distribution on $\mathcal{Y}$. The relation $\sim$ from Section 2.3.1 (see equation (2.3.4)) thus becomes

$$\begin{aligned}
(x,y) \sim (x',y') \quad &\Leftrightarrow \quad \mu(x,y)\tilde{\mu}(x',y') = \mu(x',y')\tilde{\mu}(x,y) \\
&\Leftrightarrow \quad \mu(x,y)\mu(x')\frac{1}{|\mathcal{Y}|} = \mu(x',y')\mu(x)\frac{1}{|\mathcal{Y}|} \\
&\Leftrightarrow \quad \mu(y|x) = \mu(y'|x').
\end{aligned}$$

Note that $\mathcal{E}_{\text{ce}}$ coincides with the hierarchical model of probabilities on $\mathcal{X} \times \mathcal{Y}$ that actually depend only on $\mathcal{X}$. Borrowing from the geometric approach to complexity (Ay et al., 2011), we thus interpret $D(\mu||\mathcal{E}_{\text{ce}})$ as the "degree to which the system (X,Y) is more than just X", or equivalently, "the amount of information that $\mu(Y|X)$ carries about (X,Y)". More precisely, $\mathcal{E}_{\text{ce}}$ is the family of probabilities $\nu(X,Y)$ such that $\nu(Y|x)$ (when well-defined) is always equal to the maximum entropy distribution $\nu = \upsilon_{\mathcal{Y}}$. This condition means, intuitively, that the channel $\nu(Y|X)$ is "maximally uninformative", i.e., that it "carries no information" about the system (X,Y). Through the decomposition $\nu(X,Y) = \nu_{\mathcal{X}}\nu(Y|X)$, this is equivalent to the intuition that all the information about (X,Y) is contained in the marginal $\nu_{\mathcal{X}}$.

The divergence $D(\mu||\mathcal{E}_{\text{ce}})$ then quantifies "how far" μ is from such distributions. Thus the set of solutions to the DIB problem (2.3.1) with $\mathcal{A} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{E} = \mathcal{E}_{\text{ce}}$ and parameter λ can be interpreted as that of optimal compressions *preserving the information carried by the channel* $\mu(Y|X)$ (to the degree λ). We denote by $\text{DIB}_{\text{ce}}(\lambda)$ this set of solutions, and pr_{ce} the projection on the equivalence classes defined by the relation $\sim$ in (2.3.6).

Theorem 2.3.4. *The following holds, for* $\Lambda := D(\mu||\mathcal{E}_{\text{ce}})$ *and all* $(\phi,\psi) \in \text{Bij}(\mathcal{X}) \times \text{Bij}(\mathcal{Y})$:[3]

(i) $\text{DIB}_{ce}(\Lambda) = \left\{\iota \circ \mathsf{pr}_{\text{ce}} \ :\ \iota \in \mathscr{K}_{\mathsf{cong}}(\mathcal{X} \times \mathcal{Y}, \mathcal{T})\right\}$.

[3] Appendix B.3.5 clarifies how this work relates to our previous results in (Charvin et al., 2023b).

(ii) *For all* $0 \leq \lambda \leq \Lambda$ *and* $\kappa \in \mathrm{DIB}_{ce}(\lambda)$*, there exists a channel* $\gamma \in \mathscr{K}(\mathcal{X} \times \mathcal{Y}, \mathcal{T})$ *such that* $\kappa = \gamma \circ \mathsf{pr}$.

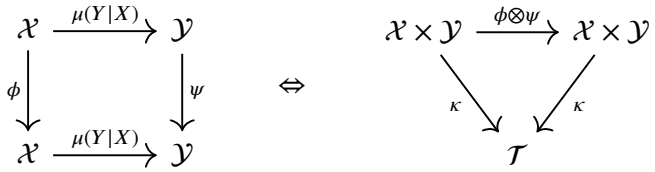
(iii) *Let* $\kappa \in \mathrm{DIB}_{\mathrm{ce}}(\Lambda)$. *Then* $(\phi, \psi) \in \mathcal{G}_{\mathrm{ce}}$ *if and only if* $\kappa \circ (\phi \otimes \psi) = \kappa$.

(iv) *If* $(\phi, \psi) \in \mathcal{G}_{\mathrm{ce}}$*, then* $\kappa \circ (\phi \otimes \psi) = \kappa$ *also holds for all* $0 \leq \lambda \leq \Lambda$ *and* $\kappa \in \mathrm{DIB}_{\mathrm{ce}}(\lambda)$.

(v) *The projection* pr *defined by* $\sim$ *in equation* (2.3.6) *does not, in general, coincide with* $\mathsf{pr}_{\mathrm{ce}}$.

Proof. See Appendix B.3.4. Points (i) and (ii) are direct applications of Theorem 2.3.1, and the remaining points follow easily, with similar proofs as for the classic IB case. □

Most importantly, point (i) means that equivariances of $\mu(Y|X)$ are those pairs of transformations (ϕ, ψ) such that the effect of simultaneously transforming $\mathcal{X}$ with ϕ and $\mathcal{Y}$ with ψ is "quotiented out" by the coarse-grainings $\kappa \in \mathrm{DIB}(\Lambda)$, making these transformations indiscernible from the identity. The $\mathrm{DIB}_{\mathrm{ce}}$ framework thus provides an *information-theoretic characterisation of equivariances*. It can be summarised with the following equivalence of commutative diagrams:



i.e., we replaced the quotient space $\mathcal{X}\times\mathcal{Y}/\mathcal{G}_{\mathrm{ce}}$ from the Diagram (2.3.5) by a bottleneck space $\mathcal{T}$, and the projection on orbits $\mathsf{pr}_{\mathrm{ce}}$ by a compression channel $\kappa \in \mathrm{DIB}_{\mathrm{ce}}(\Lambda)$ implementing a trade-off between compression and preservation of the divergence $D(\mu||\mathcal{E}_{\mathrm{ce}})$.

Note that, similarly as for the classic IB, point (i) also implies that the solutions for $\lambda := \Lambda$ only depend on the channel $\mu(Y|X)$, and not on the distribution $\mu(X)$ on the source. However, $\mu(X)$ becomes important once we consider the cases of *partial* divergence preservation $0 \leq \lambda < \Lambda$.

Moreover, point (ii) in Theorem 2.3.4 means that the "quotienting out" of equivariances happens actually for all granularities λ, even though the equivalence only holds for $\lambda = \Lambda$.

However, unfortunately, point (iii) says that the clustering pr obtained from $\mathrm{DIB}_{\mathrm{ce}}(\Lambda)$ does not always coincide with the projection on orbits $\mathsf{pr}_{\mathrm{ce}}$. See Appendix B.3.4 for mathematical details, Section 2.5.2 for a longer discussion of this limitation, and Section 2.5.3 for an outline of how this limitation could be overcome in future work.

Following again our approach for the classic IB, we can now draw upon our new information parsimony perspective on equivariances to soften this group-theoretic notion, where each granularity λ defines a corresponding set of soft equivariances:

Definition 2.3.5. Let $0 \leq \lambda \leq \Lambda$. A *$\lambda$-equivariance* is a pair of channels $(\phi, \psi) \in \mathscr{K}(\mathcal{X}) \otimes \mathscr{K}(\mathcal{Y})$ such that there exists $\kappa \in \mathrm{DIB}_{\mathrm{ce}}(\lambda)$ with $\kappa \circ (\phi \otimes \psi) = \kappa$.

In other words, a soft equivariance is defined through the very same equation that characterises exact equivariances, but where the fully information-preserving compression κ is now only a *partially* information-preserving compression. Moreover, we allow ϕ and ψ to be non-invertible and stochastic. Such soft equivariances will be explored numerically in Section 2.4.

To conclude this section, let us point out that the classic IB can be recovered as a DIB with the same exponential family $\mathcal{E}_{\mathrm{ce}}$ as for equivariances, but with shape constraints $\mathscr{K}_{\mathsf{shape}} \subsetneq \mathscr{K}(\mathcal{A}, \mathcal{T})$ which impose that κ can only compress $\mathcal{X}$ and not $\mathcal{Y}$. See Appendix B.3.6.

2.3.3 Application to distribution invariances

We can proceed similarly for transformations leaving a given distribution invariant. I.e., let $\mu \in \Delta_{\mathcal{A}}$ be full support, and define the group $\mathcal{G}_{\text{di}}$ of *distribution invariances* as the group of $\Phi \in \text{Bij}(\mathcal{A})$ such that $\mu(\Phi \cdot A) = \mu(A)$. As we do not consider any structure on $\mathcal{A}$, we choose unconstrained compression channels, i.e., $\mathscr{K}_{\text{shape}} = \mathscr{K}(\mathcal{A}, \mathcal{T})$. Moreover, as $\Phi \in \mathcal{G}_{\text{di}}$ if and only if $\mu(a) = \mu(\Phi \cdot a)$ for all a, it is natural to search for an exponential family yielding the equivalence relation $a \sim a' \Leftrightarrow \mu(a) = \mu(a')$. It can be easily verified that this is achieved by choosing only the uniform distribution: $\mathcal{E} = \mathcal{E}_{\text{di}} := \{\upsilon_{\mathcal{A}}\}$. Intuitively, here the DIB problem, which we denote by DIB_{di}, preserves (partially or wholly) the divergence $D(\mu(A)||\upsilon_{\mathcal{A}})$ of $\mu(A)$ from the uniform distribution: i.e., it preserves the "degree to which $\mu(A)$ is deterministic".

Theorem 2.3.6. *The following holds, for* $\Lambda := D(\mu||\mathcal{E}_{\text{di}})$ *and all* $\Phi \in \text{Bij}(\mathcal{A})$*:*

(i) Let $\kappa \in \text{DIB}_{\text{di}}(\Lambda)$*. Then* $\Phi \in \mathcal{G}_{\text{di}}$ *if and only if* $\kappa \circ \Phi = \kappa$*.*

(ii) If $\Phi \in \mathcal{G}_{\text{ci}}$*, then* $\kappa \circ \Phi = \kappa$ *also holds for all* $0 \leq \lambda \leq \Lambda$ *and* $\kappa \in \text{DIB}_{\text{di}}(\lambda)$*.*

(iii) The projection pr *defined by* $\sim$ *coincides with the projection on orbits* pr_{di}*.*

The proof relies on Theorem 2.3.1 (see Appendix B.3.7). Interpretations of points (i) and (ii) are analogous to those for equivariances. Point (iii) highlights that here, pr_{di} and pr do coincide. Eventually, one can directly adapt the definition of soft equivariances to one for soft distribution invariances.

2.3.4 Relevant computational and conceptual tools

Here, we present an algorithm approximating solutions to the DIB (2.3.1) for unconstrained shape of compression channels, and two relevant concepts. Consider the Lagrangian relaxation of (2.3.1),

$$\arg\min_{\kappa \in C} \Big[I_{\kappa}(A;T) - \beta D(\kappa \cdot \mu || \kappa \cdot \tilde{\mu}) \Big], \tag{2.3.7}$$

where $\beta \geq 0$. For no constraints on the shape of compression channels, i.e., for $\mathscr{K}_{\text{shape}} = \mathscr{K}(\mathcal{A}, \mathcal{T})$, deriving the Lagrangian above w.r.t. κ yields a fixed-point equation that all local minimisers must satisfy (similarly as for the classic IB). From this fixed-point equation, we obtain a Blahut-Arimoto-like (BA) algorithm with the same guarantees as BA for the classic IB (Tishby et al., 2000) (see Appendices B.4.1 and B.4.2).[4] In the following, we will write κ_β for the output of the BA algorithm (i.e., a local minimser) with parameter β, and also $I_\beta := I_{\kappa_\beta}(A;T)$ and $D_\beta := D(\kappa_\beta \cdot p || \kappa_\beta \cdot \tilde{\mu})$. Note that both I_β and D_β increase with β.

Now the *effective cardinality* (Zaslavsky et al., 2019) of some $\kappa \in \text{DIB}(\lambda)$ is defined as the minimum number of symbols t necessary to describe the output of κ (see Appendix B.4.3 for a formal definition). In all our numerical experiments, we observed that: (i) similarly as for the classic IB, effective cardinality monotically increases with β, and (ii) changes of effective cardinality coincide with discontinuities in the slope of the curve $\beta \mapsto (I_\beta, D_\beta)$, which is reminiscent of the second-order bifurcations observed for the IB (Zaslavsky et al., 2019). We will thus here refer to changes of effective cardinalities as bifurcations.

Eventually, we want to investigate whether the equation $\kappa_\beta \circ \Phi = \kappa_\beta$ is satisfied for varying β and varying $\Phi \in \mathcal{G}$, with $\mathcal{G}$ some fixed subgroup of $\text{Bij}(\mathcal{A})$. But numerically, it is also

[4] The convergence speed (and thus computational feasability) of this and other BA algorithms is difficult to estimate — particularly as they exhibit critical slowing down near bifurcations (Agmon et al., 2021).

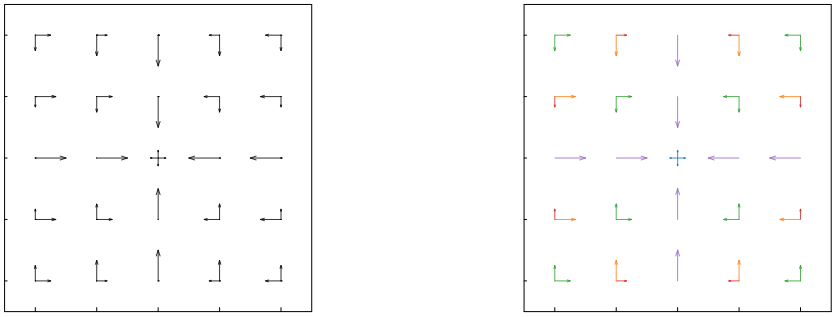


FIGURE 2.1: Left: representation of $\mu(Y|X)$, where X is the position on the grid, Y the gradient direction, and probabilities are proportional to arrow lengths. Thus equivariances are here pairs (ϕ, ψ) that send each arrow on an arrow of equal length. Right: same figure with colors representing a clustering of $\text{supp}(\mu(X,Y))$ — which defines a clustering of $\mathcal{X}\times\mathcal{Y}$ if we add the cluster $\text{supp}(\mu(X,Y))^{\complement}$, made of position-orientation pairs with probability 0, i.e., with no arrow. The latter clustering is obtained in 2 distinct ways: (i) as the projection on orbits of the equivariance group of $\mu(Y|X)$, and (ii) as the clustering defined by relation (2.3.6).

important, when this equation is not exactly satisfied, to quantify the extent of the deviation. We propose to use the divergence defined for all channel $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$ as

$$D_\mu(\kappa||\mathscr{K}_\mathcal{G}) := \min_{\nu\in\mathscr{K}_\mathcal{G}} D_\mu(\kappa||\nu) := \min_{\nu\in\mathscr{K}_\mathcal{G}} \sum_{a\in\text{supp}(\mu(A))} \mu(a) D(\kappa(T|a)||\nu(T|a)),$$

where $\mathscr{K}_\mathcal{G} := \{\nu : \forall\Phi \in \mathcal{G}, \nu\circ\Phi = \nu\}$ is the family of channels that are exactly input-symmetric w.r.t $\mathcal{G}$. Intuitively, $D_\mu(\kappa||\mathscr{K}_\mathcal{G})$ measures the divergence of the channel κ from being input-symmetric for the action of $\mathcal{G}$ on the distribution $\mu(A)$. In particular, $D_\mu(\kappa||\mathscr{K}_\mathcal{G}) = 0$ if and only if $\kappa\circ\Phi = \kappa$ for all $\Phi \in \mathcal{G}$. See Appendix B.4.4 for more details.

2.4 Synthetic numerical experiments on equivariances

The concept of soft λ-equivariance from Section 2.3.2 is motivated by the case of full information preservation $\lambda = \Lambda$, where our new definition of λ-equivariances coincides with that of classic group equivariances. However, it is not a priori clear that, once $\lambda < \Lambda$, our generalisation is consistent with the intuition of approximate symmetry. Here, we provide a sanity check in this direction, in a simple synthetic grid-world scenario. We start from an exactly equivariant channel $\mu(Y|X)$, which we synthetically perturb in such a way to render some of its equivariances more perturbed than others. We then expect that (i) all the perturbed equivariances should be recovered as soft λ-equivariances by the DIB_{ce} framework, once the reduction of the parameter λ enforces sufficient compression, and (ii) more perturbed equivariances should need more compression before being recovered as soft equivariances — because it means that more noise needs to be "overlooked".

We compute the DIB_{ce} solutions with the BA algorithm described in Appendix B.4.2, combined with reverse deterministic annealing, starting from $T = (X, Y)$ for large β (similarly as for the classic IB in (Zaslavsky et al., 2019)). All our experiments were simulated on a PC having 32GB of RAM and a 2.3GHz 12th generation i7 CPU.

Here, $\mathcal{X}$ stands for positions on a 5×5 grid, and $\mathcal{Y}$ for a gradient with 4 possible directions. Thus $\mu(Y|X)$ describes the probability of a direction at a given position, which can be thought of, e.g., as a nutrient gradient sensed by a bacteria. We choose uniform $\mu(X)$ (choosing non-uniform $\mu(X)$ resulted in similar results). As seen in Figure 2.1, left, $\mu(Y|X)$ has many

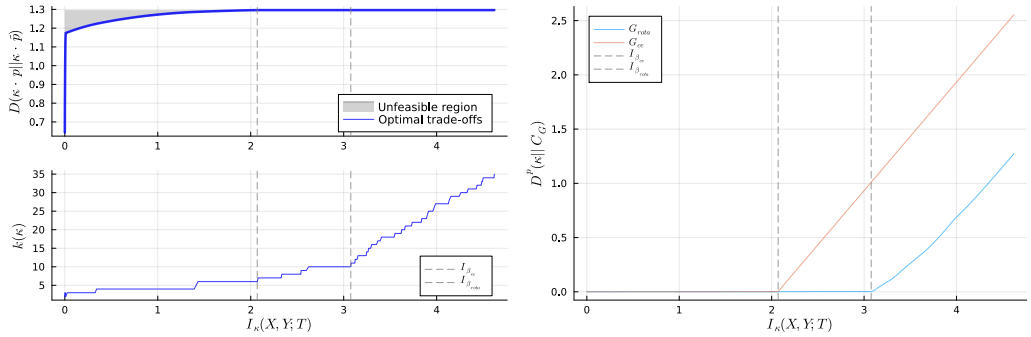


FIGURE 2.2: $D_\beta := D(\kappa_\beta \cdot p || \kappa_\beta \cdot \tilde{\mu})$ as a function of $I_\beta := I_{\kappa_\beta}(X,Y;T)$. Bottom left: Effective cardinality $k(\kappa)$ as a function of I_β. Right: Divergence of compression channels κ_β as a function of I_β, for the groups G_{ce} and G_{rota}. The vertical dashed lines represent specific bifurcations of the parameter β at which $D_\mu(\kappa_\beta || \mathscr{K}_{G_{\text{rota}}})$, resp. $D_\mu(\kappa_\beta || \mathscr{K}_{\mathcal{G}_{\text{ce}}})$, approximately vanishes (in decreasing order of I_β).

symmetries: it can be verified that the equivariance group of $\mu(Y|X)$ has 6 distinct orbits (one is $\text{supp}(\mu)^{\complement}$), represented in Figure 2.1, right. Moreover, even though we saw in Theorem 2.3.4, point (iii), that the projection on orbits pr_{ce} and the clustering pr defined by relation (2.3.6) do not generally coincide, here they do coincide. Thus Figure 2.1, right, also represents pr.

From Section 2.3.4, we have $D_\mu(\kappa_\beta || \mathscr{K}_{\mathcal{G}_{\text{ce}}}) = 0$ if and only if $\kappa_\beta \circ (\phi, \psi) = \kappa_\beta$ for all $(\phi, \psi) \in \mathcal{G}_{\text{ce}}$. Theorem 2.3.4, point (ii), suggests that this equation may indeed hold for all β.[5] As a sanity check, we thus computed the DIB_{ce} bottlenecks κ_β for $0 \leq D_\beta \leq \Lambda$, and indeed obtained $D_\mu(\kappa_\beta || \mathscr{K}_{\mathcal{G}_{\text{ce}}}) \leq 3 \cdot 10^{-16}$ for all β. We also noted that the bottlenecks' effective cardinality monotonically increases from 1 for $D_\beta = 0$ to 6 for $D_\beta = \Lambda$.

We then perturb $\mu(Y|X)$ with two random perturbations. The first one, of larger amplitude, breaks some equivariances in $\mathcal{G}_{\text{ce}}$, but not all of them. More precisely, after the perturbation, $\mu(Y|X)$ still satisfies the equivariances from its subgroup $G_{\text{rota}} \subsetneq \mathcal{G}_{\text{ce}}$ generated by rotating both the positions and the gradient directions by 90 degrees. The second perturbation applied to $\mu(Y|X)$, of smaller amplitude, breaks all the remaining equivariances from G_{rota}. We thus obtain a new $\mu(Y|X)$ which, intuitively, is still "approximately" equivariant, but where the approximate equivariances in $\mathcal{G}_{\text{ce}} \setminus G_{\text{rota}}$ are *coarser* than those in G_{rota}, because the perturbation was larger for the former than for the latter.

We compute 1000 DIB_{ce}-bottlenecks for varying β (which took 339 seconds). The resulting information curve (I_β, D_β), along with the corresponding effective cardinalities, are shown in Figure 2.2, left. As for the classic IB, we obtain a non-decreasing and concave information curve, and an increasing effective cardinality (except for small I_β, which could be due to numerical errors).

Crucially, we then observe (see Figure 2.2, right) that for decreasing β, the divergences $D_\mu(\kappa_\beta || \mathscr{K}_{G_{\text{rota}}})$ and $D_\mu(\kappa_\beta || \mathscr{K}_{\mathcal{G}_{\text{ce}}})$ successively vanish, at bifurcation values β_{rota}, resp. $\beta_{\text{ce}} < \beta_{\text{rota}}$. Thus the perturbed equivariances are here recovered by the DIB_{ce} method as *soft* equivariances, for large enough compression. Moreover, as the equivariances from G_{rota} have been less perturbed that those in the remaining of G_{ce}, here the degree of compression required to recover an approximate equivariance scales with the "coarseness" of that equivariance.

Eventually, note that, in Figure 2.2, left, the gain in divergence D_β from $I_\beta = I_{\beta_{\text{ce}}}$ to the maximum value $I_{\max}$ of I_β is negligible, whereas $I_{\max} - I_{\beta_{\text{ce}}}$ is large. This resonates with

[5] Here, the full support assumption, which is required in Theorem 2.3.4, does not hold for $\mu(X,Y)$. We leave to future work the theoretical study of the case $\text{supp}(\mu(X,Y)) \subsetneq \mathcal{X} \times \mathcal{Y}$.

the intuition that underlying symmetries in raw data afford a potentially drastic informational compression (I_β here), under a negligible loss in informational accuracy (D_β here).

2.5 Discussion

2.5.1 Summary

We started from an explicit formalisation of the idea that the IB implictily extracts invariances. I.e., we showed that for full preservation of the information that the source X carries about the relevancy variable Y, the IB compression channels coincide with the projection on orbits of the invariance group of $\mu(Y|X)$ — and that for lower granularity, the compression channel is always a post-processing of this projection on orbits. This motivated us to search for a similar information-theoretic reformulation of the *equivariances* of a channel $\mu(Y|X)$. As in this case, the projection on orbits compresses X and Y *jointly*, the bottleneck channel must also do so — while, on the other hand, it seems clear that the information-theoretic quantity to be preserved should depend only on (X, Y). We thus had to discard the notions of "source" and "relevancy" variables, and design a framework that goes beyond the preservation of mutual information. The new preserved quantity is, rather, a divergence beween the respective projections, through the compression channel, of the data distribution and a well-chosen exponential family. For equivariances, this results in a generalised IB problem whose solutions, informally, trade-off compression with the preservation of the information carried by the channel $\mu(Y|X)$ about the system (X, Y).

Moreover, the trade-off between compression and divergence preservation can actually be formulated for arbitrary exponential families on the probability simplex of an arbitrary finite alphabet. This yields a highly versatile Divergence IB framework, which can be interpreted — when the exponential family is a hierarchical model — as implementing compressions preserving the complexity of a given set of stochastic interdependencies (Ay et al., 2011). Therefore, this framework suggests a principled route to capture the data's underlying symmetries through *complexity-preserving coarse-grainings* of it, thus exposing the data's "platonic core", so to say. Crucially, expressing symmetries through such IB-like trade-offs yields a natural softening of these stringent group-theoretic notions.

Eventually, this new bridge between probabilistic symmetries and the language of information theory is also a bridge with information-theoretic models of adaptive behaviour. In particular, the Divergence IB framework could be specified to equivariances in Markov Decision Processes, e.g., of the kind defined in equation (1.2.1). This would make our framework directly relevant to the study of sensorimotor perception, and, specifically, to what has been referred to as *apparatus-related sensorimotor contingencies* — for a detailed discussion of this point, see the subsection "Representation learning, reinforcement learning & apparatus-related SMCs" in Section 1.2.3.

2.5.2 Limitations

The current framework, however, has important limitations. First, our core results are of theoretical nature, and hold in the discrete and full support case. At this stage, it is still unclear whether and how they extend to continuous and non fully supported distributions. Numerically, the BA class of algorithms addresses only the discrete case and generally scales unfavourably in larger scenarios. Future work could make the DIB problem amenable to deep network optimisation by adapting the classic IB's variational bounds (Alemi et al., 2017). The link between our soft equivariances and other proposals of generalised equivariances (Ashman et al., 2024; Romero et al., 2022; Song et al., 2023; Wang et al., 2022b) also deserves clarification.

More fundamentally, let us recall that we started from the intuition that a given group's projection on orbits can be seen as an informational compression, which motivated us to search for IB-like problems capturing these projections as their solution — for specific groups of channel and distribution symmetries. This was achieved quite straightforwardly for the classic IB and channel invariances (see Theorem 2.2.3). But for equivariances, the situation is more subtle: even though the solutions to our DIB_{ce} problem (for maximum parameter $\lambda = \Lambda$) do characterise group equivariances, the corresponding projection pr does not necessarily coincide with the projection on orbits under the equivariance group's action (see Theorem 2.3.4, points (i) and (iii)). More precisely, the projection pr "overshoots" the projection on orbits, in the sense that the equivalence relation $\sim$ defining pr (see equation (2.3.6)) induces a potentially *coarser* partition than the partition in orbits. This is because for (x, y) and (x', y') to be in the same orbit, we must have $(x', y') = (\phi \cdot x, \psi \cdot y)$ with (ϕ, ψ) such that the relation $(x'', y'') \sim (\phi \cdot x'', \psi \cdot y'')$ holds *for any pair* (x'', y'') (see Appendix B.3.4).

The reason for this problem thus seems to be that we disregarded the *transformations* of $\mathcal{X} \times \mathcal{Y}$ that induce the projection on orbit. This approach leads to another key limitation of our framework, which holds both for invariances and equivariances: whether or not the compression κ coincides with the projection on orbits, it is unclear how to leverage it to recover the symmetry group itself. In other words, while the focus on the projection on orbits is the crucial step that brings information parsimony into the picture, we might have thrown the baby out with the bathwater by defining our DIB compressions with no reference, whatsoever, to transformations of the data space.

This suggests that the "compression" and "transformation" point of views should rather be considered *jointly*: i.e., that we should formulate information-theoretic optimisation problems over both the compression channel κ *and* a channel ρ describing transformations of $\mathcal{X}\times\mathcal{Y}$. Let us now briefly outline how such a "transformation-based" bottleneck problem could look like in the case of equivariances. While we will dive back into a limited amount of mathematical details, our point is not to present a full-fledge, explicit extension of the DIB framework, but to identify the main stumbling block that prevents us from doing so.

2.5.3 Towards co-discovery of transformations and corresponding invariants

Here, we want to jointly discover both a compression channel κ capturing the projection on orbits of the equivariance group $\mathcal{G}_{\text{ce}}$, and a channel ρ capturing the action of this group on $\mathcal{X} \times \mathcal{Y}$. As each equivariance has the *split* form $\phi \otimes \psi$, it seems natural to require that ρ defines transformations with a similar form. To formalise this, let us consider a finite set $\mathcal{G}$ of transformation labels, and for each $\rho \in \mathscr{K}(\mathcal{X} \times \mathcal{Y} \times \mathcal{G}, \mathcal{X} \times \mathcal{Y})$ and $g \in \mathcal{G}$, let us denote by $\rho_g \in \mathscr{K}(\mathcal{X} \times \mathcal{Y})$ the channel defined by $\rho_g(x', y'|x, y) := \rho(x', y'|x, y, g)$. We then define

$$\mathscr{K}^{\otimes}_{\mathcal{G}}(\mathcal{X} \times \mathcal{Y}) := \left\{ \begin{array}{l} \rho \in \mathscr{K}(\mathcal{X} \times \mathcal{Y} \times \mathcal{G},\, \mathcal{X} \times \mathcal{Y}) : \\ \forall g \in \mathcal{G},\, \exists \phi_g \in \mathscr{K}(\mathcal{X}),\, \exists \psi_g \in \mathscr{K}(\mathcal{Y}) :\; \rho_g = \phi_g \otimes \psi_g \end{array} \right\},$$

i.e., the above set defines collections ρ of split stochastic transformations ρ_g of $\mathcal{X} \times \mathcal{Y}$.

Let us then set ourselves the following goal: defining an information-theoretic problem similar to the instance of the DIB problem (2.3.1) corresponding to equivariances, but where we now optimise over both compression channels $\kappa \in \mathscr{K}(\mathcal{X} \times \mathcal{Y}, \mathcal{T})$ and "split action" channels $\rho \in \mathscr{K}^{\otimes}_{\mathcal{G}}(\mathcal{X} \times \mathcal{Y})$, such that, for trade-off parameters requiring "full information preservation" (in a sense to be defined), the solutions (κ, ρ) satisfy the following:

(i) As for the DIB_{ce} problem, we have $D(\kappa \cdot \mu || \kappa \cdot \mathcal{E}_{\text{ce}}) = D(\mu || \mathcal{E}_{\text{ce}})$,

(ii) For all $g \in \mathcal{G}$, the channels ϕ_g and ψ_g such that $\rho_g = \phi_g \otimes \psi_g$ are both bijections,

(*iii*) κ coincides (up composition by congruent channels at the output) with the projection on orbits of the group generated (through composition) by the bijections $(\rho_g)_{g\in\mathcal{G}}$.

As far as the compression channel $\kappa \in \mathscr{K}(\mathcal{X}\times\mathcal{Y},\mathcal{T})$ is concerned, the only constraint that was added, as compared to the problem $\mathrm{DIB}_{\mathrm{ce}}(\Lambda)$, is that κ must be the projection on orbits of a group of transformations of the form $\phi_g \otimes \psi_g$ where $\phi_g \in \mathrm{Bij}(\mathcal{X})$, $\psi_g \in \mathrm{Bij}(\mathcal{Y})$. We expect this additional constraint to make the partition of $\mathcal{X}\times\mathcal{Y}$ defined by κ coincide exactly with the projection on orbits of $\mathcal{G}_{\mathrm{ce}}$, instead of "overshooting it" as before (see Section 2.5.2). On the other hand, while the above conditions do not guarantee that $(\phi_g \otimes \psi_g)_{g\in\mathcal{G}}$ describes the whole group $\mathcal{G}_{\mathrm{ce}}$, we expect them to require that $(\phi_g \otimes \psi_g)_{g\in\mathcal{G}}$ generates at least a subgroup of $\mathcal{G}_{\mathrm{ce}}$ that has the same projections on orbits as $\mathcal{G}_{\mathrm{ce}}$. While these arguments do need to be made explicit, here we only mention them to suggest that the above goal is valuable. So let us rather focus on how to translate conditions (*i*) to (*iii*) into information-theoretic constraints.

For condition (*i*), we will use the same constraint as for the $\mathrm{DIB}_{\mathrm{ce}}$ problem — see equation (2.3.1) and Section 2.3.2. Moreover, condition (*ii*) is relatively straightforward to characterise information-theoretically. Indeed, if we define the joint law $q_\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}\times\mathcal{G}\times\mathcal{X}\times\mathcal{Y}}$ by, say,

$$q_\rho(x,y,g,x',y') := \mu(x,y)\frac{1}{|\mathcal{G}|}\rho(x',y'|x,y,g), \tag{2.5.1}$$

and denote by $I_\rho(X,Y;X',Y'|G)$ the corresponding conditional mutual information, then:

Proposition 2.5.1. *Assume that $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ is full-support and $\rho \in \mathscr{K}^{\otimes}_{\mathcal{G}}(\mathcal{X}\times\mathcal{Y})$. Then*

$$I_\rho(X,Y;X',Y'|G) \le H(X,Y)$$

with equality if and only if for all $g\in\mathcal{G}$, the channels $\phi_g \in \mathscr{K}(\mathcal{X})$ and $\psi_g \in \mathscr{K}(\mathcal{Y})$ such that $\rho_g = \phi_g \otimes \psi_g$ are both defined by a bijective function.

Proof. See Appendix B.5. □

Interestingly, the converse condition $I_\rho(X,Y;X',Y'|G) = 0$ means that we have the Markov chain $(X,Y) - G - (X',Y')$, i.e., that the output of each stochastic transformation ρ_g does not depend on its input. The quantity $I_\rho(X,Y;X',Y'|G)$ thus parametrises the amount to which, on average, the output of stochastic transformations ρ_g depends on their input.

Overall, Proposition (2.5.1) and the discussion above suggest to consider a problem of the form

$$\underset{\substack{\kappa\in\mathscr{K}(\mathcal{X}\times\mathcal{Y},\mathcal{T}),\,\rho\in\mathscr{K}^{\otimes}_{\mathcal{G}}(\mathcal{X}\times\mathcal{Y})\\ D(\kappa\cdot\mu||\kappa\cdot\mathcal{E}_{\mathrm{ce}})\ge\lambda_1\\ I_\rho(X,Y;X',Y'|G)\ge\lambda_2\\ \boldsymbol{F}(\mu,\kappa,\rho)\ge\lambda_3}}{\arg\min} I_\kappa(X,Y;T), \tag{2.5.2}$$

where $\lambda_1 \in [0, D(\mu||\mathcal{E}_{\mathrm{ce}})]$, $\lambda_2 \in [0, H(X,Y)]$, and $\lambda_3 \in [0,\Lambda_3]$ for some $\Lambda_3 > 0$ to be determined. Crucially, here $\boldsymbol{F}$ should be an information-theoretic functional quantifying "how much" the point (*iii*) above is satisfied. The remaining question being: how should we define $\boldsymbol{F}$? Which kind of "information parsimony" parametrises the amount to which κ is the "projection on orbits" corresponding to ρ? I.e., we want the constraint $\boldsymbol{F}(\mu,\kappa,\rho)\ge\lambda_3$ to *bind* the compression channel κ and the action channel ρ in a such way that κ captures the "invariants" of the transformations implemented by ρ — where smaller parameter λ_3 would correspond to "softer" invariants of ρ.

This question goes beyond our current focus on equivariances of stochastic channels: it is of crucial importance for the broader program of formalising the duality between symmetry

and information parsimony (see Chapter 1). It will be addressed in the next chapter, where we will reframe the partition in orbits as a *decomposition into ergodic components*. Our investigations will also take a new turn, as we will consider, besides the bottleneck/orbit variable, a second "complementary" variable which "parsimoniously" tracks the changes induced by the action.

Chapter 3

Minimal Class-Pose Parametrisation in Markov Decision Processes

3.1 Introduction

3.1.1 Capturing the structure of changes

In Chapter 2, we focused on a specific family of group symmetries — mostly, invariances and equivariances of stochastic channels — and reformulated them in an information-theoretic language. But this effort ended up stumbling upon an unresolved question that does not depend on these specific symmetries (see Section 2.5.3): how to generalise the notion of projection on orbits — which requires a *group* action — to the action of stochastic transformations, and how to capture such generalised projections as solutions to an IB-like problem? These questions will be answered in this chapter (in resp. Sections 3.4.2 and 3.6.1). However, even before starting to answer them, they suggest a new round of unresolved questions.

Let us recall that our general approach is to use group theory as a scaffold to design incrementally "softer" notions of structure — better fitted to real-world adaptive behaviour — while retaining the essence of what makes the group-theoretic treatment interesting (see Section 1.3.1). Here, we want to generalise the notion of projection on orbits of a group action. As this requires generalising the notion of group action itself, this leads to the question: what should we retain from the latter? I.e., *what has the concept of group action to essentially offer to adaptive behaviour* — in particular, to the study of sensorimotor perception — and *which aspects of the formalism are nothing but the trace of its use by generations of scientists with their own, unrelated aims*? This question has of course many possible answers, depending on the use case. As we saw in Section 1.2.3, an important distinction to keep in mind is whether we are using the concept of group action to model *abstract symmetries* of a given system (which might or might not include a space modeling an agent's actions), or to model (a subset of) the *agent's own actions*. If we are dealing with abstract symmetries, the transformations of the symmetry group and the corresponding invariants — i.e., orbits — might well be all that we really care about: this was our approach in Chapter 2, where the problem was, ultimately, to discover these symmetries.

In contrast, in the case where the group action models (a subset of) an agent's actions, the group action is what we start from, and the problem is to analyse the perceptually relevant structure that it induces. Orbits are part of that perceptually relevant structure, as they can be interpreted, to a certain extent, as capturing object-related SMCs, in the sense of (O'Regan et al., 2001). Even though this idea was already reviewed in Section 1.2.3, let us briefly recall some formal examples, inspired from vision. First, if the surfaces of 3D rigid objects are modeled as closed surfaces in $\mathbb{R}^3$, then their shape can be captured as an invariant under rigid transformations (Marchetti et al., 2023): i.e., here we define the shape of a surface as the equivalence class of all surfaces that it can be transformed into through rigid transformations (this example will be treated in detail in Section 3.1.2). If we augment the group of rigid

transformations to that of affine transformations, we obtain a scale-invariant notion of shape.[1] Closer to the modeling of real-world agents with a visual apparatus, a groundbreaking result has shown that the segmentation of distinct contiguous surfaces can be achieved, in short, by identifying invariants under changes of perspective in the space of light rays itself (Tsao et al., 2022).[2]

This suggests that, under the (drastically) simplifying assumption that the effect of an agent's action on its own sensory space can be formalised as a group action, the corresponding orbits capture a fundamental dimension of sensorimotor perception. However, crucially, this does not formalise adequately the notion that perception is based on the exploration of the *structure of changes* induced by actions on the sensory space — the main claim of SMC theory, non-formalised in its original formulation (O'Regan et al., 2001), with subsequent work operationalising it in different directions (see Section 1.2). Indeed, here, the partition in orbits only describes what changes *cannot* be (the group action cannot send a point on a different orbit than its own). Our focus on orbits and information parsimony, however, suggests to complement this picture by searching for a *parsimonious description* of how the group action transforms *distinct* orbits: are these transformations inherently similar, irreconciliably different, or something in between? More precisely, can we formalise the "mastery" of an SMC (see Section 1.1.3) by deriving a group action on an "abstract" system that would simultaneously describe all the "concrete" actions on each orbit, while capturing as much as possible their common structure?

This chapter is primarily motivated by this agent-centric interpretation of group actions, and aims at a formalism capturing this "structure of changes" in a way that indeed encompasses all distinct orbits while capturing their common dynamical structure. As a starting point, let us see the partition in orbits as a coordinate, and search for a second, "complementary" coordinate, that would simultaneously describe, in the simplest way possible, the changes induced by the group action on each orbit. This would yield a special kind of *coordinatization* (Biehl et al., 2013) of the state-space that would be "adapted" to the group action.

For instance, consider the group $\mathcal{G}$ of rotations of any angle around the origin, on the infinite plane without the origin, i.e., $\mathcal{X} := \mathbb{R}^2 \setminus \{(0,0)\}$. A well-known decomposition of $\mathcal{X}$ into coordinates is "adapted" to this action: the *polar coordinates* $(r, \theta) \in]0, +\infty[\times \mathbb{S}$, where $\mathbb{S}$ is the unit circle. Here, the radius r fixes on which circle around the origin a given point lies — i.e., which *orbit* it belongs to, w.r.t. the action of $\mathcal{G}$ — while the angle θ determines on which half-line starting from the origin the point lies. Crucially, in polar coordinates, the action of a rotation g of angle θ_g on a point (r, θ) takes the simple form

$$g \cdot (r, \theta) = (r, g \cdot \theta), \tag{3.1.1}$$

where

$$g \cdot \theta := \theta + \theta_g. \tag{3.1.2}$$

I.e., both the coordinates r and θ define a *factor* of the action of $\mathcal{G}$ on $\mathcal{X}$ (see Definition A.0.5 in Chapter 1), where the action of $\mathcal{G}$ on r is trivial (i.e., it is the identity for all $g \in \mathcal{G}$), while its action on θ is non-trivial. More precisely, the latter action happens to be *isomorphic* to that of $\mathcal{G}$ on itself by multiplication on the left (see again Definition A.0.5 in Chapter 1). Indeed,

[1] Affine transformations also allow for reflections while rigid transformations do not, but this is not really relevant to our main point here.

[2] To be precise, the invariant classes defined in (Tsao et al., 2022) are based on a *pseudogroup* rather than an a group.

fix some $\theta_0 \in \mathbb{S}^1$ and, using the action from equation (3.1.2), define the map

$$\begin{aligned} \phi : \mathcal{G} &\to \mathbb{S}^1 \\ g &\mapsto g \cdot \theta_0. \end{aligned}$$

Then ϕ is clearly bijective, and it yields the equivariance relation $\phi(gg') = g \cdot \phi(g')$: i.e. ϕ induces an isomorphim between the action of the rotation group $\mathcal{G}$ on itself and its action on the unit circle $\mathbb{S}^1$. Eventually, while the coordinate r parametrises the partition in orbits, which is the finest partition in invariant subsets (see Proposition A.0.4 in Appendix A of Chapter 1), the coordinate θ does not capture any invariant of the group action: i.e., the only non-empty subset of $\mathbb{S}^1$ that is invariant under all rotations is $\mathbb{S}^1$ itself. In this sense, $\mathcal{X}$ decomposes into the "strictly invariant" coordinate r, defined by the projection on orbits, and the "strictly equivariant" coordinate θ, parametrised by a copy of the group itself — and such that the action of $\mathcal{G}$ on θ is isomorphic to the action of $\mathcal{G}$ on itself.

To which extent can this simple example be generalised, and how can the decomposition be discovered from "sensorimotor" data — interpreted here as samples of the effect of some transformations $g \in \mathcal{G}$ on some states $x \in \mathcal{X}$? It turns out that the ongoing line of work on *class-pose decomposition* (Marchetti et al., 2023; Oizumi et al., 2025; Pérez Rey et al., 2023; Winter et al., 2022) contributes, intentionally or not, to investigate this question. Indeed, even though the narrative underlying this framework is often not that of sensorimotor perception and always heavily representational, at the formal level, it resonates strongly with the notion of object-related SMC. This is exemplified, e.g., by the fact that the "class-pose" terminology is inspired by the above example of rigid transformations of rigid objects (Marchetti et al., 2023), which directly aligns with an example from the foundational paper (O'Regan et al., 2001) of SMC theory (see Section 1.2). Here, we take the mathematical object underlying this framework as a starting point, but generalise it to bring it closer to our aims. Namely, we want to:

(i) Understand the conditions under which class-pose decomposition is possible. This will show that it actually requires a highly non-generic assumption, which will lead to a more general structure which we call *minimal class-pose parametrisation*, where the group action on the class-pose space can now be "richer" than that on the state-space, but is also required to be "as simple as possible";

(ii) Generalise the group-theoretic setting into one that is inherently dynamical, and allows for non-invertible, stochastic, and closed-loop actions. This will bring the formalism closer to the modeling of embodied agents' natural behaviour;

(iii) Make the framework amenable to an information-theoretic treatment. This will exhibit a conceptual link between information and symmetry, provide a bridge with the information-theoretic analysis of embodied agents, and lay the groundwork for information theory-based discovery, from (sensorimotor) data, of the mathematical objects defined here.

We start, in Section 3.1.2 below, by addressing point (i) from a pure group-theoretic point of view, i.e., without any additional structure than set theory and the axioms defining group actions. This allows us to present the core idea of this generalisation while keeping the mathematical details as light as possible. We then suggest, in Section 3.1.3, interpretations of this new formal structure in terms of sensorimotor perception research. We will then be better equipped to explain, in Section 3.1.4, the limitations of the group-theoretic language — both from a dynamical and a sensorimotor perspective. This lays the ground for Section 3.1.5, which outlines the core content of this chapter, i.e., how we will generalise these group-theoretic ideas to a more flexible framework — thus addressing points (ii) and (iii) above.

3.1.2 From isomorphic decompositions to minimal joinings of the orbits

We consider an action

$$\begin{aligned} \mathcal{X} \times \mathcal{G} &\to \mathcal{X} \\ (x, g) &\mapsto \rho_g(x) = g \cdot x \end{aligned}$$

of a group $\mathcal{G}$ on a set $\mathcal{X}$, whose partition into orbits is denoted[3] by $\{\mathcal{X}^c\}_{c\in C}$ (see Appendix A for definitions). Our (set-theoretic) interpretation of class-pose decomposition is the following mathematical object:[4]

Definition 3.1.1. A *(set-theoretic) class-pose decomposition* of $\mathcal{X}$ w.r.t. ρ is a tuple (κ, θ, ξ) where $\kappa : \mathcal{X} \to \mathcal{C}$, $\theta : \mathcal{X} \to \mathcal{P}$ and $\xi : \mathcal{P} \times \mathcal{G} \to \mathcal{P}$ for some sets $\mathcal{C}$, $\mathcal{P}$, such that:

(*i*) The set of pre-images $\{\kappa^{-1}(c)\}_{c\in C}$ coincides with the partition in orbits $\{\mathcal{X}^c\}_{c\in C}$,

(*ii*) ξ is a group action of $\mathcal{G}$ on $\mathcal{P}$, and defining the corresponding action $\mathtt{Id}_C \otimes \xi$ of $\mathcal{G}$ on $\mathcal{C} \times \mathcal{P}$ as:

$$\forall g \in \mathcal{G}, \quad (\mathtt{Id}_C \otimes \xi)_g(c, p) := (\mathtt{Id}_C \otimes \xi_g)(c, p) := (c, \xi_g(p)), \tag{3.1.3}$$

then, for all $g \in \mathcal{G}$, we have $(\kappa, \theta) \circ \rho_g = (\mathtt{Id}_C \otimes \xi)_g \circ (\kappa, \theta)$; i.e., the following diagram commutes:

$$\begin{array}{ccc} \mathcal{X} & \xrightarrow{\rho_g} & \mathcal{X} \\ {\scriptstyle (\kappa,\theta)}\downarrow & & \downarrow{\scriptstyle (\kappa,\theta)} \\ \mathcal{C} \times \mathcal{P} & \xrightarrow{\mathtt{Id}_C \otimes \xi_g} & \mathcal{C} \times \mathcal{P} \end{array} \tag{3.1.4}$$

(*iii*) $(\kappa, \theta) : \mathcal{X} \to \mathcal{C} \times \mathcal{P}$ is a bijection.

In this section, we will often drop the adjective "set-theoretic", whose point is only to distinguish this definition from the measure-theoretic ones in further sections. Points (*ii*) and (*iii*) means that ρ and $\mathtt{Id}_C \otimes \xi$ are isomorphic, through the map (κ, θ) (see Definition A.0.5). Moreover, the group action $\mathtt{Id}_C \otimes \xi$ leaves the classes $c \in \mathcal{C}$ invariant but applies the non-trivial action ξ to the poses $p \in \mathcal{P}$. In that sense, a class-pose decomposition does "decompose" the group action ρ on $\mathcal{X}$ into its "strictly invariant" part (the identity $\mathtt{Id}_C$ on the "class" space $\mathcal{C}$) and its "strictly equivariant" part (the non-trivial action ξ on the "pose" space $\mathcal{P}$), obtaining along the way the change of coordinates (κ, θ) "revealing" this decomposition of the action ρ. Note that as (κ, θ) is a bijection, all class-pose pairs (c, p) in the full "rectangle" $\mathcal{C} \times \mathcal{P}$ correspond to a state $x \in \mathcal{X}$ — so that $\mathcal{C}$ and $\mathcal{P}$ can be seen as "independent" coordinates. Note also that our discussion on polar coordinates in Section 3.1.1 proves that they provide an example of class-pose decomposition (recall that we excluded the origin $(0, 0)$ from the state-space $\mathcal{X}$).

It is then natural to ask: under which conditions is such a decomposition possible? Previous work, when it makes its mathematical assumptions explicit, has mostly focused on what is known as *free* group actions (Marchetti et al., 2023; Oizumi et al., 2025), with (Pérez Rey et al., 2023) considering also some non-free cases. However, it turns out that the decomposition is only possible under a very stringent condition: if, in short, the group action is the same on each orbit.

[3] The distinction between c and $\mathcal{X}^c$ might at first seem unnecessary, but it will turn out to be a convenient notation.

[4] In (Oizumi et al., 2025), though, $\mathcal{X}$ is only *locally* decomposed, through the geometric structure of *principal bundle*.

To formalise this statement, note first that the action ρ of $\mathcal{G}$ on $\mathcal{X}$ restricts, for all $c \in \mathcal{C}$, to an action ρ^c of $\mathcal{G}$ on the orbit $\mathcal{X}^c$, i.e., $\rho^c_g(x) := \rho_g(x) \in \mathcal{X}^c$ for all $x \in \mathcal{X}^c$, $g \in \mathcal{G}$. Denote also by $\theta^c : \mathcal{X}^c \to \mathcal{P}$ the restriction of θ to each orbit $\mathcal{X}^c$. It can then be easily verified that given point (i) in Definition 3.1.1, points (ii) and (iii) can be replaced by the requirement that each θ^c is an isomorphism between ρ^c and ξ, i.e.:

$(ii)'$ For all $c \in \mathcal{C}$, $g \in \mathcal{G}$, we have $\theta^c \circ \rho^c_g = \xi_g \circ \theta^c$; i.e., the following diagram commutes:

$$\begin{array}{ccc} \mathcal{X}^c & \xrightarrow{\rho^c_g} & \mathcal{X}^c \\ {\scriptstyle\theta^c}\downarrow & & \downarrow{\scriptstyle\theta^c} \\ \mathcal{P} & \xrightarrow{\xi_g} & \mathcal{P} \end{array} \tag{3.1.5}$$

$(iii)'$ For all $c \in \mathcal{C}$, the map θ^c is a bijection between the orbit $\mathcal{X}^c$ and the pose space $\mathcal{P}$.

This reformulation will be convenient to prove the following characterisation:

Proposition 3.1.2. *There exists a (set-theoretic) class-pose decomposition w.r.t. ρ if and only if the restricted actions ρ^c and $\rho^{c'}$ are isomorphic for all $c, c' \in \mathcal{C}$.*

Proof. See Appendix C.1.1. □

While the requirement that orbits are pairwise isomorphic is more general than that of a free actions often considered in the literature (Marchetti et al., 2023; Oizumi et al., 2025), it is still a highly non-generic case. For instance, a simple counter-example is given by disjoint finite cycles. I.e., consider the sets $\mathcal{X}_m := \{1, \dots, m\}$ and $\mathcal{X}_n := \{1, \dots, n\}$; the action ρ_m, resp. ρ_n, of the cyclic group $\mathcal{G}_m$ of length m on $\mathcal{X}_m$, resp. the cyclic group $\mathcal{G}_n$ of length n on $\mathcal{X}_n$; and define the action ρ of $\mathcal{G} := \mathcal{G}_m \times \mathcal{G}_n$ on the state-space $\mathcal{X} := \mathcal{X}_m \sqcup \mathcal{X}_n$ as, for all $(g_m, g_n) \in \mathcal{G}$, $x \in \mathcal{X}$,

$$\rho_{(g_m,g_n)}(x) := \begin{cases} \rho_{g_m}(x) & \text{if } x \in \mathcal{X}_m, \\ \rho_{g_n}(x) & \text{if } x \in \mathcal{X}_n. \end{cases}$$

Clearly, the orbits of this action are $\mathcal{X}_m$ and $\mathcal{X}_n$, and the restrictions of the group action to resp. $\mathcal{X}_m$ and $\mathcal{X}_n$ are isomorphic to resp. $\mathcal{G}_m$ and $\mathcal{G}_n$. But for $m \neq n$, the latter groups are not isomorphic, which from Proposition 3.1.2 means that there is no class-pose decomposition.

This problem also applies to the rigid motions of rigid objects, which can be seen as a paradigmatic example of class-pose decomposition, as it even provides the metaphor inspiring the termininology "pose" (Marchetti et al., 2023). Let us formalise (the surface of) rigid 3D objects as the space $\mathcal{X}$ of closed topological surfaces — defined here as the 2-dimensional topological submanifolds of $\mathbb{R}^3$ that are compact and have no boundary. Consider the group $\mathcal{G}$ of rigid transformations: i.e., the group generated by rotations and translations in $\mathbb{R}^3$. This group induces an action ρ on closed topological surfaces, defined by applying the element $g \in \mathcal{G}$ to all points of the surface. The "class" or orbit of a surface under ρ is thus the set of all surfaces into which it can be transformed, which is here seen as a formalisation of the "shape" of the surface (i.e., the shape is precisely the information that does not depend on the position and orientation of the surface). But then, for instance, any rotation around the origin leaves the unit sphere $\mathbb{S}^2$ invariant; while there exists a surface $\mathcal{M}$ such that $g \cdot \mathcal{M} \neq \mathcal{M}$ for all rigid transformation $g \in \mathcal{G}$ (actually, this is the generic case: "most" objects are have no perfect symmetry[5]). This implies that the restriction of the action ρ to the orbit of $\mathcal{M}$

[5]Making this statement formal would derail us from our main point, which only requires the existence of one non-symmetric surface: take, e.g., a 3D version of the letter "L".

cannot be isomorphic to its restriction to the orbit of $\mathbb{S}^2$. Thus from Proposition 3.1.2, there is no class-pose decomposition in the sense of Definition 3.1.1. In other words, the latter definition provides a concept of class-pose decomposition that does not even "decompose" what is supposed to be a paradigmatic example.

Beyond these formal examples, it is overly restrictive, from a sensorimotor perception perspective, to require all classes (seen as corresponding to object-related SMCs) to behave exactly the same under the agent's actions (or a specific subset of it). To take a classic example from the SMC literature (O'Regan, 2011): assuming that the percept of a sponge consists in the dynamical process of "skillfully" handling it while receiving the corresponding sensory feedback, this process will not be exactly the same for all sponges (or even all interactions with a given sponge). Yet, there is a sense in which all these processes are specific instances of a broadly defined "sponge" percept/dynamical process — a "know-how" of how to hold and use a sponge. In our current group-theoretic setting, this suggests that we need a concept of pose that can not only capture the common structure of the restriction of the group action to distinct orbits, but also accomodate for variations across the orbits.[6]

How, then, should we generalise Definition 3.1.1, so as to also capture the case of non-isomorphic orbits? Let us come back to the example of polar coordinates. Maybe our generalisation of the angular coordinate θ was asking too much. Here, θ is a *single* angle setting the angles on *all circles (centered on 0) at the same time*, and the action of the rotation group on θ is enough to "simulate" the action of the same group on each individual circle. In other words, *the* pose (*singular*), and the action of $\mathcal{G}$ on it, can be seen as a "minimal collective description" of *all* poses (*plural*) on all classes, and the respective actions of $\mathcal{G}$ on them. In the case of polar coordinates, this "minimal collective description" turns out to be just as simple as the group action on any fixed individual orbit — in the sense that the action on the pose space is isomorphic to that on a fixed orbit, e.g., the unit circle. Definition 3.1.1 focuses on similar situations. But in general, it is natural to expect a "minimal way to collectively describe the simultaneous action on all orbits" to not be isomorphic to the action on a single orbit.

To formalise this more general situation, we propose to *"reverse the arrows and break the bijectivity"* in Diagram (3.1.5). I.e, for each $c \in \mathcal{C}$, instead of considering an isomorphism from ρ^c to ξ, we consider a *factor* from ξ to ρ^c (see Definition A.0.5 in Chapter 1). More precisely:

Definition 3.1.3. Let $(\rho^c)_{c\in\mathcal{C}}$ be family of group actions of the same group $\mathcal{G}$ on resp. sets $(\mathcal{X}^c)_{c\in\mathcal{C}}$. A *(set-theoretic) joining* of $(\rho^c)_{c\in\mathcal{C}}$ is a group action ξ of $\mathcal{G}$ on a set $\mathcal{P}$, together with a family of maps $(\phi^c)_{c\in\mathcal{C}}$ such that for all $c \in \mathcal{C}$, the action ρ^c on $\mathcal{X}^c$ is a factor of the action ξ on $\mathcal{P}$, with factor map $\phi^c : \mathcal{P} \to \mathcal{X}^c$; i.e., explicitly, ϕ^c is surjective and for all $g \in \mathcal{G}$, the following diagram is commutative:

$$\begin{array}{ccc} \mathcal{P} & \xrightarrow{\xi_g} & \mathcal{P} \\ {\scriptstyle \phi^c}\downarrow & & \downarrow{\scriptstyle \phi^c} \\ \mathcal{X}^c & \xrightarrow{\rho^c_g} & \mathcal{X}^c \end{array} \tag{3.1.6}$$

The family $(\phi^c)_{c\in\mathcal{C}}$ is called the joining's family of *marginalisation maps*. Alternatively, the map $\phi : \mathcal{C} \times \mathcal{P} \to \mathcal{X}$ defined by $\phi(c, p) := \phi^c(p)$ will be called the *marginalisation map* of the joining (ξ, ϕ).

Technical remark. Our terminology takes inspiration from ergodic theory's notion of joining (Glasner, 2003), which is usually stated in a measure-theoretic setting. Definition 3.1.3 can be

[6] Of course, the group setting does not capture the dynamical and closed-loop aspects of the sponge example — which we will focus on later. But it has the merit of isolating the problem of "common structure across SMCs".

seen as a set-theoretic version of it. While later sections will move on to a measure-theoretic definition as well, in this section, we will drop the qualificative "set-theoretic".

Here, we are interested in the case where each $\mathcal{X}^c$ is the orbit under the action ρ of $\mathcal{G}$ on $\mathcal{X}$, and $(\rho^c)_{c\in C}$ is the corresponding family of restrictions to the orbits. In this case, it easy to verify that an action ξ of $\mathcal{G}$ on $\mathcal{P}$ defines a joining of $(\rho^c)_{c\in C}$ with marginalisation map ϕ if and only if:

$(i)''$ For all $c \in \mathcal{C}$, the map $\phi^c : \mathcal{P} \to \mathcal{X}^c$ is surjective;

$(ii)''$ For all $g \in \mathcal{G}$, defining $\mathsf{Id}_C \otimes \xi_g$ as in (3.1.3), the following diagram is commutative:

$$\begin{array}{ccc} \mathcal{C}\times\mathcal{P} & \xrightarrow{\mathsf{Id}_C\otimes\xi_g} & \mathcal{C}\times\mathcal{P} \\ \downarrow{\scriptstyle\phi} & & \downarrow{\scriptstyle\phi} \\ \mathcal{X} & \xrightarrow{\rho_g} & \mathcal{X} \end{array}$$

The map ϕ can thus be seen as a parametrisation of $\mathcal{X}$ that, similarly as for class-pose decomposition, "separates" the invariant coordinate defined by the partition in orbits from a second, equivariant coordinate. With however the crucial difference that now, this parametrisation is not required to be isomorphic. For instance, a joining can always be obtained by "running all the orbits in parallel", i.e., by choosing $\mathcal{P} := \bigtimes_{c\in C} \mathcal{X}^c$ and for all $(x^c)_{c\in C} \in \mathcal{X}$, $g \in \mathcal{G}$, $c_0 \in \mathcal{C}$,

$$\begin{aligned} \xi_g((x^c)_{c\in C}) &:= (\rho_g(x^c))_{c\in C}, \\ \phi^{c_0}((x^c)_{c\in C}) &:= x^{c_0}. \end{aligned}$$

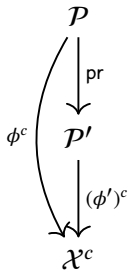
This joining is intuitively, the "least isomorphic" one: e.g., for the action of the rotation group on $\mathbb{R}^2$, it would correspond to keeping track of the angle on each circle with a different coordinate — while we saw that it is enough to keep track of a single number, the angular part θ of the polar coordinates. Here we are interested, on the contrary, in the "most isomorphic" joinings.[7] To formalise this intuition, we need a way of comparing joinings. It will be based on the following relation:

Definition 3.1.4. Let (ξ, ϕ) and (ξ', ϕ') be two joinings of a family of group actions $(\rho^c)_{c\in C}$, with resp. marginalisation maps ϕ and ϕ'. We say that the joining (ξ', ϕ') is a *joining factor*, or *j-factor* for short, of the joining (ξ, ϕ), if the group action ξ' is a factor of the group action ξ through a factor map pr such that for all $c \in \mathcal{C}$, we have $\phi^c = (\phi^c)' \circ \mathsf{pr}$. I.e., (ξ', ϕ') is a j-factor of (ξ, ϕ) with factor map pr if the diagram

$$\begin{array}{ccc} \mathcal{P} & \xrightarrow{\xi_g} & \mathcal{P} \\ \downarrow{\scriptstyle\mathsf{pr}} & & \downarrow{\scriptstyle\mathsf{pr}} \\ \mathcal{P}' & \xrightarrow{\xi'_g} & \mathcal{P}' \end{array} \tag{3.1.7}$$

[7] We will see in Section 3.6 that "most isomorphic" joinings can actually be seen as *"most parsimonious"* joinings, but for now we focus on the algebraic aspect.

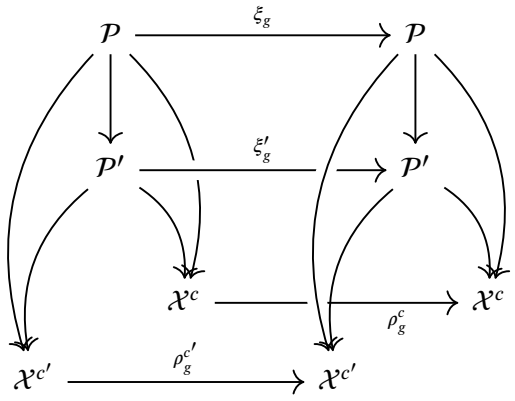
commutes for all $g \in \mathcal{G}$, and the diagram



(3.1.8)

commutes for all $c \in C$.

Informally, the joining (ρ', ϕ') is a j-factor of the joining (ρ, ϕ) if it is a "coarse-grained version" of the latter, with each corresponding marginalisation map $(\phi')^c$ being thus "more injective" than ϕ^c. This relation is better visualised when the family of group actions is made of only two elements. Indeed, if $C = \{c, c'\}$, it can be easily verified (using the definition of a joining) that (ρ', ϕ') is a j-factor of (ρ, ϕ) if and only if the following diagram commutes for all $g \in \mathcal{G}$:



where for simplicity, we ommited the labels of the marginalisation maps ϕ^c, $\phi^{c'}$, $(\phi')^c$, $(\phi')^{c'}$ and the projection map $\mathsf{pr} : \mathcal{P} \to \mathcal{P}'$.

Importantly, the j-factor relation yields a pre-order:

Proposition 3.1.5. *Let $(\rho^c)_{c\in C}$ be a family of actions of the same group $\mathcal{G}$ on resp. state spaces $(\mathcal{X}^c)_{c\in C}$. The j-factor relation between the joinings of $(\rho^c)_{c\in C}$ is a pre-order: i.e., it is reflexive and transitive.*

Proof. See Appendix C.1.2. □

We then formalise the intuition of a "maximally isomorphic" joining — with "maximally injective" marginalisation maps — through this pre-order:

Definition 3.1.6. A *(set-theoretic) minimal joining* of the group actions $(\rho^c)_{c\in C}$ is a joining (ξ_*, ϕ_*) of $(\rho^c)_{c\in C}$ that is minimal for the joining factor pre-order: i.e., such that (ξ_*, ϕ_*) is a j-factor of any other joining of $(\rho^c)_{c\in C}$. [8]

As noted in (de la Rue, 2006, 2023), the theory of joinings, a branch of ergodic theory, explores some kind of "arithmetics of dynamical systems" — and measurable group actions (Glasner, 2003). This point of view is particularly relevant to us, as the notion of minimal

[8] While we do not prove that set-theoretic minimal joinings always exist, we will do so for the finite case of their measure-theoretic counterpart (see Theorem 3.5.13).

joining is a generalisation of integers' *least common multiple* to group actions. Indeed, in the above example of disjoint finite cycles, it can be verified that the minimal joining of the action of $\mathcal{G}_m$ on $\mathcal{X}_m$ and that of $\mathcal{G}_n$ on $\mathcal{X}_n$ is a cycle with length the least common multiple between m and n. Surprisingly, however, the notion of minimal joining (in its set-theoretic formulation above, or its measure-theoretic one in later sections) has, to the best of our knowledge, not yet been considered in the literature.[9]

These considerations suggest to step back from class-pose decompositions, i.e., tuples (κ, θ, ξ) satisfying conditions (i), (ii), (iii) from Definition 3.1.1, and generalise them into *minimal joinings of the orbits*, i.e., pairs (ϕ, ξ) satisfying conditions $(i)''$, $(ii)''$ above, with the additional requirement that $(iii)''$ the joining of the orbits is minimal. Here, the "class" is still the orbit under the group action, but the pose coordinate now provides, intuitively, a "maximally isomorphic" group action that "simultaneouslly simulates" the restriction of the group actions to each orbit.

Note that the terminology "class-pose *decomposition*" does not fit this new notion. Indeed, the map ϕ is now only a surjective map from the class-pose space $\mathcal{C} \times \mathcal{P}$ to the state $\mathcal{X}$, which might not be injective. It is thus more appropriate to refer to minimal joinings of the orbits as *(set-theoretic) minimal class-pose parametrisations*.

3.1.3 Sensorimotor interpretation

The point of view outlined in Section 3.1.2 captures several related ideas that have been put forward by sensorimotor approaches to perception (see Section 1.1.3). On the one hand, SMC theory claims that perception can only unfold through the *mastery*, also called *know-how*, of sensorimotor contingencies that are relevant to an agent's behaviour, where percepts are reframed as specifics sets of (potential or actually enacted) actions and resulting sensations, to which brain dynamics have been *attuned* through development and learning (Buhrmann et al., 2013; O'Regan et al., 2001). On the other hand, in the inside-out approach to brain dynamics, "the main emphasis is on how the brain's outputs, reflected by the animal's actions, influence incoming signals"[10], thus leading to the emergence of meaningful percepts via the *calibration* of brain dynamics through sensorimotor interaction. More precisely, the claim is here that the concrete sensorimotor experiences become gradually *internalised*, so that, after development and learning, brain dynamics not only acquire meaning through the concrete sensorimotor interactions that they participate in, but, to a certain extent, they can also "simulate" these experiences without actually triggering the activity of the sensorimotor interface (Buzsáki et al., 2019).

If the state-space $\mathcal{X}$ is seen as an agent's sensory space and group actions as a model of how a specific subset of agent's own actions affects this sensory space (which is admitedly a very crude model), the minimal class-pose parametrisation defined above captures some aspects of both these frameworks, combined with a certain notion of parsimony. First, as argued in Section 1.2, each orbit can then be seen as a specific set of sensorimotor interactions, formalising the notion of object-related SMC — e.g., the shape of a given surface is here understood as all the possible ways in which it could be transformed into another surface through rigid transformations. The minimal joining of the orbits then becomes an *abstraction that collectively describes, in the simplest way possible, the set of all possible concrete sensorimotor interactions* under the given actions. If the action on the pose space is seen as corresponding to brain dynamics, the learning of the orbits' minimal joining thus corresponds to a process of

[9] The notion of *minimal self-joining* (de la Rue, 2006) is not directly related to minimal joinings as defined here.

[10] (Buzsáki et al., 2019), p. 21.

"calibration"/"attunement"[11] of brain dynamics to concrete sensorimotor interactions, where crucially, *different sensorimotor interactions are captured, as much as possible, through the same brain dynamics*. This induces those dynamics to capture the *common structure* across different sensorimotor interactions, which can be understood as the development of a certain *know-how* or *mastery* of these sensorimotor interactions.

Of course, this formalisation only touches upon certain aspects of these theories: e.g., it does not capture the full coupling between "calibrated"/"attuned" brain dynamics and the actually enacted movement on the time-scale of perception itself — while mounting experimental evidence suggests that such ongoing movement is a fundamental part of the perceptual process, and is often coupled to ongoing neural activity (e.g., *saccades* (Rolfs et al., 2025) and *fixational eye movements* (Rucci et al., 2015) in the case of vision, see Section 1.1.3). However, to address this crucial question — which will remain beyond the scope of this thesis — it is instrumental to first have a clearer concept of calibration/attunement (Buhrmann et al., 2014). We hope that our novel minimal joining perspective — in the group-theoretic form above or the richer MDP form presented further in this chapter — can help for this purpose.

3.1.4 Towards a more flexible framework

The new perspective outlined in Section 3.1.2 seems like a good candidate to address the limitation of class-pose decomposition identified by Proposition 3.1.2. However, even though the group-theoretic setting was convenient to address this purely algebraic limitation, minimal class-pose parametrisation as defined above remains in many ways ill-adapted to the adaptive behaviour aims presented in Section 3.1.1, and in particular, to the sensorimotor perception interpretation from Section 3.1.3. Here, we review some of these remaining limitations, which will naturally lead to the framework developed in the rest of this chapter.

First of all, while the concept of orbit seems to acurately capture the intuition of "class" in the examples from Section 3.1.2 and others from the class-pose decomposition literature (which often considers classic Lie groups), other examples cast doubt on the ultimate relevance of seeing classes as orbits. In particular, it is often ill-adapted to invertible discrete-time dynamical systems, which can be seen as actions of the group $\mathbb{Z}$ of integers (representing time shifts) on a given state-space. E.g., consider the group $\mathcal{G} = \{g^n, n \in \mathbb{Z}\}$ generated by a rotation g of irrational angle on the unit circle $\mathbb{S}^1$. It is well-known that for the action of $\mathcal{G}$ on $\mathbb{S}^1$, each orbit is dense (Coudène, 2016). But, as orbits define a partition of $\mathbb{S}^1$ and each orbit is countable while $\mathbb{S}^1$ is uncountable, there must be an uncountable number of orbits.[12] In this case, should a good notion of "class" really capture an uncontable number of equally dense classes, or should we rather have only one class, given by the whole circle? It seems natural to choose the second option; yet defining classes as orbits leads to the first one.

The limitation arising in this example is far from an edge case — as, among all planar rotations, rotations of irrational angle are clearly the generic case.[13] More generally, for any group action generated by an invertible transformation, one of the core features of dynamical systems is that orbits (i.e., here, bi-sided temporal trajectories) might asymptotically "approach" a given "attractor" without containing it. This suggests that a concept of class that has any hope to be relevant to dynamical systems should capture the dynamics' *asymptotic behaviour*.

[11] We are here lumping the concept of "calibration" from (Buzsáki et al., 2019) and that of "attunement" from (O'Regan et al., 2001) into a common one, as even though they are distinct, they seem to describe different aspects of a common phenomenon.

[12] Indeed: if there was a countable number of countable orbits, then $\mathbb{S}^1$ would be included in $\mathbb{N} \times \mathbb{N}$, which is in bijection with $\mathbb{N}$. But $\mathbb{S}^1$ is in bijection with $[0, 2\pi[$, which is in bijection with $\mathbb{R}$; and $|\mathbb{R}| > |\mathbb{N}|$.

[13] In the sense (*i*) that the set of rational angles is countable, while that of irrational angles is uncountable, or (*ii*) that the former has Lebesgue measure 0 on the circle, while the latter has measure 1.

It turns out that the concept of *ergodic component* does exactly that, and in addition — just like orbits — also provides a *partition* of the state-space (or a generic subset of it). Moreover, invertible dynamical systems' decomposition into ergodic components generalises to a large class of group actions,[14] including those usually considered in the class-pose decomposition literature (Marchetti et al., 2023; Oizumi et al., 2025; Pérez Rey et al., 2023; Winter et al., 2022).

However, more fundamentally, modeling (even a subset of) the agent's actions with a group action is a limitation in itself. In particular, realistic models of adaptive behaviour require non-invertible and often stochastic actions. Moreover, while ergodicists might find it helpful, for their own purposes, to abstract away the time evolution of dynamics into a group action (Glasner, 2003; Kerr et al., 2016), it is at this stage not clear at all that disregarding the arrow of time is any helpful to understand the structure of embodied agents' behaviour. It seems thus more adapted to replace the map $\rho : \mathcal{X} \times \mathcal{G} \to \mathcal{X}$ implementing the group action by an arbitrary stochastic channel from $\mathcal{X} \times \mathcal{G}$ to $\mathcal{X}$ that models the effect of the agent's action over a single time-step.[15] Eventually, crucially, we want to understand the structure of the actual, potentially *closed-loop* behaviour of the agent, rather than treating all the actions that it can possibly do equally. This suggests considering a *policy* π, i.e., a stochastic map from the state-space space $\mathcal{X}$ to the action space $\mathcal{G}$. All together, these considerations make *Markov Decision Processes* (MDPs) a desirable direction in which to generalise minimal class-pose parametrisation.[16] Note that here we are not interested in any kind of policy learning, but rather in the *structure induced by a given agent behaviour*: we will thus consider a fixed policy and no reward function.

The discrete case is of course relevant to these questions, but we are here aiming for a framework that also generalises actions of continuous groups on continuous spaces — which requires continuous MDP action and state spaces. To carry out the generalisations mentioned above, we thus need to do *ergodic theory on Markov Decision Processes defined on either discrete or continuous spaces*. This requires using a measure-theoretic language: more precisely, here, we will work with *standard Borel* spaces — which encompass countable spaces, Euclidean spaces, or differential manifolds, and have been extensively studied by ergodic theory both in the deterministic and stochastic cases (Coudène, 2016; Glasner, 2003; Worm et al., 2011).

We are ultimately interested in reformulating the minimal class-pose parametrisation framework in an information-theoretic language — similarly as, in the previous chapter, we reformulated channel invariances and equivariances with the Information Bottleneck (IB) and generalisations of it. Our motivations are here similar as well. First, even an MDP version of minimal class-pose parametrisation is unlikely to be found in real-world data. In particular, our working assumption is that, at the level of an embodied agent's sensorimotor interface, structure often only exists as "soft structure", i.e., as structure that only becomes salient under appropriate "simplifications" of that interface. Our aim is here to capture this "simplification" through trade-offs between information parsimony constraints and the preservation of other, well-chosen notions of "behaviourally relevant information". I.e., here, the class and pose coordinates would emerge from the "informational tension" induced by such trade-offs. Importantly, making these links explicit could provide a bridge with a rich line of work that investigates the principles of adaptive behaviour through the information-theoretic constraints that underlie it (Ay, 2015; Ay et al., 2012; Krakauer et al., 2020; Ortega et al., 2013; Pezzulo et al., 2024; Salge et al., 2014; Tishby et al., 2011). Eventually, if class and pose are solutions to informational trade-offs — i.e., if they define Pareto fronts of multi-objective optimisation

[14] E.g., for locally compact, second countable group acting on a standard Borel space (Greschonig et al., 2000).
[15] The notation $\mathcal{G}$ then stands for "generator" (of strings of stochastic actions) rather than "group".
[16] Generalisations to non-Markovian processes would be crucial, but will not be considered here.

problems with information measures as objectives — then this opens the way to the design of information theory-based algorithms discovering these structures from data. This is a crucial aspect, as to the best of our knowledge, there exists no ready-made algorithm to concretely compute the novel mathematical objects that we will define here.

The question that we will focus on in the rest of this chapter is thus the following:

Question. How to generalise minimal class-pose parametrisation to Markov Decision Processes on standard Borel spaces, and how to characterise it (at least in the finite case) in terms of information parsimony?

3.1.5 Plan for the rest of this chapter

We start, in Section 3.2, by introducing the measure-theoretic notions that we will rely on. In Section 3.3, we then present previous results on the decomposition into ergodic components of standard Borel Markov chains (Sections 3.3.1 and 3.3.2), which we fine-tune so as to obtain the exact form that we need (Section 3.3.3). The latter yields a partition of a "generic" (in a measure-theoretic sense) and invariant subset $\mathcal{X}_{\text{erg,inv}}$ into a family of invariant subsets $(\mathcal{X}^c)_{c\in C}$, each invariant and carrying a unique ergodic distribution ϵ^c. We also prove that, under a continuity assumption, the space of ergodic components C has itself a standard Borel structure (Section 3.3.4); and that in the finite case, the projection on ergodic components is a minimal sufficient statistic between the initial state and the time-average of the resulting trajectory (Section 3.3.5) — a fact that will be instrumental in the information-theoretic treatment in later sections.

Our fine-tuned version of the Markov chain result from (Worm et al., 2011) is leveraged, in Section 3.4.1, to obtain an *ergodic decomposition theorem for standard Borel MDPs*. More precisely, consider an MDP defined as a pair (π, ρ), where π is a policy (i.e., a channel from the state-space $\mathcal{X}$ to the action space $\mathcal{G}$) and ρ a transition channel (i.e., a channel from $\mathcal{X}\times\mathcal{G}$ to $\mathcal{X}$). We prove that for any stationary initial distribution, the resulting process distribution obtained using the MDP (π, ρ) can be decomposed as an integral over process distributions of MDPs $(\pi^c, \rho^c)_{c\in C}$ resp. defined on each ergodic component $\mathcal{X}^c$, and equiped with the corresponding ergodic initial distribution ϵ^c. As the decomposition depends on the policy π, this yields a notion of *behaviour-induced class* — a promising tool to capture the invariants of closed-loop behaviour in MDPs. In Section 3.4.2, we then specialise this result to measurable group actions of standard Borel groups that have a stationary probability (which includes the group actions usually considered in the class-pose decomposition literature). In particular, we show that, for an MDP naturally defined by the group action and corresponding stationary probability on the group, *the partition in orbits of the group action coincides with the MDP's decomposition into ergodic components*. Building upon the arguments from Section 3.1.4 above, we then show that for groups with no stationary probability, ergodic components provide, in general, a more natural concept of class than orbits, as they capture the *asymptotic properties* of the MDP's dynamics.

Once this generalisation of the class coordinate achieved, we turn, in Section 3.5, to the generalisation of the pose coordinate, i.e., of minimal joinings of orbits — see Section 3.1.2. A substantial amount of work is required to adapt to *stationary MDPs*[17] the notions of *factor* and *isomorphism* (Section 3.5.1), which lead to those of *joining* and, crucially, *minimal joining* (Section 3.5.2). We prove that for a finite number of stationary MDPs each with finite alphabets, there always exist minimal joinings. We then eventually define, in Section 3.5.3, our generalised version of minimal class-pose parametrisation: i.e., as a minimal joining of

[17] State and action spaces are here only assumed measurable. This is because some spaces will be uncountable Cartesian products, which are not standard Borel even if each coordinate is (see Prop. C.2.5 below).

the family of "ergodic" MDPs $(\epsilon^c, \pi^c, \rho^c)_{c\in C}$ obtained in Section 3.4.1. We prove, in Section 3.5.4, that for the same group actions as those considered in Section 3.4.2 and for a countable number of ergodic components, minimal class-pose parametrisation generalises (a measured space version of) the group-theoretic class-pose decomposition that we started from in Section 3.1.2. Importantly, however, just like in the latter section, measure-theoretic class-pose decompositions only exist in highly non-generic cases — while, in the finite case at least, the existence of minimal joinings implies that of minimal class-pose decompositions (of stationary MDPs, and in particular group actions).

While it might not seem a priori clear how the group-theoretic setting presented in Section 3.1.2 lends itself to an information-theoretic characterisation, the bridge between these two languages is provided precisely by our ergodic-theoretic reformulation from Sections 3.3 to 3.5. In Section 3.6, we thus exhibit — for finite systems — two information trade-offs that our generalised classes and poses are respectively solution to.

On the one hand, the projection on ergodic components is characterised, in Section 3.6.1, as an optimal compression of the state-space $\mathcal{X}$ under the contraint of preserving the *mutual information between the initial state and the time-average of the resulting trajectory*. When the MDP implements a group action, this yields an *information-theoretic characterisation of the projection on orbits*. — thus addressing the main limitation identified in Section 2.5.3. More generally, this novel characterisation of ergodic components is a stepping stone in the program of formalising the intuition of "duality" between symmetry and information parsimony that underlies this whole thesis. Indeed, it exhibits an informational trade-off that *binds* a compression channel and a corresponding set of transformation channels, in a way that requires the coarse-graining implemented by the compression channel to capture the transformations' invariants.

On the other hand, we show in Section 3.6.2 that minimal joinings coincide with *minimum entropy joinings* — or equivalently, *maximum multi-information joinings*. This provides an information-theoretic characterisation of the algebraic notion of minimal joining. Even though at this stage, the latter does not provide a full multi-objective information-theoretic optimisation problem, and we do not propose an algorithm to solve the one corresponding to ergodic components, our reformulations of classes and poses lay the groundwork for the information-theoretic agenda outlined in Section 3.1.4.

Let us stress that the results from this chapter are exclusively mathematical. Section 3.7 describes the steps that would be required for the computational implementation of this mathematical framework, as well as other limitations that should be addressed by future work. Eventually, we conclude and come back to the relevancy of our novel formalism to sensorimotor perception in Section 3.8. In particular, our stationary MDP version of minimal class-pose parametrisation formalises the intuition of a "*parsimonious fiction*" that makes sense of concrete sensorimotor interactions by capturing the "structure of change" common to all these interactions while still expressing the variations across them. This interpretation provides a new builing block for the formalisation of sensorimotor theories of perception, while also introducing a *conceptual* innovation on the notions of "attunement" (O'Regan et al., 2001) and "calibration" (Buzsáki et al., 2019).

3.2 Measure-theoretic setting

As explained in Section 3.1.4, we will here do ergodic theory on Markov Decision Processes (MDPs) with *standard Borel* state and action spaces — so as to encompass a broad array of examples (e.g., countable spaces, Euclidean spaces, differential manifolds), while relying on previous results of ergodic theory on standard Borel Markov chains (Worm et al., 2011). However, for the study of joinings, we will need to consider *uncountable* products of standard

Borel spaces — which are *not* standard Borel (see Proposition C.2.5 below). We will thus also need, to a certain extent, to work with general measurable spaces. In this section, we introduce the measure-theoretic background required for these purposes. Along the way, we establish the notations that we will rely on.

We start by succintly recalling some basic measure-theoretic notions, relating to: measurable spaces (Section 3.2.1), positive and signed measures (Section 3.2.2), Lebesgue and Bochner integrals (Section 3.2.3). Fully explicit definitions of all the above notions can be found in, resp., Appendices C.2.1, C.2.2 and C.2.3. We then introduce the notions involving measure-theoretic morphisms (Section 3.2.4), as well as stochastic transformations: first the measure-theoretic definition of channels — i.e., Markov kernels — and related notions such as push-forwards or hook-ups (Section 3.2.5). In Section 3.2.6, we present different notions of tensor products of distributions and channels which will be pivotal in the study of joinings, before defining, in Section 3.2.7, Markov chains and MDPs in this general measure-theoretic setting. Eventually, we collect and prove some technical properties useful in computations (Section 3.2.8).

3.2.1 Measurable spaces (short version of Appendix C.2.1)

In Appendix C.2.1, we present detailed definitions relating to measurable spaces. In this section, we provide a condensed version of this content, where almost all the definitions are assumed to be known, and we mostly limit ourselves to setting up notations and conventions.

Sets are denoted with calligraphic letters, e.g., $\mathcal{A}$, and collections of subsets with gothic letters, e.g., $\mathfrak{A}$. A measurable space $(\mathcal{A}, \mathfrak{A})$, where $\mathfrak{A}$ is a σ-algebra, is only denoted by $\mathcal{A}$ when there is no ambiguity on the σ-algebra. The algebra, resp. σ-algebra induced by a collection of subsets $\mathfrak{A}$ is denoted by $\mathsf{Alg}(\mathfrak{A})$, resp. $\sigma(\mathfrak{A})$. The induced σ-algebra on a measurable subset E of a measurable space $(\mathcal{A}, \mathfrak{A})$ is denoted by $\mathfrak{A}_E$. If not mentioned explicitly, any measurable subset of a measurable space, when regarded itself as a measurable space, is equipped with the induced σ-algebra. The Borel σ-algebra $\sigma(\mathcal{T})$ of a topological space $(\mathcal{A}, \mathcal{T})$ will often be written $\mathrm{Bor}_{\mathcal{A}}$.

We will focus our attention, as much as possible, on a class of "nice" measurable spaces: *standard Borel spaces*, i.e., Borel spaces whose topology is separable and completely metrisable. Standard Borel spaces are "nice" because they encompass many important examples (e.g., countable spaces, Euclidean spaces, separable Banach spaces, differential manifolds), but still have enough structure for many desirable properties to hold. Most importantly for us, one can take conditional probabilities (see Proposition C.2.16) and do ergodic and information theory (see (Gray, 2009, 2011) and Section 3.4 below).

We will often not refer explicitly to the underlying topology of a standard Borel space, and denote it by $(\mathcal{A}, \mathfrak{A})$ where $\mathfrak{A} := \mathrm{Bor}_{\mathcal{A}}$, or just $\mathcal{A}$ when this yields no confusion. The set $\mathbb{R}$ of real numbers is always equipped with its usual topology and corresponding Borelians, with measurable subsets equipped with the induced topology and σ-algebra.

Let us now turn to products of measurable spaces. Let $(\mathcal{A}_i, \mathfrak{A}_i)_{i\in\mathcal{I}}$ a family of measurable spaces. For subsets of indices $\mathcal{J}' \subseteq \mathcal{J} \subseteq \mathcal{I}$, we denote by $\mathsf{pr}_{\mathcal{J}\to\mathcal{J}'}$ the corresponding projection between Cartesian products. The set of finite-dimensional rectangles is denoted by Rect. The *product measurable space* of the family $(\mathcal{A}_i, \mathfrak{A}_i)_{i\in\mathcal{I}}$ is then

$$(\mathcal{A}, \mathfrak{A}) := \left(\bigtimes_{i\in\mathcal{I}} \mathcal{A}_c, \bigotimes_{i\in\mathcal{I}} \mathfrak{A}_i \right),$$

where $\bigotimes_{i\in\mathcal{I}} \mathfrak{A}_i := \sigma(\mathsf{Rect})$ is the *product σ-algebra*. A Cartesian product of measurable spaces will, unless stated otherwise, always be equipped with the product σ-algebra. Moreover, when $\mathcal{I} = \mathbb{N}$ models time for $\mathcal{A}_n = \mathcal{A}$ for all $n \in \mathbb{N}$ with $\mathcal{A}$ a measurable space, we write $\overrightarrow{\mathcal{A}} := \mathcal{A}^{\mathbb{N}} = \bigtimes_{n\in\mathbb{N}} \mathcal{A}_n$.

Let us recall that while a countable product of standard Borel spaces is standard Borel spaces, this is *not* the case if the product is uncountable (see Proposition C.2.5). Below, we will indeed need to consider uncountable products of standard Borel spaces — more precisely, of what we will define as "ergodic components of a Markov Decision Process". This is the reason why we need to deal also with general measurable spaces, not only standard Borel ones.

3.2.2 Measures (short version of Appendix C.2.2)

In this chapter, we focus on *signed measures* and *finite positive measures*, where **unless specified otherwise, the term "measure" denotes a finite positive measure**. Signed or finite positive measures are typically denoted with the letters μ, ν, ϵ or υ. The sets of signed measures, finite positive measures and probability measures on a measurable space $(\mathcal{A}, \mathfrak{A})$ are denoted by resp. $\mathscr{M}_{\mathcal{A}}$, $\mathscr{M}^+_{\mathcal{A}}$ and $\Delta_{\mathcal{A}}$. A measure space $(\mathcal{A}, \mathfrak{A}, \mu)$ is only denoted by $(\mathcal{A}, \mu)$ when there is no ambiguity on the σ-algebra $\mathfrak{A}$.

For a subset F of a set $\mathcal{A}$, the complement $\mathcal{A} \setminus F$ of F is denoted by $F^{\complement}$. This is not to be confused with superscripts of the form F^c, which we will heavily rely on and do *not* denote complements.

Let $(\mathcal{A}, \mathfrak{A})$ be a measurable space. For $a \in \mathcal{A}$, the *Dirac measure* on a is denoted by δ_a. For $\tilde{\mathcal{A}} \in \mathfrak{A}$, if $\mu \in \mathscr{M}^+_{\mathcal{A}}$ is such that $\mu(\tilde{\mathcal{A}}^{\complement}) = 0$, then it is said *concentrated on* $\tilde{\mathcal{A}}$, and the measure $\tilde{\mu} \in \mathscr{M}^+_{\tilde{\mathcal{A}}}$ defined by $\tilde{\mu}(A \cap \tilde{\mathcal{A}}) := \mu(A)$ for all $A \in \mathfrak{A}$ is called the *restriction* of μ to $\tilde{\mathcal{A}}$. For $\tilde{\mathcal{A}} \in \mathfrak{A}$ and $\tilde{\mu} \in \mathscr{M}^+_{\tilde{\mathcal{A}}}$, the *extension* of $\tilde{\mu}$ to $(\mathcal{A}, \mathfrak{A})$ is the measure μ defined by $\mu(F) := \tilde{\mu}(F \cap \tilde{\mathcal{A}})$ for all $F \in \mathfrak{A}$. With a slight abuse of notation, the restriction or extension of a measure μ will often be denoted with the same symbol μ. For another measurable space $(\mathcal{B}, \mathfrak{B})$, the push-forward of a measure $\mu \in \mathscr{M}^+_{\mathcal{A}}$ through a measurable map $f : \mathcal{A} \to \mathcal{B}$ is denoted by $f \cdot \mu$.

In our proofs, we will make repeated use of the fact that two measures coincide if and only if they coincide on a generating algebra (see Proposition C.2.9), and of the *Kolmogorov extension theorem*, which states that for a (non-necessarily countable) product of standard Borel spaces, if a family of probabilities on each finite subset of coordinates is consistent, then it uniquely defines a probability on the whole product space (see Theorem C.2.10).

3.2.3 Lebesgue and Bochner integrals (short version of Appendix C.2.3)

We will assume familiarity with Lebesgue integration theory , and refer to, e.g., (Gray, 2009) for basic definitions and results. For $\mathcal{A}$ a measurable space, $f : \mathcal{A} \to \mathbb{R}$ a measurable function and $\mu \in \mathscr{M}_{\mathcal{A}}$, whenever the integral makes sense, we write

$$\langle \mu, f \rangle := \int_{\mathcal{A}} f \, d\mu(a).$$

Now, in Section 3.4, we will need to consider integrals valued in probability spaces, for which *Bochner integrals* are a natural language. Indeed, the latter generalise the Lebesgue integrals, defined for scalar-valued functions, to maps valued in potentially infinite-dimensional spaces — more precisely, *Banach spaces*.[18] We assume familiarity with basic facts about

[18] I.e., vector spaces equipped with a norm that induces a complete metric. Banach spaces can be seen as infinite-dimensional generalisations of Euclidean spaces, and play a central role in functional analysis.

these spaces, and we refer to, e.g., Appendix E in (Cohn, 2013) for a complete presentation of Bochner integrals. Here, let us just highlight the following:

- While Lebesgue integration considers measurable scalar-valued functions, integrating Banach space-valued functions requires adding a requirement of separability, leading to a notion of *strong measurability* (see Definition C.2.12). A function is then called *Bochner integrable* if it is strongly measurable and its norm is Lebesgue integrable.
- One can then construct a notion of *Bochner integral* of any Bochner integrable function (Cohn, 2013), where the integral is valued in the same Banach space as the function's output space. Lebesgue integral's dominated convergence theorem then generalises to Bochner integrals for sequences of strongly measurable functions (see Theorem C.2.13).

3.2.4 Measure-theoretic morphisms

The theoretical framework developed in this chapter is based on a series of notions of morphisms and isomorphisms, where each new notion adds more structure to previous ones. Here, let us start by introducing the most fundamental ones, which are standard objects in ergodic theory.

Definition 3.2.1. Let $\mathcal{A}$ and $\mathcal{B}$ be measurable spaces. A *measurable isomorphism* from $\mathcal{A}$ to $\mathcal{B}$ is a measurable map $f : \mathcal{A} \to \mathcal{B}$ which is bijective with measurable inverse. If such a map exists, $\mathcal{A}$ and $\mathcal{B}$ are then said *measurably isomorphic*. If $\mu_{\mathcal{A}} \in \Delta_{\mathcal{A}}$ and $\mu_{\mathcal{B}} \in \Delta_{\mathcal{B}}$, a *measured morphism* from $(\mathcal{A}, \mu_{\mathcal{A}})$ to $(\mathcal{B}, \mu_{\mathcal{B}})$ is a measurable map $f : \mathcal{A} \to \mathcal{B}$ such that $f \cdot \mu_{\mathcal{A}} = \mu_{\mathcal{B}}$. It is a *measured isomorphism* if there there exist $\tilde{\mathcal{A}} \subseteq \mathcal{A}$ and $\tilde{\mathcal{B}} \subseteq \mathcal{B}$ with $\mu_{\mathcal{A}}(\tilde{\mathcal{A}}) = \mu_{\mathcal{B}}(\tilde{\mathcal{B}}) = 1$ such that f induces a measurable isomorphism from $\tilde{\mathcal{A}}$ to $\tilde{\mathcal{B}}$. For $f : (\mathcal{A}, \mu_{\mathcal{A}}) \to (\mathcal{B}, \mu_{\mathcal{B}})$ a measurable isomorphism, a *mod 0 inverse of f w.r.t. the distributions* $(\mu_{\mathcal{A}}, \mu_{\mathcal{B}})$, or *mod 0 inverse of f* for short, is a measurable function $f^{-1} : \mathcal{B} \to \mathcal{A}$ such that there exist $\tilde{\mathcal{A}} \subseteq \mathcal{A}$ and $\tilde{\mathcal{B}} \subseteq \mathcal{B}$ with $\mu_{\mathcal{A}}(\tilde{\mathcal{A}}) = \mu_{\mathcal{B}}(\tilde{\mathcal{B}}) = 1$ such that f, f^{-1} induce maps $\widetilde{f} : \tilde{\mathcal{A}} \to \tilde{\mathcal{B}}$, $\widetilde{f^{-1}} : \tilde{\mathcal{B}} \to \tilde{\mathcal{A}}$ satisfying $\widetilde{f^{-1}} \circ \widetilde{f} = \mathtt{Id}_{\tilde{\mathcal{A}}}$ and $\widetilde{f} \circ \widetilde{f^{-1}} = \mathtt{Id}_{\tilde{\mathcal{B}}}$.

In short, a measured isomorphism is a measurable isomorphism up to a null measure set, and a mod 0 inverse is an inverse up to a null measure set. Let us stress that, while all mod 0 inverses of a measured isomorphism $f : (\mathcal{A}, \mu_{\mathcal{A}}) \to (\mathcal{B}, \mu_{\mathcal{B}})$ coincide $\mu_{\mathcal{B}}$-a.e, there is no uniqueness of mod 0 inverses, and **despite our notation f^{-1}, they might not be set-theoretic inverses of f**. Note also that our use of the terminology "measured isomorphism" is slightly unconventional (the conventional terminology would be "mod 0 isomorphism" (Coudène, 2016)).

Measured isomorphisms are "well-behaved" w.r.t. the operation of taking mod inverses, w.r.t composition, and w.r.t. almost everywhere equality (see Proposition C.2.14).

3.2.5 Channels

We now turn to notions that will be the core of our analysis below. Let us start with the measure-theoretic formalisation of stochastic maps — channels — and related concepts.

In the definitions below, $(\mathcal{A}, \mathfrak{A})$, $(\mathcal{B}, \mathfrak{B})$ and $(\mathcal{C}, \mathfrak{C})$ denote measurable spaces.

Definition 3.2.2. A *channel*[19] γ from $\mathcal{A}$ to $\mathcal{B}$ is a map

$$\begin{aligned} \gamma : \mathcal{A} \times \mathfrak{B} &\to [0, 1] \\ (a, F_{\mathcal{B}}) &\mapsto \gamma(F_{\mathcal{B}}|a) \end{aligned}$$

[19] Also known as "Markov kernel", "conditional probability", "transition probability" or "stochastic map".

such that $\gamma(\cdot|a) \in \Delta_{\mathcal{A}}$ for all $a \in \mathcal{A}$, and for every $F_{\mathcal{B}} \in \mathfrak{B}$, the map

$$\begin{aligned} \mathcal{A} &\to [0,1] \\ a &\mapsto \gamma(F_{\mathcal{B}}|a) \end{aligned}$$

is measurable. The set of channels from $\mathcal{A}$ to $\mathcal{B}$ is denoted by $\mathscr{K}(\mathcal{A},\mathcal{B})$, and we write $\mathscr{K}(\mathcal{A}) := \mathscr{K}(\mathcal{A},\mathcal{A})$. For a measurable function $f : \mathcal{A} \to \mathcal{B}$, the deterministic channel defined by f will often also be denoted by f: i.e., for all $a \in \mathcal{A}$, $F_{\mathcal{B}} \in \mathfrak{B}$, we define $f(F_{\mathcal{B}}|a) := \delta_{f(a)\in F_{\mathcal{B}}}$ where the left-hand-side uses the "channel" notation while the right-hand-side uses the "function" notation (in the following, if it is clear from context, we will not specify which of the two notations we are using). For $\mu \in \Delta_{\mathcal{A}}$, we say that two channels $\gamma, \gamma' \in \mathscr{K}(\mathcal{A},\mathcal{B})$ coincide *μ-almost everywhere*, or *μ-a.e.* for short, if there exists $F_{\mathcal{A}} \in \mathfrak{A}$ with $\mu(F_{\mathcal{A}}) = 1$ such that $\gamma(\cdot|a) = \gamma'(\cdot|a)$ for all $a \in F_{\mathcal{A}}$. The composition of two channels $\gamma \in \mathscr{K}(\mathcal{A},\mathcal{B})$ and $\gamma' \in \mathscr{K}(\mathcal{B},\mathcal{C})$ is the channel $\gamma' \circ \gamma$ from $\mathcal{A}$ to $\mathcal{C}$ defined, for all $a \in \mathcal{A}$, $F_{\mathcal{C}} \in \mathfrak{C}$, by

$$(\gamma' \circ \gamma)(F_{\mathcal{C}}|a) := \int_{\mathcal{B}} \gamma'(F_{\mathcal{C}}|b) d\gamma(b|a).$$

Eventually, for any measurable subset $\tilde{\mathcal{A}} \in \mathfrak{A}$, the (necessarily unique) *restriction* of $\gamma \in \mathscr{K}(\mathcal{A},\mathcal{B})$ to $\tilde{\mathcal{A}}$ is the channel $\tilde{\gamma} \in \mathscr{K}(\tilde{\mathcal{A}},\mathcal{B})$ defined by $\tilde{\gamma}(\cdot|a) := \gamma(\cdot|a)$ for all $a \in \tilde{\mathcal{A}}$. If we rather assume $\gamma \in \mathscr{K}(\tilde{\mathcal{A}},\mathcal{B})$, a (non-necessarily unique) *extension* of γ to $\mathcal{A}$ is a channel $\tilde{\gamma} \in \mathscr{K}(\mathcal{A},\mathcal{B})$ such that $\tilde{\gamma}(\cdot|a) := \gamma(\cdot|a)$ for all $a \in \tilde{\mathcal{A}}$

Let us now turn to the **hook-up notation, which we will heavily rely one along the whole chapter**. In short, it consists in combining channels and distributions on their input to obtain *joint* input-output distributions — or in combining only channels, to obtain new channels whose output space is the Cartesian product of the combined channels' output spaces.

Definition 3.2.3. The *hook-up*[20] of $\mu \in \Delta_{\mathcal{A}}$ and a channel γ is the joint distribution $\mu\gamma \in \Delta_{\mathcal{A}\times\mathcal{B}}$ defined by, for all $F_{\mathcal{A}} \in \mathfrak{A}$, $F_{\mathcal{B}} \in \mathfrak{B}$,

$$\mu\gamma(F_{\mathcal{A}} \times F_{\mathcal{B}}) := \int_{F_{\mathcal{A}}} \gamma(F_{\mathcal{B}}|a) d\mu(a).$$

We also define the *hook-up of two channels* $\gamma \in \mathscr{K}(\mathcal{A},\mathcal{B})$ and $\gamma' \in \mathscr{K}(\mathcal{A}\times\mathcal{B},\mathcal{C})$ as the channel $\gamma\gamma' \in \mathscr{K}(\mathcal{A},\mathcal{B}\times\mathcal{C})$ defined, for all $a \in \mathcal{A}$, $F_{\mathcal{B}} \in \mathfrak{B}$, $F_{\mathcal{C}} \in \mathfrak{C}$, by

$$\gamma\gamma'(F_{\mathcal{B}} \times F_{\mathcal{C}}|a) := \int_{F_{\mathcal{B}}} \gamma'(F_{\mathcal{C}}|a,b) d\gamma(b|a).$$

When this yields no ambiguity, we use the same notation if the left-hand-side of the hook-up is not defined on the whole input space of the right-hand-side. E.g., if $\mu \in \Delta_{\mathcal{A}}$ with $(\mathcal{A},\mathfrak{A}) = (\mathcal{A}_1 \times \mathcal{A}_2, \mathfrak{A}_1 \otimes \mathfrak{A}_2)$ and $\gamma \in \mathscr{K}(\mathcal{A}_2,\mathcal{B})$, then $\mu\gamma \in \Delta_{\mathcal{A}_1\times\mathcal{A}_2\times\mathcal{B}}$ is defined, for $F_{\mathcal{A}_1} \in \mathfrak{A}_1$, $F_{\mathcal{A}_2} \in \mathfrak{A}_2$, $F_{\mathcal{B}} \in \mathfrak{B}$, by

$$\mu\gamma(F_{\mathcal{A}_1} \times F_{\mathcal{A}_2} \times F_{\mathcal{B}}) := \int_{F_{\mathcal{A}_1}\times F_{\mathcal{A}_2}} \gamma(F_{\mathcal{B}}|a_2) d\mu(a_1,a_2),$$

[20] We take this terminology from (Gray, 2011).

and similarly, if $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$, $\gamma' \in \mathscr{K}(\mathcal{B}, \mathcal{C})$, then $\gamma\gamma' \in \mathscr{K}(\mathcal{A}, \mathcal{B} \times \mathcal{C})$ is defined, for $F_{\mathcal{B}} \in \mathfrak{B}$, $F_{\mathcal{C}} \in \mathfrak{C}$, $a \in \mathcal{A}$, by

$$\gamma\gamma'(F_{\mathcal{B}} \times F_{\mathcal{C}}|a) := \int_{F_{\mathcal{B}}} \gamma'(F_{\mathcal{C}}|b) d\gamma(b|a).$$

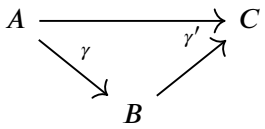
When we write expressions involving several hook-ups at the same time, we remove the brackets if it yields no ambiguity: e.g., for $\mu \in \Delta_{\mathcal{A}}$, $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$, and $\gamma' \in \mathscr{K}(\mathcal{A} \times \mathcal{B}, \mathcal{C})$, we write

$$\mu\gamma\gamma' := (\mu\gamma)\gamma' = \mu(\gamma\gamma').$$

To clarify the meaning of the hook-up notation, let us use graphical representations. For $\mu \in \Delta_{\mathcal{A}}$, $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$, the hook-up $\mu\gamma \in \Delta_{\mathcal{A}\times\mathcal{B}}$ can be represented by the Bayesian network

$$0 \xrightarrow{\mu} A \xrightarrow{\gamma} B$$

where in addition to the variables (A, B) with the joint distribution $\mu\gamma$, we add a trivial variable 0 on the single element set $\{0\}$, and see the distribution $\mu \in \Delta_{\mathcal{A}}$ as a channel from this trivial set to $\mathcal{A}$ — which makes it possible to graphically represent the fact that A has the distribution μ.[21] On the other hand, the hook-up $\gamma\gamma' \in \mathscr{K}(\mathcal{A}, \mathcal{B} \times \mathcal{C})$ of two channels $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$ and $\gamma' \in \mathscr{K}(\mathcal{A} \times \mathcal{B}, \mathcal{C})$, which is a channel rather than a joint distribution, can be represented by *not* specifying any distribution on $\mathcal{A}$, through the Bayesian network



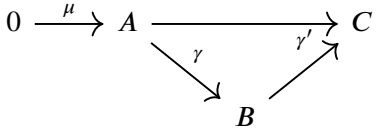
Now, if we have $\mu \in \Delta_{\mathcal{A}_1\times\mathcal{A}_2}$ and $\gamma \in \mathscr{K}(\mathcal{A}_2, \mathcal{B})$, Bayesian networks are less suited for a graphical representation of $\mu\gamma$,[22] but if, e..g., we can decompose μ as $\mu = \mu_{A_1}\mu_{A_2|A_1}$ where $\mu_{A_1} \in \Delta_{\mathcal{A}_1}$ is the marginal of μ on A_1 and $\mu_{A_2|A_1} \in \mathscr{K}(\mathcal{A}_1, \mathcal{A}_2)$, then we can represent $\mu\gamma = (\mu_{A_1}\mu_{A_2|A_1})\gamma$ with the Bayesian network

$$0 \xrightarrow{\mu_{A_1}} A_1 \xrightarrow{\mu_{A_2|A_1}} A_2 \xrightarrow{\gamma} B$$

If $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$ and $\gamma' \in \mathscr{K}(\mathcal{B}, \mathcal{C})$, then the Bayesian network representing $\gamma\gamma' \in \mathscr{K}(\mathcal{A}, \mathcal{B}\times\mathcal{C})$ is

$$A \xrightarrow{\gamma} B \xrightarrow{\gamma'} C$$

Eventually, for $\mu \in \Delta_{\mathcal{A}}$, $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$, and $\gamma' \in \mathscr{K}(\mathcal{A} \times \mathcal{B}, \mathcal{C})$, the "double" hook-up $\mu\gamma\gamma' = (\mu\gamma)\gamma' = \mu(\gamma\gamma')$ can be represented by the Bayesian network



On an arbitrary subset $F \in \mathfrak{A} \otimes \mathfrak{B}$ of the product space $\mathcal{A} \times \mathcal{B}$, hook-ups of the form

[21] This is loosely inspired from string diagrams in categorical probability (Perrone, 2024).

[22] Categorical probability's string diagrams could address this limitation — see Section 2.2 in (Perrone, 2024).

$\mu\gamma \in \Delta_{\mathcal{A}\times\mathcal{B}}$ can be computed by integrating conditional probabilities of *sections* of the output space $\mathcal{B}$ w.r.t. the input space distribution μ. See Proposition C.2.15 for a formal statement.

The next definition is convenient to compare hook-ups of the form $\mu_{\mathcal{A}}\gamma_{\mathcal{A}\to\mathcal{B}} \in \Delta_{\mathcal{A}\times\mathcal{B}}$ and $\mu_{\mathcal{B}}\gamma_{\mathcal{B}\to\mathcal{A}} \in \Delta_{\mathcal{B}\times\mathcal{A}}$.

Definition 3.2.4. Let $(\mathcal{A}, \mathfrak{A})$, $(\mathcal{B}, \mathfrak{B})$ measurable and $\mu \in \Delta_{\mathcal{A}\times\mathcal{B}}$. The *transpose* of μ, denoted by μ^{T}, is the distribution on $\mathcal{B}\times\mathcal{A}$ defined, for all $F_{\mathcal{A}} \in \mathfrak{A}$, $F_{\mathcal{B}} \in \mathfrak{B}$, by $\mu^{\mathsf{T}}(F_{\mathcal{B}} \times F_{\mathcal{A}}) := \mu(F_{\mathcal{A}} \times F_{\mathcal{B}})$.

Channels can be used to transform distribution on their input space into distrbution on their output space — this operation is called the channel's *push-forward*:

Definition 3.2.5. The *push-forward* of $\mu \in \Delta_{\mathcal{A}}$ through $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$, denoted by $\gamma \cdot \mu$, is the probability on $\mathcal{B}$ defined for all $F_{\mathcal{B}} \in \mathfrak{B}$ by

$$(\gamma \cdot \mu)(F_{\mathcal{B}}) := \mu\gamma(\mathcal{A} \times F_{\mathcal{B}}) = \int_{\mathcal{A}} \gamma(F_{\mathcal{B}}|a) d\,\mu(a).$$

We denote by γ_* the corresponding *push-forward operator*[23]

$$\begin{aligned} \gamma_* : \Delta_{\mathcal{A}} &\to \Delta_{\mathcal{B}} \\ \mu &\mapsto \gamma_*\mu := \gamma \cdot q. \end{aligned}$$

If $f : \mathcal{A} \to \mathcal{B}$ is measurable, then the push-forward $f \cdot \mu$ of $\mu \in \Delta_{\mathcal{A}}$ is the push-forward of μ through the deterministic channel defined by f, i.e., $(f \cdot \mu)(F_{\mathcal{B}}) := \mu(f^{-1}(F_{\mathcal{B}}))$ for all $F_{\mathcal{B}} \in \mathfrak{B}$.

Note that a channel can always be recovered from its push-forward, through the formula $\gamma(\cdot|a) = (\gamma \cdot \delta_a)$.

For probabilities on finite spaces, we can always use joint distributions to define corresponding conditional distributions. One of the "nice" features of standard Borel spaces is that they provide a more expressive framework in which the property also holds true (see Proposition C.2.16 for a formal statement).

3.2.6 Tensor products

Tensor products will be basic tools for our generalised pose coordinate in Section 3.5.

Definition 3.2.6. Let $(\mathcal{A}_i, \mathfrak{A}_i)_{i\in\mathcal{I}}$, $(\mathcal{B}_i, \mathfrak{B}_i)_{i\in\mathcal{I}}$ two families of measurable spaces. We assume that either $\mathcal{I}$ is finite, or the spaces $\mathcal{A}_i$ and $\mathcal{B}_i$ are standard Borel for all $i \in \mathcal{I}$. Then:

(i) Let $\mu_i \in \Delta_{\mathcal{A}_i}$ for all $i \in \mathcal{I}$. The *tensor product of* $(\mu_i)_{i\in\mathcal{I}}$, denoted by $\bigotimes_{i\in\mathcal{I}} \mu_i$, is the distribution on $(\bigtimes_{i\in\mathcal{I}} \mathcal{A}_i, \bigotimes_{i\in\mathcal{I}} \mathfrak{A}_i)$ defined, for all $\mathcal{J} \subseteq \mathcal{I}$ finite and $\bigtimes_{j\in\mathcal{J}} F_j \in \bigotimes_{j\in\mathcal{J}} \mathcal{A}_j$, by

$$\left(\bigotimes_{i\in\mathcal{I}} \mu_i\right)\left(\bigtimes_{j\in\mathcal{J}} F_j\right) = \prod_{j\in\mathcal{J}} \mu(F_j). \tag{3.2.1}$$

(ii) Let $\gamma_{\mathcal{A}_i\to\mathcal{B}_i} \in \mathscr{K}(\mathcal{A}_i, \mathcal{B}_i)$ for all $i \in \mathcal{I}$. The *tensor product of* $(\gamma_{\mathcal{A}_i\to\mathcal{B}_i})_{i\in\mathcal{I}}$, denoted by $\bigotimes_{i\in\mathcal{I}} \gamma_{\mathcal{A}_i\to\mathcal{B}_i}$, is the channel from $(\bigtimes_{i\in\mathcal{I}} \mathcal{A}_i, \bigotimes_{i\in\mathcal{I}} \mathfrak{A}_i)$ to $(\bigtimes_{i\in\mathcal{I}} \mathcal{B}_i, \bigotimes_{i\in\mathcal{I}} \mathfrak{B}_i)$ defined, for all $(a_i)_{i\in\mathcal{I}} \in \bigtimes_{i\in\mathcal{I}} \mathcal{A}_i$, finite set of indices $\mathcal{J} \subseteq \mathcal{I}$ and $\bigtimes_{j\in\mathcal{J}} F_j \in \bigotimes_{j\in\mathcal{J}} \mathcal{B}_j$,

[23] We will mostly use the notation $\gamma \cdot \mu$ and not $\gamma_*\mu$, in part. to avoid confusions with the hook-up notation.

by

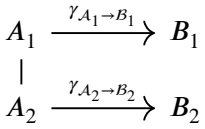
$$\left(\bigotimes_{i\in\mathcal{I}} \gamma_{\mathcal{A}_i\to\mathcal{B}_i}\right)\left(\bigtimes_{j\in\mathcal{J}} F_j \middle| (a_i)_{i\in\mathcal{I}}\right) = \prod_{j\in\mathcal{J}} \gamma_{\mathcal{A}_j\to\mathcal{B}_j}(F_j|a_j) \tag{3.2.2}$$

(iii) Let $(\mathcal{A}, \mathfrak{A})$ another measurable space, and $\gamma_{\mathcal{A}\to\mathcal{B}_i} \in \mathscr{K}(\mathcal{A}, \mathcal{B}_i)$ for all $i \in \mathcal{I}$. The *output tensor product of* $(\gamma_{\mathcal{A}\to\mathcal{B}_i})_{i\in\mathcal{I}}$, denoted by $\bigcirc\!\!\!\!\!\!\times_{i\in\mathcal{I}} \gamma_{\mathcal{A}\to\mathcal{B}_i}$, is the channel from $(\mathcal{A}, \mathfrak{A})$ to $(\bigotimes_{i\in\mathcal{I}} \mathcal{B}_i, \bigotimes_{i\in\mathcal{I}} \mathfrak{B}_i)$ defined, for all $a \in \mathcal{A}$, finite set of indices $\mathcal{J} \subseteq \mathcal{I}$ and $\bigtimes_{j\in\mathcal{J}} F_j \in \bigotimes_{j\in\mathcal{J}} \mathcal{B}_j$, by

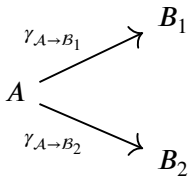
$$\left(\bigcirc\!\!\!\!\!\!\times_{i\in\mathcal{I}} \gamma_{\mathcal{A}\to\mathcal{B}_i}\right)\left(\bigtimes_{j\in\mathcal{J}} F_j \middle| a\right) = \prod_{j\in\mathcal{J}} \gamma_{\mathcal{A}\to\mathcal{B}_j}(F_j|a) \tag{3.2.3}$$

Moreover, the (output) tensor product of measurable functions is defined as the measurable function defined by the (output) tensor product of their corresponding deterministic channels: e.g., for $f_1 : \mathcal{A}_1 \to \mathcal{B}_1$ and $f_2 : \mathcal{A}_2 \to \mathcal{B}_2$, we have $(f_1 \otimes f_2)(a_1, a_2) := (f(a_1), f(a_2))$.

Tensor products (of distributions or channels) and output tensor products of channels are well-defined (see Appendix C.2.6). In plain words, the tensor product $\bigotimes_{i\in\mathcal{I}} \mu_i$ of probability distributions μ_i is the unique joint distribution whose marginals are the μ_i and that makes the coordinates of $\mathcal{I}$ independent. The tensor product $\bigotimes_{i\in\mathcal{I}} \gamma_{\mathcal{A}_i\to\mathcal{B}_i}$ formalises the parallel processing of the distinct coordinates $i \in \mathcal{I}$, which can be represented, e.g., for $\mathcal{I} = \{1, 2\}$, by the Bayesian network



where the line between A_1 and A_2 indicates that these "input" variables are not necessarily assumed independent. The output tensor product $\bigcirc\!\!\!\!\!\!\times_{i\in\mathcal{I}} \gamma_{\mathcal{A}\to\mathcal{B}_i}$ formalises the independent processing of different copies of the input space $\mathcal{A}$, which can be represented, e.g., for $\mathcal{I} = \{1, 2\}$, by the Bayesian network



Note that the operation of taking the tensor products of two channels is bilinear, and that the composition of tensor products of channels coincides with the tensor product of the compositions on each coordinates (see Proposition C.2.17 for a formal statement).

3.2.7 Markov chains and Markov Decision Processes

We eventually define the object whose structure we want to investigate in this work: Markov Decision Processes (MDPs). Let us however start with specific cases of MDPs that are also central to the broader study of time-evolving dynamics: dynamical systems and Markov chains.

Definition 3.2.7. A *measurable dynamical system* is a tuple $(\mathcal{X}, \mathfrak{X}, \tau)$ with $(\mathcal{X}, \mathfrak{X})$ a measurable space and $\tau : \mathcal{X} \to \mathcal{X}$ a measurable transformation — we often only write $(\mathcal{X}, \tau)$ when

there no ambiguity on the σ-algebra $\mathfrak{X}$. A *stationary dynamical system*[24] $(\mathcal{X}, \mu_0, \tau)$ is a measurable dynamical system equipped with a probability $\mu_0 \in \Delta_{\mathcal{X}}$ which is stationary under τ, i.e., $\tau \cdot \mu_0 = \mu_0$.

If the deterministic transformation τ becomes stochastic, we obtain a Markov chain:

Definition 3.2.8. A *measurable Markov chain* is a tuple $(\mathcal{X}, \mathfrak{X}, \tau)$, where $(\mathcal{X}, \mathfrak{X})$ is a measurable space, and $\tau \in \mathscr{K}(\mathcal{X})$. When it yields no ambiguity, we refer to a measured Markov chain only as the pair $(\mathcal{X}, \tau)$, or even just the channel τ. The Markov chain is called *standard Borel* when $\mathcal{X}$ is a standard Borel space. A *measured Markov chain* is a measurable Markov chain equipped with an initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$, i.e. a tuple $(\mathcal{X}, \mathfrak{X}, \mu_0, \tau)$ — or, e.g., (μ_0, τ) for short. A distribution $\mu_0 \in \Delta_{\mathcal{X}}$ is said *stationary under* τ if $\tau \cdot \mu_0 = \mu_0$. A *stationary Markov chain* is a measured Markov chain (μ_0, τ) such that μ_0 is stationary.

Technical remark. A (measurable, resp. stationary) dynamical system is equivalent to a (measurable, resp. stationary) Markov chain whose transformation τ is a deterministic channel (replace the measurable function from Definition 3.2.7 by the corresponding deterministic channel).

On the other hand, if we introduce actions into this picture, we obtain a Markov Decision Process:

Definition 3.2.9. A *measurable Markov Decision Process* (or *measurable MDP* for short) is a tuple $(\mathcal{X}, \mathfrak{X}, \mathcal{G}, \mathfrak{G}, \pi, \rho)$, where $(\mathcal{X}, \mathfrak{X})$ and $(\mathcal{G}, \mathfrak{G})$ are measurable spaces, $\pi \in \mathscr{K}(\mathcal{X}, \mathcal{G})$ and $\rho \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X})$. The space $\mathcal{X}$ is called the *state-space*. $\mathcal{G}$ the *action space*,[25] ρ the *transition channel* and π the *policy*. For all $g \in \mathcal{G}$, the *action* $\rho_g \in \mathscr{K}(\mathcal{X})$ is defined, for all $x \in \mathcal{X}$, $F \in \mathrm{Bor}_{\mathcal{X}}$, by

$$\rho_g(F|x) := \rho(F|x, g),$$

and the *update channel*, denoted by $\bar{\rho} \in \mathscr{K}(\mathcal{X})$ is the average of the actions ρ_g w.r.t. the policy π, i.e., for all $x \in \mathcal{X}$, $F \in \mathrm{Bor}_{\mathcal{X}}$,

$$\bar{\rho}(F|x) := \big(\rho \circ (\mathtt{Id}_{\mathcal{X}} \rtimes \pi)\big)(F|x) = \int_{g \in \mathcal{G}} \rho_g(F|x) d\pi(g|x).$$

When it yields no ambiguity, we refer to a measurable MDP only as the tuple $(\mathcal{X}, \mathcal{G}, \pi, \rho)$, or even just the pair of channels (π, ρ). The MDP is called *standard Borel* when both $\mathcal{X}$ and $\mathcal{G}$ are standard Borel spaces. A *measured MDP* is a measurable MDP equipped with an initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$, i.e., a tuple $(\mathcal{X}, \mathfrak{X}, \mathcal{G}, \mathfrak{G}, \mu_0, \pi, \rho)$, or, e.g., (μ_0, π, ρ) for short. A *stationary MDP* is a measured MDP (μ_0, π, ρ) such that the initial distribution is stationary under the update channel, i.e., $\bar{\rho} \cdot \mu_0 = \mu_0$.

Technical remark. A (measurable, resp. measured, resp. stationary) Markov chain is equivalent to a (measurable, resp. measured, resp. stationary) with only one action, i.e., with $|\mathcal{G}| = 1$. In particular, a (measurable, resp. stationary) dynamical system can be seen as a (measurable, resp. stationary) MDP with one action and deterministic transition function. Moreover, importantly, our definition of stationarity for MDPs does not require the stationarity of the initial distribution μ_0 w.r.t. each action ρ_g for all $g \in \mathcal{G}$, but *only w.r.t. to the average* $\bar{\rho}$ *of all actions* $(\rho_g)_{g \in \mathcal{G}}$ *w.r.t. the policy* π.

Let us write $\mathcal{X}_0^n = \mathcal{X}_0 \times \cdots \times \mathcal{X}_n$, resp. $\mathcal{G}_0^n = \mathcal{G}_0 \times \cdots \times \mathcal{G}_n$, the product of $n+1$ copies of $\mathcal{X}$, resp. of $\mathcal{G}$ — each copy corresponding to one time-step; with σ-algebras denoted by $\mathfrak{X}_0^n$ and

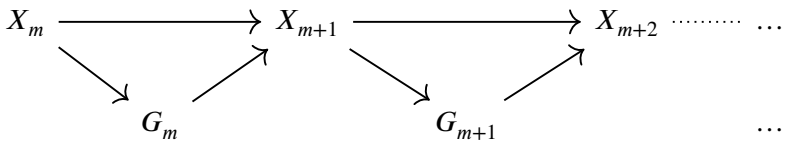
[24] More commonly known as *probability measure preserving transformation* (Coudène, 2016).

[25] Our notation $\mathcal{G}$ stands for "generators", as the elements of $\mathcal{G}$ generate strings of multiple time-step actions.

$\mathfrak{G}_0^n$. A measured MDP (μ_0, π, ρ) defines, for each pair of finite times $m < n$, a distribution

$$q_m^n := q_m^n(X_m, G_m, \dots, X_{n-1}, G_{n-1}, X_n) \in \Delta_{\mathcal{X}_m^n \times \mathcal{G}_m^{n-1}} \tag{3.2.4}$$

satisfying the Bayesian network



where the first time-steps drawn here are iterated up to time n. I.e., formally: we define iteratively, using the hook-up notation from Definition 3.2.2,

$$\begin{aligned} q_0^0 &:= \mu_0 \in \Delta_{\mathcal{X}_0}, \\ \forall n \geq 1, \quad q_0^n &:= q_0^{n-1}\pi\rho := (q_0^{n-1}\pi)\rho \in \Delta_{\mathcal{X}_0^n \times \mathcal{G}_0^{n-1}}, \end{aligned} \tag{3.2.5}$$

where here π is seen as a channel from $\mathcal{X}_n$ to $\mathcal{G}_n$, and ρ as a channel from $\mathcal{X}_n \times \mathcal{G}_n$ to $\mathcal{X}_{n+1}$. The distribution q_m^n is then the marginal of q_0^n on the last $(n-m)$ coordinates. For standard Borel spaces, the Kolmogorov extension theorem (Theorem C.2.10) ensures that there exists a unique distribution

$$q := q(\overrightarrow{X, G}) := q((X_n, G_n)_{n\in\mathbb{N}})$$

on $\overrightarrow{\mathcal{X} \times \mathcal{G}} := (\mathcal{X} \times \mathcal{G})^{\mathbb{N}}$ such that for all $m, n \in \mathbb{N}$, we have $\mu(X_m, G_m, \dots, X_n, G_n) = \quad q_m^n$. Similarly, each measured Markov chains (μ, τ) on a standard Borel space $\mathcal{X}$ uniquely defines a distribution $\mu \in \Delta_{\overrightarrow{\mathcal{X}}}$.

Definition 3.2.10. The *process distribution* of a standard Borel measured MDP (μ_0, π, ρ) is the distribution $q(\overrightarrow{X, G})$ defined above. The *process distributions* of a standard Borel measurable MDP (π, ρ) are all the process distributions of the measured MDPs (μ_0, π, ρ) for all initial distributions $\mu_0 \in \Delta_{\mathcal{X}}$. We make similar definitions for process distribution(s) of standard Borel Markov chains.

Using the Kolmogorov extension theorem again, it is easy to verify the following:

Proposition 3.2.11. *Let (π, ρ) be a standard Borel measurable MDP. There exists a unique channel $\overrightarrow{\pi\rho} \in \mathscr{K}(\mathcal{X}, \overrightarrow{\mathcal{G} \times \mathcal{X}})$ such that*

$$\begin{aligned} \overrightarrow{\pi\rho}_0^1 &:= \pi\rho, \\ \forall n \geq 2, \quad \overrightarrow{\pi\rho}_0^n &:= (\overrightarrow{\pi\rho}_0^{n-1}\pi)\rho, \end{aligned}$$

where for all $n \geq 1$, the channel $\overrightarrow{\pi\rho}_0^n \in \mathscr{K}(\mathcal{X}, (\mathcal{G} \times \mathcal{X})^n)$ is such that for all $x \in \mathcal{X}$, the distribution $\overrightarrow{\pi\rho}_0^n(\cdot|x)$ is the marginal of $\overrightarrow{\pi\rho}(\cdot|x)$ on the first coordinates $(\mathcal{G} \times \mathcal{X})^n$. Moreover, this channel satisfies $q = \mu_0\overrightarrow{\pi\rho}$ for all process distribution q of the measurable MDP (π, ρ) equipped with an initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$.

Definition 3.2.12. The *process channel* of a standard Borel measurable MDP (π, ρ) is the channel $\overrightarrow{\pi\rho}$ from Proposition 3.2.11.

In short, the process channel transforms an initial state $x \in \mathcal{X}$ into the resulting process $\overrightarrow{\pi\rho}(\cdot|x)$ of actions (starting from time 0) and resulting states (starting from time 1) obtained with the MDP (π, ρ) starting from x at time 0; and combining any initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$ with $\overrightarrow{\pi\rho}$, we obtain a process distribution $\mu_0\overrightarrow{\pi\rho}$ starting from time 0 for both states and actions.

The point of this definition is that it will turn out to be a convenient mathematical object to state and prove theorems (see, in particular, Theorem 3.4.1 below).

On our use of the MDP formalism w.r.t. embodied agents Our definition of measurable MDP is a more abstract, "measurable space" version of the usual definition of MDP, with however a crucial difference. Usually, the definition of an MDP involves a reward function instead of a policy — and the main problem is then to learn an optimal policy for the given reward function. Here, we do not address any kind of "policy learning" problem, rather working with an already given, fixed policy π, and disregarding any reward function that might have led to learning that policy. This is because our focus is on understanding the *structure of the dynamics* induced by a given behaviour in a given environment — **e.g., if the state-space $\mathcal{X}$ models a fully observed environment, and the action space $\mathcal{G}$ the agent's own actions**. Let us also stress that, despite the "Markov *Decision* Process" terminology, the policy is here just thought of as a statistical description of the agent's closed-loop behaviour, rather than as a model of decision-making — which, in an embodied setting, some have argued happens *along* the unfolding of movement rather than in a separate stage (Thura et al., 2022). However, even when not interpreted as sequential models of decision-making, MDPs are ultimately very limited formalisations of real-world agents — either biological or artificial. From an adaptive behaviour perspective, the work presented here should thus be seen as only a building block for more realistic models: e.g. causal Bayesian network models of the perception-action loop (Ay et al., 2014; Polani et al., 2009), coupled Moore and Mealy machines (Virgo et al., 2025), or input-output processes (Barnett et al., 2015; Rosas et al., 2025) (in the latter, one only assumes an "interface" channel describing how any "input process" of agent's actions yields a simultaneously occuring "output process" of resulting sensations).

3.2.8 Some useful rules

In our proofs (especially in Section 3.5), we will manipulate different combinations of the concepts of hook-up, channel composition, and push-forward defined above. In Appendix C.3, we collect algebraic rules that will be useful for that purpose. Some of them might well be scattered or implicitly used across the measure-theoretic literature. But for the sake of completeness, we include all the detailed proofs. These results will only be useful in appendices. We recommend skipping them at the first reading and coming back to them when necessary.

3.3 Ergodic decomposition of standard Borel Markov chains

Previous work on class-pose decomposition has formalised "classes" as orbits under the action of a given group. Here, following the considerations from Section 3.1.4, our aim is to show that the notion of class can be made relevant to much more general cases by recasting it in terms of *ergodic components* of the Markov chain defined by the update channel $\bar{\rho}$ of a measurable MDP (see Definition 3.2.9). For that purpose, after an informal introduction in Section 3.3.1, we introduce in Section 3.3.2 previous results on the *ergodic decomposition of (measurable, standard Borel) Markov chains* (Worm et al., 2011) that yield, in short, a partition of the state-space into ergodic components. However, these previous results turn out to not be exactly the ones that we need. This is the reason why we present at length the technical machinery from (Worm et al., 2011), which allows us, in Section 3.3.3, to fine-tune this previously established framework to obtain a form of the decomposition into ergodic components that is adapted to our needs. In Section 3.3.4, we prove that under a continuity assumption, the space of ergodic components has a standard Borel structure — which is useful for the long-term aim of softening the notion of class with information theory in the non-countable case. We then show in Section 3.3.5 that, in the finite case, ergodic components

can be seen as a mean-asymptotic minimal sufficent statistic — which will be useful for their information-theoretic characterisation in Section 3.6.1.

3.3.1 Informal introduction

We want to describe what is known as the *decomposition into ergodic components* of a standard Borel state-space $\mathcal{X}$ (or a subset of "generic points" of $\mathcal{X}$, in some measure-theoretic sense) w.r.t. a Markov chain τ. Let us first broadly outline this concept, and related ones. The decomposition can be seen as the *finest* partition of (the set of "generic points" of) $\mathcal{X}$ into *invariant subsets*, i.e., subsets that contain their "stochastic image" through the action of the Markov chain τ. Each element of this partition is thus, intuitively, an "invariant subset that cannot be broken down into smaller invariant subsets", and is technically called an *ergodic component*. This partition of the state-space is closely linked to the notion of *ergodic measure*, which means, roughly, a *stationary* measure that "puts all the weight" on a single ergodic component. To each ergodic component corresponds a unique ergodic measure, which can be obtained the following way: for any point x in the ergodic component, the sequence of pushed-forward distributions $\left(\frac{1}{n}\sum_{i=0}^{n-1}\tau^i\cdot\delta_x\right)_n \subseteq \Delta_{\mathcal{X}}$ of time averages of the Markov chain's distribution when starting from x (technically known as *Césaro means*) converges, for $n\to+\infty$, to the ergodic component's unique ergodic measure. Ergodic components can thus be seen as sets of points that "have the same attractor", in the sense that the time-averages of the trajectories starting from these points yield the same distribution. Moreover, each stationary measure μ_0 on $\mathcal{X}$ then decomposes as an "average over all ergodic measures", where the "weights" defining this average uniquely define μ_0.

In the case of deterministic transformations, the ergodic decomposition of both stationary measures and the underlying state-space are classic results of ergodic theory (Coudène, 2016). For Markov chains with a countable state-space, analogous results are standard knowledge as well: the set of what is known as *positive recurrent points* (i.e., points that, if visited once, are visited again with a positive asymptotic frequency) can be partitioned into *communicating classes* (i.e., classes of points that can be reached with positive probability in finite time from one another). This partition of the set of positive recurrent points yields a corresponding ergodic decomposition of any stationary probability (Sericola, 2013).

However, in the case of Markov chains on standard Borel spaces, a formulation of the ergodic decomposition theorem that provides the partition of the underlying state-space has been obtained only relatively recently (Worm et al., 2011). In the next section 3.3.2, we start by quoting the relevant results. Even though we aim to communicate the essence of this section to a wider audience, a close reading of it will require some familiarity with functional analysis — i.e., with Banach spaces, their dual spaces, weak topologies, and Bochner integrals. Note also that, for the sake of consistency with other sections, our notations and terminology will not always follow exactly those of (Worm et al., 2011) (we will point out the most important differences).

3.3.2 Previous results

This section does not present any novel contribution of ours, but only previous results from (Worm et al., 2011). We present the underlying technical machinery in extensive details because we will need to fine-tune these results in Section 3.3.3.

The setting of (Worm et al., 2011) sits within an old tradition of seeing *measures as integrals of continuous functions* (Rudin, 1987). In short, for a "well-chosen" set $\mathscr{M}$ of measures on a "well-behaved" Borel measurable space $\mathcal{X}$, one can choose a Banach space $\mathscr{C}$ of "well-behaved" continuous functions $f:\mathcal{X}\to\mathbb{R}$ such that each measure $\mu\in\mathscr{M}$ is uniquely defined by the linear form $f\mapsto\int_{\mathcal{X}} f\,d\mu$ that integrates functions w.r.t. μ. Intuitively, this

allows one to investigate "what a measure is" by studying only "what it does" — i.e., here, we forget the formal definition of a measure, and we only look at how it integrates each function in $\mathscr{C}$. Technically, this means identifying a space of measures with some subset of the *dual* of the Banach space $\mathscr{C}$, which brings in the full power of functional analysis to measure theory.

This "measures as integrals" perspective happens to be particularly relevant to ergodic decomposition theorems. For instance, it defines a natural topology on the set of probability measures, which yields a well-defined notion of convergence of the sequence of Césaro means $\left(\frac{1}{n}\sum_{i=0}^{n-1}\tau^i\cdot\delta_x\right)_n$ mentioned above. Moreover, it is natural to formalise the ergodic decomposition of stationary probabilities as a probability distribution-valued integral. This is made possible precisely by the Banach space structure on probability measures, which allows using the Bochner integral (see Section C.2.3).

Ref. (Worm et al., 2011) relies on a version of this "duality" perspective on measures where:

- $\mathcal{X}$ is a standard Borel space,
- $\mathscr{M} = \mathscr{M}_{\mathcal{X}}$ is the set of signed measures on $\mathcal{X}$,
- $\mathscr{C} = \mathsf{BL}_{\mathcal{X}}$ is the set of *bounded real-valued Lipschitz functions* on $\mathcal{X}$.

The set $\mathsf{BL}_{\mathcal{X}}$ is a Banach space with norm $\|f\|_{\mathsf{BL}} := |f|_{\mathsf{Lip}} + \|f\|_\infty$, where $|f|_{\mathsf{Lip}}$ is the global Lipschitz constant and $\|f\|_\infty$ the sup norm. The dual of $\mathsf{BL}_{\mathcal{X}}$ is denoted by $\mathsf{BL}^*_{\mathcal{X}}$, and it is regarded as a Banach space with the usual dual norm, denoted by $\|\cdot\|^*_{\mathsf{BL}}$. Note that the Dirac measures δ_x belong to $\mathsf{BL}^*_{\mathcal{X}}$ for all $x\in\mathcal{X}$. We can thus consider the subspace

$$\mathcal{X}_{\mathsf{BL}} := \overline{\mathsf{Span}\{\delta_x,\ x\in\mathcal{X}\}} \subseteq \mathsf{BL}^*_{\mathcal{X}},$$

i.e., $\mathcal{X}_{\mathsf{BL}}$ is the topological closure, w.r.t. the norm $\|\cdot\|^*_{\mathsf{BL}}$ in $\mathsf{BL}^*_{\mathcal{X}}$, of the linear span of the Dirac measures δ_x. As a closed subspace of a Banach space, the space $\mathcal{X}_{\mathsf{BL}}$ is itself a Banach space with norm the restriction of $\|\cdot\|^*_{\mathsf{BL}}$ to $\mathcal{X}_{\mathsf{BL}}$, which we will simply denote by $\|\cdot\|$ when there is no ambiguity. The space $\mathcal{X}_{\mathsf{BL}}$ is also separable (for the topology induced by $\|\cdot\|$), which makes it a Polish space — the complete metric being the one induced by the norm $\|\cdot\|$ (see Definition C.2.2). Here, the "measures as integrals" approach mentioned above is embodied by the following fact: each signed measure $\mu\in\mathscr{M}_{\mathcal{X}}$ defines a unique element in $\mathcal{X}_{\mathsf{BL}}$, still denoted by μ, and defined by

$$\langle\mu, f\rangle := \int_{\mathcal{X}} f\,d\mu$$

for all $f\in\mathsf{BL}_{\mathcal{X}}$. More precisely, $\mathscr{M}_{\mathcal{X}}$ identifies in this way to a dense subspace of $\mathcal{X}_{\mathsf{BL}}$. In the following, we will not write explicitly the corresponding bijection, and regard $\mathscr{M}_{\mathcal{X}}$ as a subset of $\mathcal{X}_{\mathsf{BL}}$.

The set $\mathscr{M}^+_{\mathcal{X}}$ of finite positive measures identifies to a closed convex cone of $\mathcal{X}_{\mathsf{BL}}$, which we denote by $\mathcal{X}^+_{\mathsf{BL}}$. The set of probability distributions $\Delta_{\mathcal{X}}\subseteq\mathscr{M}^+_{\mathcal{X}}$ identifies to the closed convex subset of $\mathcal{X}^+_{\mathsf{BL}}$ made of elements of norm 1. In particular, the restriction of the norm metric on $\mathcal{X}_{\mathsf{BL}}$ defines a complete metric on its closed subsets $\mathscr{M}^+$ and $\Delta_{\mathcal{X}}$ — and thus defines a corresponding standard Borel structure. In the following, we see $\Delta_{\mathcal{X}}$ as a standard Borel space with this induced standard Borel structure. Moreover, the push-forward operator $\tau_*:\Delta_{\mathcal{X}}\to\Delta_{\mathcal{X}}$ of a Markov chain $\tau\in\mathscr{K}(\mathcal{X})$ can be uniquely extended to a positive bounded linear operator on the Banach space of signed measures $(\mathscr{M}_{\mathcal{X}},\|\cdot\|)$, still denoted by τ_*: more precisely, for all $\mu\in\mathscr{M}_{\mathcal{X}}$, we have $\|\tau_*\mu\|\leq\|\mu\|$, and $\mu\in\mathscr{M}^+$ implies $\tau_*\mu\in\mathscr{M}^+$. (See Sections 2.1 and 2.2 in (Worm et al., 2011), and references therein, for statements from this paragraph and the previous one; note that in (Worm et al., 2011), channels $\tau\in\mathscr{K}(\mathcal{X})$ are

called *transition probabilities*, and push-forward operators τ_* are *regular Markov operators* — see Section 2.2 there.)

With this framework in place, we now turn to the ergodic properties of stochastic transformations. We recall that $\mu \in \Delta_{\mathcal{X}}$ is *stationary* under τ if $\tau \cdot \mu = \mu$ (see Definition 3.2.8).

Definition 3.3.1. Let $\tau \in \mathscr{K}(\mathcal{X})$. A Borel subset $F \subseteq \mathcal{X}$ is called *invariant* (w.r.t. τ) if $(\tau \cdot \delta_x)(F) = 1$ for all $x \in F$: i.e., intuitively, if "the stochastic image of each point of F through τ remains in F with probability one". A probability measure $\mu \in \Delta_{\mathcal{X}}$ is called *ergodic* (w.r.t. τ) if it is stationary and $\mu(F) = 1$ or $\mu(F) = 0$ whenever F is an invariant set: i.e., if "any invariant subset of $\mathcal{X}$ has either full measure or zero measure". The *n-th Césaro mean* (w.r.t. τ) is the channel $\tau^{(n)} := \frac{1}{n}\sum_{i=0}^{n-1} \tau^n$.

For each $n \in \mathbb{N}$ and $x \in \mathcal{X}$, the Césaro mean $\tau^{(n)} \cdot \delta_x$ of δ_x is the n-time-step average of the distribution of a trajectory starting from x. In particular, if τ is deterministic, this can be seen as an "empirical distribution", computed by an experimentalist recording measurements of the trajectory. The starting point of ergodic theory was to understand under which conditions such an "empirical distribution", computed from time-averages of trajectory, could accurately approximate the underying "spatial" stationary distribution $\mu_0 \in \Delta_{\mathcal{X}}$ over the state-space.[26] The formalism of ergodic theory provided, in short, the following answer for *deterministic* dynamics: this is indeed the case when, up to zero probability sets, there is only one non-trivial invariant set — which can be understood, informally, as the fact that "there is only one attractor". I.e., rephrasing this in the language introduced in Definition 3.3.1: if the initial distribution μ_0 is ergodic w.r.t. τ where τ is deterministic, then for μ_0-a.e. $x \in \mathcal{X}$, the Césaro mean $\tau^{(n)} \cdot \delta_x$ does indeed converge (in some sense) to a unique stationary distribution (Coudène, 2016).

Now, for non-ergodic deterministic dynamics, the situation is more subtle: in short, distinct starting points might lead to distinct limits of the Césaro means, each of them ergodic — which unveils a rich structure of deterministic dynamical systems, described precisely by the notion of *ergodic decomposition* (Coudène, 2016). This suggests that, to exhibit a similar structure for stochastic dynamics — i.e., here, for Markov chains on standard Borel spaces — we may start by considering the set of points whose Césaro means converge to an ergodic measure, i.e.:[27]

$$\mathcal{X}_{\text{erg}} := \left\{ x \in \mathcal{X} \;:\; \left(\tau^{(n)} \cdot \delta_x\right)_{n\in\mathbb{N}} \text{converges in } \mathcal{X}_{\mathsf{BL}} \text{ to an ergodic probability measure } \epsilon_x \right\}. \tag{3.3.1}$$

This set is "generic" in the following, measure-theoretic/dynamic sense:

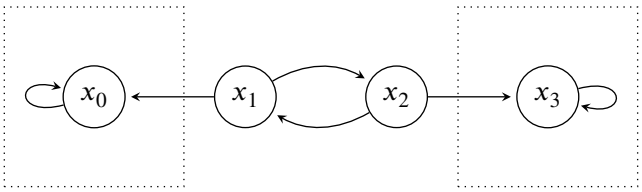
Theorem 3.3.2 ((Worm et al., 2011), Theorem 3.12)**.** *$\mathcal{X}_{\text{erg}}$ is measurable, and $\mu(\mathcal{X}_{\text{erg}}) = 1$ for any τ-stationary measure μ.*

Remark 3.3.3. Assume that $\mathcal{X}$ is countable (i.e., finite or countably infinite). Then it can be verified, using standard knowledge,[28] that $\mathcal{X}_{\text{erg}}$ is made of the positive recurrent points, and the transient points that lead with probability one to a single communicating class of positive recurrent points. The important point, here, is that we encompass all positive recurrent points — indeed, these are the points on which the stationary measures are supported. The fact that $\mathcal{X}_{\text{erg}}$ captures some transient points as well can be seen as an artifact of the tools developed in (Worm et al., 2011) (but, for our purposes at least, a harmless artifact). See Figure 3.1 for illustrative examples — note that the figure also references the objects $\mathcal{C}$, $(\mathcal{X}^c)_{c\in\mathcal{C}}$ and $\mathcal{X}_{\text{erg,inv}}$ which are defined further below.

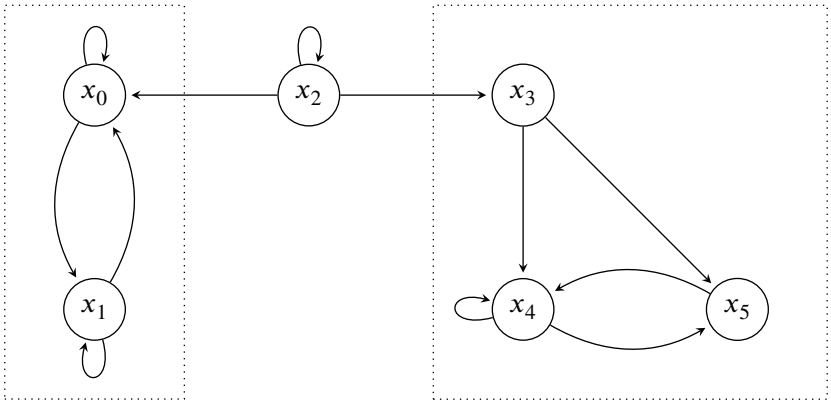
[26] Actually, an experimentalist usually rather records *functions* of the state-space trajectory, whose time-average approximate their integral over the state-space — but here, we simplify for the sake of conciseness.

[27] The set $\mathcal{X}_{\text{erg}}$ is denoted by Γ_{cpie} in (Worm et al., 2011).

[28] In this remark, we assume familiarity with countable Markov chains theory.



(A) Here $\mathcal{X}_{\text{erg}} = \mathcal{X}_{\text{erg,inv}} = \{x_0, x_3\}$, with $\mathcal{C} = \{\mathcal{X}^c\}_{c\in\mathcal{C}} = \big\{\{x_0\}, \{x_3\}\big\}$. The states x_1 and x_2 are transient, and their resp. Césaro means do converge to stationary distributions. But as these states lead to two distincts communicating classes of positive recurrent states, the resp. invariant distributions are not ergodic — which is why they are not included in $\mathcal{X}_{\text{erg}}$ (see Remark 3.3.3). Here $\mathcal{X}_{\text{erg,inv}}$ coincides with the set of positive recurrent points.



(B) Here $\mathcal{X}_{\text{erg}} = \mathcal{X}_{\text{erg,inv}} = \{x_0, x_1, x_3, x_4, x_5\}$, with $\mathcal{C} = \{\mathcal{X}^c\}_{c\in\mathcal{C}} = \big\{\{x_0, x_1\}, \{x_3, x_4, x_5\}\big\}$. The state x_2 is transient but leads to two distincts communicating classes of positive recurrent states, so it is not in $\mathcal{X}_{\text{erg}}$. In contrast, the state x_3 is transient and leads to the unique communicating class $\{x_4, x_5\}$, so its Césaro means converge to the same ergodic distribution as the Césaro means of x_4 and x_5. Therefore it is in $\mathcal{X}_{\text{erg}}$ and, more precisely, in the same equivalence class c as $\{x_4, x_5\}$. This holds even though x_3 is not contained in the set on which invariant probability distributions are supported — i.e., the set of positive recurrent points $\{x_0, x_1, x_4, x_5\}$.

FIGURE 3.1: Examples of Markov chains τ with corresponding generic set and partition in ergodic components, in the sense of (Worm et al., 2011). An arrow from state x_i to state x_j means that $\tau(x_j|x_i) > 0$. Each dotted box is an equivalence class $c \in \mathcal{C}$ (see equation (3.3.2)), which here always coincides with its invariant subset $\mathcal{X}^c$ (see Theorem 3.3.7). Here, their union always coincides with the generic set $\mathcal{X}_{\text{erg}}$ (see equation (3.3.1)), which is the same as its invariant subset $\mathcal{X}_{\text{erg,inv}}$ (see equation (3.3.3)). Note that the captions of Figures 3.1b and 3.1a assume familiarity with countable Markov chains theory.

We then cluster elements of $\mathcal{X}_{\text{erg}}$ according to the limit of their Césaro means. I.e., for each $x \in \mathcal{X}_{\text{erg}}$, denoting by ϵ_x the limit in $\mathcal{X}_{\mathsf{BL}}$ of the Césaro means $\left(\tau^{(n)} \cdot \delta_x\right)_{n\in\mathbb{N}}$, we define the following equivalence relation on $\mathcal{X}_{\text{erg}}$:

$$x \sim x' \text{ if and only if } \epsilon_x = \epsilon_{x'} \tag{3.3.2}$$

The set of equivalence classes defined by $\sim$ on $\mathcal{X}_{\text{erg}}$ is denoted by $\mathcal{C}$, and for $x \in c \in \mathcal{C}$, the probability ϵ_x is also denoted by ϵ^c. See Figure 3.1 for illustrative examples — note that the figure also references the objects $(\mathcal{X}^c)_{c\in\mathcal{C}}$ and $\mathcal{X}_{\text{erg,inv}}$ which are defined further below.

The next theorem lumps together results from Theorem 4.3, Theorem 4.6, Corollary 4.7 and Corollary 4.8 in (Worm et al., 2011). Before stating it, let us point out that for a given $\tau \in \mathscr{K}(\mathcal{X})$, the set of τ-stationary probabilities define a convex subset $\Delta_{\mathcal{X}}^{\tau} \subseteq \Delta_{\mathcal{X}}$, so that it makes sense to consider its extreme points (i.e., those that cannot be obtained as a non-trivial convex combination of distinct points of $\Delta_{\mathcal{X}}^{\tau}$).

Theorem 3.3.4. *Let $\mathcal{X}$ standard Borel, and $\tau \in \mathscr{K}(\mathcal{X})$. Then:*

(i) The set of ergodic measures w.r.t. τ coincides with the set of extreme points of $\Delta_{\mathcal{X}}^{\tau}$.

(ii) Any $c \in \mathcal{C}$ is measurable and satisfies $\epsilon^c(c) = 1$. In particular, $\epsilon^{x'}(c) = 0$ for $c' \neq c$.

(iii) Any ergodic probability μ is of the form $\mu = \epsilon^c$ for some $c \in \mathcal{C}$.

(iv) There is a bijection between the space of ergodic probability measures and the set $\mathcal{C}$.

(v) $\mathcal{X}_{\text{erg}}$ is non-empty if and only if there exists a stationary measure.

Point (ii) is often referred to as the fact that the ergodic measures corresponding to distinct ergodic components are *mutually singular*. Let us now state the Markov chain version of the integral decomposition of stationary probabilities into ergodic probabilities, using the language of Bochner integrals (see Section C.2.3).

Theorem 3.3.5 ((Worm et al., 2011), Theorem 4.10)**.** *Let $\mu \in \Delta_{\mathcal{X}}$ be stationary. Then the map*

$$\begin{aligned} \mathcal{X} &\to \mathcal{X}_{\mathsf{BL}} \\ x &\mapsto \begin{cases} \epsilon_x & \text{if } x \in \mathcal{X}_{\text{erg}} \\ 0 & \text{if } x \notin \mathcal{X}_{\text{erg}} \end{cases} \end{aligned}$$

is Bochner integrable, and

$$\mu = \int_{\mathcal{X}} \epsilon_x d\mu(x) = \int_{\mathcal{X}_{\text{erg}}} \epsilon_x d\mu(x).$$

The next theorem involves the notion of restriction of a Markov chain to an invariant subset:

Definition 3.3.6. Let $(\mathcal{X}, \mathfrak{X})$ be a measurable space, $\tau \in \mathscr{K}(\mathcal{X})$, and E a τ-invariant measurable subset, seen as measurable space with the induced σ-algebra $\mathfrak{X}_E := \{F \cap E,\ F \in \mathfrak{X}\}$. Then the channel $\gamma \in \mathscr{K}(E)$ defined by $\gamma(F \cap E|x) := \tau(F|x)$ for all $F \cap E \in \mathfrak{X}_E$ is called the *restriction* of τ to E.[29]

Theorem 3.3.7 ((Worm et al., 2011), Theorem 4.13)**.** *Let $\mathcal{X}$ standard Borel, $\tau \in \mathscr{K}(\mathcal{X})$, and assume that there exists a stationary probability measure (or equivalently that $\mathcal{X}_{\text{erg}}$ is not empty). Then for all $c \in \mathcal{C}$:*

[29] For details on the fact that γ is a well-defined channel, see Section 4.3 in (Worm et al., 2011).

(i) The set $\mathcal{X}^c := \bigcap_{n\in\mathbb{N}} c_n$, *where* $c_0 := c$, *and*

$$c_n := \{x \in c_{n-1} \,:\, (\tau \cdot \delta_x)(c_{n-1}) = 1\},$$

is an invariant measurable subset of c, *and* $\epsilon^c(\mathcal{X}^c) = 1$.

(ii) ϵ^c *is the only stationary probability measure of* τ^c, *where* τ^c *is the restriction of* τ *to* $\mathcal{X}^c$.

(iii) c *cannot be written as the union of two disjoint* τ^c*-invariant sets* $\mathcal{X}^c_1$ *and* $\mathcal{X}^c_2$ *with* $\epsilon^c(\mathcal{X}^c_1) > 0$ *and* $\epsilon^c(\mathcal{X}^c_2) > 0$.

Each set $\mathcal{X}^c$ can be seen as the "largest invariant subset of c", as will be discussed in further detail in Section 3.3.3.

Proof of Theorem 3.3.7. This is Theorem 4.13 in (Worm et al., 2011), with the only difference that the explicit form of the invariant measurable set $\mathcal{X}^c \subseteq c$ is not stated in the latter result. However, looking at its proof — and those of Lemma 4.1 and Corollary 4.2 in (Worm et al., 2011), on which the theorem's proof relies — shows that we can indeed choose $\mathcal{X}^c$ as stated in point (i) above. □

In a sense, Theorem 3.3.7 completes the formalisation of the statements informally stated in Section 3.3.1. Indeed, it somehow provides a decomposition of (a generic subset $\mathcal{X}_{\text{erg}}$ of) the state-space $\mathcal{X}$ into "minimally τ-invariant subsets $\mathcal{X}^c$" (see points (i) and (iii)), which makes them good candidates for being called "ergodic components". Moreover, to each "ergodic component" $\mathcal{X}^c$ corresponds a unique ergodic measure ϵ^c (see point (ii)), which is concentrated on $\mathcal{X}^c$. Each such ergodic measure is obtained as a limit of Césaro means (by definition of $\mathcal{X}_{\text{erg}}$), i.e., of time-averages of the process' spatial distribution over $\mathcal{X}$ when it starts from a given point. Eventually, Theorem 3.3.5 does provide a decomposition of stationary probabilities as averages over ergodic probabilities.

However, point (i) in Theorem 3.3.7 is somehow unsatisfying. Indeed, the union $\bigsqcup_{c\in C} \mathcal{X}^c$ of the "ergodic components" might not coincide with the "generic" set $\mathcal{X}_{\text{erg}}$, as the partition of the latter is $\bigsqcup_{c\in C} c$, and even though $\epsilon^c(\mathcal{X}^c) = \epsilon^c(c) = 1$ for all $c \in C$, we might have $\mathcal{X}^c \subsetneq c$. So on the one hand, we have a partition $\mathcal{X}_{\text{erg}} = \bigsqcup_{c\in C} c$ of a "generic" set that might not be τ-invariant into subsets that might not be τ-invariant; and on the other hand, we have a τ-invariant disjoint union $\bigsqcup_{c\in C} \mathcal{X}^c$ of τ-invariant subsets $\mathcal{X}^c$, but it is not clear that this union is "generic" in the same way as $\mathcal{X}_{\text{erg}}$ is — or that it is even measurable if C is uncountable. In the next section, we will show that the latter facts do actually hold. In Appendix C.4.1, we gather additional technical results from (Worm et al., 2011) that will be useful for that purpose, or further down in the proofs of results from Sections 3.3 and 3.4.

3.3.3 Fine-tuning of previous results

In Section 3.3.2, we presented in detail the mathematical machinery developed in (Worm et al., 2011). Our own contribution (from Section 3.3) starts here. In the current Section 3.3.3, we fine-tune the results from (Worm et al., 2011) to obtain exactly the facts that we need for our purposes of class-pose decomposition: namely, that there is an *invariant* "generic" set that can be decomposed into an exact, set-theoretic partition in invariant ergodic components. These facts are obtained as relatively straightforward consequences of the results in (Worm et al., 2011), or small changes to their proofs.

First, note that each set $\mathcal{X}^c$ from Theorem 3.3.7 can be seen as the unique *largest invariant subset of* c. Indeed, more generally, let us define, for any measurable set $F \in \mathfrak{X}$, the set

$$\mathtt{Inv}(F) := \bigcap_{n\in\mathbb{N}} F_n, \tag{3.3.3}$$

where each F_n is iteratively defined as $F_0 := F$, and

$$F_n := \{x \in F_{n-1} \,:\, (\tau \cdot \delta_x)(F_{n-1}) = 1\}.$$

Intuitively, each F_n is made of those elements in F_{n-1} whose stochastic image through τ is also in F_{n-1} with probability one. By iteration on n, it is clear that for any invariant subset $F' \subseteq F$, we have $F' \subseteq F_n$ for all $n \in \mathbb{N}$, and thus $F' \subseteq \bigcap_{n\in\mathbb{N}} F_n = \mathtt{Inv}(F)$. Moreover, $\mathtt{Inv}(F)$ is, by construction, invariant; therefore it is an invariant subset of F containing all invariant subsets of F. It is straightforward that a subset satisfying this property must be unique, so that $\mathtt{Inv}(F)$ is the unique maximal element for inclusion among invariant subsets of F: i.e., in short, it is indeed the largest invariant subset of F.

While (Worm et al., 2011) considers the largest invariant subset $\mathcal{X}^c = \mathtt{Inv}(c)$ of each equivalence class $c \subseteq \mathcal{X}_{\mathrm{erg}}$ to obtain Theorem 3.3.7, we will here show that it is better to consider the largest invariant subset *before* decomposing $\mathcal{X}_{\mathrm{erg}}$, i.e., to decompose

$$\mathcal{X}_{\mathrm{erg,inv}} := \mathtt{Inv}(\mathcal{X}_{\mathrm{erg}}),$$

instead of $\mathcal{X}_{\mathrm{erg}}$, into equivalences classes w.r.t. the relation $\sim$ (see (3.3.2)). Indeed, we then obtain the desirable properties mentioned after Theorem 3.3.7: i.e., that $\mathcal{X}_{\mathrm{erg,inv}} = \bigsqcup_{c\in C} \mathcal{X}^c$ and that $\mathcal{X}_{\mathrm{erg,inv}}$ is as "generic" as $\mathcal{X}_{\mathrm{erg}}$. Let us first show the latter fact (compare Proposition 3.3.8 with Theorem 3.3.2).

Proposition 3.3.8. *For any $F \in \mathfrak{X}$, the set $\mathtt{Inv}(F)$ is measurable, and if $\mu(F) = 1$ for some $\mu \in \Delta_{\mathcal{X}}$ then $\mu(\mathtt{Inv}(F)) = 1$. In particular, for all τ-stationary measure $\mu \in \Delta_{\mathcal{X}}$, we have $\mu(\mathcal{X}_{\mathrm{erg,inv}}) = 1$*

Proof. See Appendix C.4.3. □

The equivalence relation $\sim$ on $\mathcal{X}_{\mathrm{erg}}$ (see (3.3.2)), whose classes are the elements $c \in C$, restricts to an equivalence relation on $\mathcal{X}_{\mathrm{erg,inv}} \subseteq \mathcal{X}_{\mathrm{erg}}$, whose classes are the elements $c \cap \mathcal{X}_{\mathrm{erg,inv}}$. These restricted classes happen to coincide with the invariant sets $\mathcal{X}^c := \mathtt{Inv}(c)$ from Theorem 3.3.7:

Proposition 3.3.9. *Let $\mathcal{X}$ standard Borel, $\tau \in \mathscr{K}(\mathcal{X})$ and assume that there exists a stationary measure. Then*

(i) $c \cap \mathcal{X}_{\mathrm{erg,inv}} = \mathcal{X}^c \neq \emptyset$.

(ii) In particular, $\bigsqcup_{c\in C} \mathcal{X}^c = \mathcal{X}_{\mathrm{erg,inv}}$.

Proof. See Appendix C.4.4. □

To summarise the facts presented up to here (i.e., those established in (Worm et al., 2011), combined with the fine-tuning that we added in the current section):

- We have a set $\mathcal{X}_{\mathrm{erg,inv}}$ which is "generic", in the sense of Proposition 3.3.8, but which is now also invariant, so that the Markov chain τ can be restricted to $\mathcal{X}_{\mathrm{erg,inv}}$.
- The elements $(\mathcal{X}^c)_{c\in C}$ provide a partition of $\mathcal{X}_{\mathrm{erg,inv}}$ into invariant measurable subsets $\mathcal{X}^c$. Importantly, we do not need, as in Theorem 3.3.7, to consider subsets of each element of the partition to get the invariance property.

- Each element $\mathcal{X}^c$ of the partition satisfies the properties stated in Section 3.3.2 about $\mathcal{X}^c$ — in particular, it is *minimally invariant*, i.e., it cannot be "broken down" into strictly smaller but non-trivial invariant subsets, and it corresponds to a unique ergodic measure ϵ^c concentrated on $\mathcal{X}^c$ (see Theorem 3.3.7).

From these properties, the partition $(\mathcal{X}^c)_{c\in\mathcal{C}}$ of $\mathcal{X}_{\text{erg,inv}}$ fully deserves the following name:

Definition 3.3.10. The partition $(\mathcal{X}^c)_{c\in\mathcal{C}}$ of the invariant generic set $\mathcal{X}_{\text{erg,inv}}$ is called the *decomposition (or partition) into ergodic component w.r.t.* τ. The *projection on ergodic components* is the map $\kappa : \mathcal{X} \to \mathcal{C}$ such that for all $c \in \mathcal{C}$ and $x \in \mathcal{X}^c$, we have $\kappa(x) = c$.

Technical remark. The set $\mathcal{C}$ itself was defined as a family of subsets of $\mathcal{X}$ (see equation (3.3.2)). But we now "forget" these underlying subsets, and only see $\mathcal{C}$ as the space of "labels" indexing $(\mathcal{X}^c)_{c\in\mathcal{C}}$.

Importantly, this definition of ergodic components is a generalisation of the deterministic notion of ergodic component in standard Borel spaces: see, e.g., Section 14.3 in (Coudène, 2016) for a definition of the latter, and results showing that, on $\mathcal{X}_{\text{erg,inv}}$, it coincides with Definition 3.3.10 if τ is a deterministic channel.

3.3.4 Standard Borel structure on $\mathcal{C}$ for continuous Markov chains

Let us also recall that our broader aim, in this work, is to leverage information theory to discover — and "soften" — classes and poses. In particular, we need the class space $\mathcal{C}$ to have a "nice enough" structure to do information theory on it — namely, a standard Borel structure (Gray, 2009; Worm et al., 2011). In this thesis, we will only do information theory with finite alphabets (see Section 3.6), which are of course always standard Borel. But our perspective is also to lay the groundwork for an information-theoretic treatment in arbitrary standard Borel spaces, which requires a proof that $\mathcal{C}$ is indeed standard Borel.

It turns out that if the Markov chain τ satisfies a specific continuity assumption (which generalises continuity for deterministic functions), the space $\mathcal{C}$ can be seen as a countable intersection of open sets in the topological space of stationary measures, where the latter is identified to a compact subset of the space $\mathcal{X}_{\text{BL}}$. In particular, the subset $\mathcal{C}$ is measurable in the standard Borel space of stationary measures, and thus, from Theorem C.2.6, it is standard Borel itself.

These facts are stated more formally and proven in Appendix C.4.5 — note that this result will not be needed elsewhere. We leave to future work an investigation of the standard Borel structure on $\mathcal{C}$ for general standard Borel Markov chains.

3.3.5 Ergodic components as a mean-asymptotic minimal sufficient statistic

In this section, we turn to a property that will be pivotal for characterising — and softening — ergodic components with the language of information theory in Section 3.6.1. Namely, we show that the decomposition into ergodic components is a minimal sufficient statistic of the mean-asymptotic distribution between the initial point and an iterated point. Proposition 3.3.12 below is a first step towards this result in the general standard Borel case. The remainder of this section, however, introduces the assumption that $\mathcal{X}$ is countable. We leave to future work a complete generalisation of the result to the standard Borel setting.

Remark 3.3.11. The definitions and results from Sections 3.3.2 and 3.3.3 actually only assumed that $\mathcal{X}$ is standard Borel. Thus, the results about the Banach space $\mathcal{X}_{\text{BL}}$ still apply if we replace $\mathcal{X}$ by any other standard Borel space — e.g., $\mathcal{X} \times \mathcal{X}$.

Proposition 3.3.12. *Let τ be a measurable Markov chain with standard Borel state-space $\mathcal{X}$. Then for any probability $\mu \in \Delta_{\mathcal{X}}$ such that $\mu(\mathcal{X}_{\mathrm{erg,inv}}) = 1$, denoting by $q := q(\overrightarrow{X})$ the corresponding process distribution, we have, in the Banach space $(\mathcal{X}\times\mathcal{X})_{\mathsf{BL}}$, the convergence*

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{i=0}^{n-1} q(X_0, X_i) = \int_{\mathcal{X}} \epsilon_x \otimes \epsilon_x \, d\mu(x), \tag{3.3.4}$$

where we used the tensor product notation (see Definition 3.2.6).

Proof. See Appendix C.4.6. □

Equation (3.3.4) means that the limit, for $n \to \infty$, of the joint distribution between the initial state and the time average of the resulting trajectory (left-hand side) coincides with the spatial average, w.r.t. $x \in \mathcal{X}$, of two independent samples w.r.t. the ergodic distribution ϵ_x (right-hand side). I.e., in short: given the ergodic component, the initial state and the asymtotic mean of the resulting trajectory are i.i.d. samples w.r.t. the corresponding ergodic distribution.

We now introduce the assumption that $\mathcal{X}$ is countable. Note that this implies that the space C of ergodic components (made of subsets of $\mathcal{X}$) is countable as well. Proposition 3.3.12 then becomes:

Proposition 3.3.13. *Assume that $\mathcal{X}$ is countable, $\tau \in \mathscr{K}(\mathcal{X})$, fix a probability $\mu \in \Delta_{\mathcal{X}}$ such that $\mu(\mathcal{X}_{\mathrm{erg,inv}}) = 1$, denote by $q = q(\overrightarrow{X})$ the corresponding process distribution, and write also*

$$\bar{q}^n(X, X') := \frac{1}{n} \sum_{i=0}^{n-1} q(X_0, X_i) \in \Delta_{\mathcal{X}\times\mathcal{X}},$$

Then, for all $x, x' \in \mathcal{X}_{\mathrm{erg,inv}}$,

$$\lim_{n\to\infty} \bar{q}^n(x, x') = \bar{q}(x, x'),$$

where

$$\bar{q}(x, x') := \sum_{c\in C} \mu(\mathcal{X}^c)\, \epsilon^c(x)\epsilon^c(x')\delta_{x,x'\in\mathcal{X}^c}. \tag{3.3.5}$$

Proof. See Appendix C.4.6. □

Note that from equation (3.3.5), the distribution $\bar{q} = \bar{q}(X, X')$ is symmetric in X and X': i.e., using the transpose notation (see Definition 3.2.4), we have $\bar{q} = \bar{q}^{\mathsf{T}}$. Moreover, let us denote by $\mathsf{pr} : \mathcal{X} \to C$ the projection on ergodic components, by $(\epsilon^c)_{c\in C}$ the corresponding family of ergodic measures, and by $\epsilon \in \mathscr{K}(C, \mathcal{X})$ the channel defined by $\epsilon(x|c) := \epsilon^c(x)$. Equation (3.3.5) can then be rewritten

$$\bar{q} := \mu(\epsilon \circ \mathsf{pr}), \tag{3.3.6}$$

where we use the hook-up notation (see Definition 3.2.3). This equation means that $\bar{q}$ coincides with the joint distribution obtained by combining the input distribution $\mu \in \Delta_{\mathcal{X}}$ with the channel that composes the projection on ergodic components pr with the channel ϵ that samples according to the ergodic distribution ϵ^c defined by its input $c \in C$. As we will now see, this shows that under the assumptions of Proposition (3.3.13), the joint distribution $\bar{q}(X, X')$

between the initial state and the resulting asymptotic mean makes the decomposition into ergodic components pr a *sufficient statistics* between X and X', in the sense of the following definition.

Definition 3.3.14. Let $\mathcal{A}$, $\mathcal{B}$ countable sets and $q_{AB} \in \Delta_{\mathcal{A}\times\mathcal{B}}$, consider a map $f : \mathcal{A} \to C$ to a countable set C, and denote by $q_{AB}f \in \Delta_{\mathcal{A}\times\mathcal{B}\times C}$ the joint distribution defined by q_{AB} and f: i.e., for all $a \in \mathcal{A}$, $b \in \mathcal{B}$, $c \in C$,

$$(q_{AB}f)(a,b,c) := q_{AB}(a,b)\delta_{c=f(a)}.$$

Then the map f is a called *sufficient statistic* of A w.r.t. B if under the distribution $q_{AB}f$, we have the Markov chain $A - C - B$.

Intuitively, the function f is a sufficient statistics of A w.r.t. B if it implements a coarse-graining of A that does not loose any of the information that A carries about B. It can be easily verified that this is the case if and only if the joint distribution μ_{AB} can be obtained by combining the marginal μ_A on $\mathcal{A}$ with the composition of $f : \mathcal{A} \to C$ and a well-chosen channel $\gamma \in \mathscr{K}(C, \mathcal{B})$. I.e., more formally:

Proposition 3.3.15. *Let $\mathcal{A}$, $\mathcal{B}$, C countable sets,let $q_{AB} \in \Delta_{\mathcal{A}\times\mathcal{B}}$ and $f : \mathcal{A} \to C$. Then f is a sufficient statistic of A w.r.t. B if and only if there exists a channel $\gamma \in \mathscr{K}(C, \mathcal{B})$ such that, denoting by q_A the marginal of q_{AB} on $\mathcal{A}$ and using the hook-up notation (see Definition 3.2.3), we have $q_{AB} = q_A(\gamma \circ f)$.*

Proof. See Appendix C.4.6. □

Thus, from Proposition 3.3.15 and equation (3.3.6), under the joint distribution $\bar{q} \in \Delta_{\mathcal{X}\times\mathcal{X}}$ the projection pr on ergodic components is indeed a sufficient statistic of X w.r.t. X'. Moreover, from the symmetry $\bar{q} = \bar{q}^\top$, we obtain similarly that pr is a sufficient statistic of X' w.r.t. X. As we will now see, these sufficient statistics happen to also be *minimal*, in the following sense:

Definition 3.3.16. Let $\mathcal{A}$, $\mathcal{B}$, C, C' countable sets, $q_{AB} \in \Delta_{\mathcal{A}\times\mathcal{B}}$, and q_A the marginal of q_{AB} on $\mathcal{A}$. A sufficient statistic $f : \mathcal{A} \to C$ of A w.r.t. B is called a *minimal sufficient statistic* if for any other sufficient statistic $f' : \mathcal{A} \to C'$, there exists a function $h : C' \to C$ such that $f(a) = (h \circ f')(a)$ for all $a \in \mathcal{A}$ such that $q_A(a) > 0$.

Intuitively, a minimal sufficient statistic of A w.r.t. B is the coarsest coarse-graining of A that does not loose any of the information that A carries about B. The reason why we require the equality $f(a) = (h \circ f')(a)$ only if $q_A(a) > 0$ is because we do not want our definition to depend on the way f maps symbols $a \in \mathcal{A}$ that have probability 0 under μ.

The following statement again uses, several times, the hook-up notation (see Definition 3.2.3):

Proposition 3.3.17. *Let $\mu \in \Delta_{\mathcal{X}}$ such that $\mu(\mathcal{X}_{\mathrm{erg,inv}}) = 1$, define $\bar{q} \in \Delta_{\mathcal{X}\times\mathcal{X}}$ as in (3.3.6), and let $\mathcal{T}$ be a countable space. Then for all $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$ such that $\bar{q} = \mu(\gamma \circ \kappa)$ for some channel $\gamma \in \mathscr{K}(\mathcal{T}, \mathcal{X})$, there exists a function $h : \mathcal{T} \to C$ such that $\mu\mathsf{pr} = \mu(h \circ \kappa)$.*

Proof. See Appendix C.4.6. □

Proposition 3.3.17 says that under the assumption $\mu(\mathcal{X}_{\mathrm{erg,inv}}) = 1$ (which in particular holds if μ is τ-stationary, see Proposition 3.3.8), the equality $\bar{q} = \mu(\gamma \circ \kappa)$, which corresponds for deterministic channels $\kappa = f$ to being a sufficient statistics (see Proposition 3.3.15),

implies that the combining the initial distribution $\mu \in \Delta_{\mathcal{X}}$ with the deterministic channel $\mathsf{pr} \in \mathscr{K}(\mathcal{X}, C)$ defined by the projection on ergodic components is the same as combining μ with the composition of $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$ with a well-chosen deterministic channel $h \in \mathscr{K}(\mathcal{T}, C)$.[30]

This result yields our claim from the begining of this section:

Theorem 3.3.18. *Under the mean-asympotic distribution $\bar{q}(X, X')$, the projection on ergodic components* pr *is a minimal sufficient statistic of X w.r.t. X', and of X' w.r.t. X.*

Proof. See Appendix C.4.6. □

Crucially, it has been shown that the Information Bottleneck (IB) method implements precisely, for maximal trade-off parameter, a minimal sufficient statistic of the source variable w.r.t. the relevancy variable (Shamir et al., 2010). Thus Theorem 3.3.18 yields a characterisation of ergodic components with the IB method — or more precisely, two equivalent characterisations, due to the symmetry $\bar{q} = \bar{q}^{\mathsf{T}}$. However, choosing one of the two variables X or X' to be the "source", and the other one to be the "relevancy", seems here unnatural. Moreover, the symmetry $\bar{q} = \bar{q}^{\mathsf{T}}$ will in general break once the asymptotic time average $\bar{q}$ is replaced by a finite-time average $\bar{q}_n$ — which is likely to be necessary in concrete implementations. In Section 3.6.1, we will characterise ergodic components with another variant of the IB framework that does not require to artificially discriminate between a source and a relevancy, and that we thus expect to behave better once $\bar{q}$ is replaced by $\bar{q}_n$ for finite $n \in \mathbb{N}$.

3.4 Ergodic decomposition of MDPs with fixed policy

In this section, we apply the framework presented in Section 3.3 to the Markov chain defined by the update channel $\bar{\rho}$ of a *measurable* MDP (π, ρ) (see Definition 3.2.9). We show, in Section 3.4.1, that the decomposition into ergodic components of the state-space induces a family of *stationary* MDPs $(\epsilon^c, \pi^c, \rho^c)$, each with state-space the corresponding ergodic component $\mathcal{X}^c$, defined by a corresponding restricted policy π^c and transition channel ρ^c, and equiped with the corresponding ergodic initial distribution ϵ^c. This family of *ergodic MDPs* provides an ergodic decomposition of the original MDP's process — in a sense analogous to what it means for Markov chains. To the best of our knowledge, such an ergodic decomposition of MDPs has not yet been considered in the literature — even though it must be acknowledged that the core of the ergodic-theoretic work for this result on standard Borel MDPs resides in the results on standard Borel Markov chains previously obtained in (Worm et al., 2011) (see Section C.4.1). This ergodic decomposition of MDPs will be at the basis of our generalisation of the "pose" coordinate in Section 3.5. Before turning to it, though, we apply in Section 3.4.2 the MDP ergodic decomposition to the group-theoretic setting, which yields the equivalence of orbits and ergodic components for groups with a stationary probability.

3.4.1 General result

We fix a measurable MDP (π, ρ) whose state-space $(\mathcal{X}, \mathfrak{X})$ and action space $(\mathcal{G}, \mathfrak{G})$ are both standard Borel. Recall that the update channel is $\bar{\rho} := \rho \circ (\mathsf{Id}_{\mathcal{X}} \rtimes \pi)$, i.e., $\bar{\rho}$ is the average over actions $\mathcal{G}$ of the transition channel $\rho \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X})$, w.r.t. the policy $\pi \in \mathscr{K}(\mathcal{X}, \mathcal{G})$ (see Definition 3.2.9). As $\bar{\rho} \in \mathscr{K}(\mathcal{X})$, we can apply the results from previous section to $\tau := \bar{\rho}$. In particular, we obtain:

- A $\bar{\rho}$-invariant measurable set $\mathcal{X}_{\mathrm{erg,inv}} \subseteq \mathcal{X}$, which has probability 1 under any $\bar{\rho}$-stationary measure (see Proposition 3.3.8).

[30] Here we identify a measurable function with the deterministic channel that it defines (see Definition 3.2.2).

- A partition $\{\mathcal{X}^c\}_{c\in C}$ of $\mathcal{X}_{\text{erg,inv}} \subseteq \mathcal{X}$ in "ergodic components": i.e., each $\mathcal{X}^c$ is measurable, $\bar{\rho}$-invariant, and cannot be divided into smaller non-trivial $\bar{\rho}$-invariant subsets (points (i), (iii) in Theorem 3.3.7, and Proposition 3.3.9).
- Each $\mathcal{X}^c$ carries a unique stationary probability ϵ^c with $\epsilon^c(\mathcal{X}^c) = 1$, which is ergodic (point (i) in Theorem 3.3.7); and each stationary $\mu \in \Delta_{\mathcal{X}}$ decomposes as an average over ergodic probabilities ϵ^c (Theorem 3.3.5).
- If $\bar{\rho}$ is continuous (in a generalised sense that coincides with the usual notion of continuity for deterministic maps but also encompasses stochastic ones), then the space of "labels" C is itself standard Borel (Theorem C.4.5).

Let us now also recall that here, we are interested only in information captured by probability measures μ_0 that stationary w.r.t. the update channel $\bar{\rho}$. But from the first point above, any stationary MDP (μ_0, π, ρ) such that $\mu_0(\mathcal{X}_{\text{erg,inv}}) = 1$ defines a process distribution satisfying $\mu(X_n \in \mathcal{X}_{\text{erg,inv}}) = 1$ for all $n \in \mathbb{N}$. In this sense, the following assumption yields no loss of generality:

Assumption 1. We have $\mathcal{X} = \mathcal{X}_{\text{erg}} = \mathcal{X}_{\text{erg,inv}}$: i.e., equivalently, for all $x \in \mathcal{X}$, the sequence of Césaro means $\bar{\rho}^{(n)} \cdot \delta_x$ converges, in the space $\mathcal{X}_{\mathsf{BL}}$, to an ergodic probability distribution (see equations (3.3.1) and (3.3.3)).

Under this new assumption, $(\mathcal{X}^c)_{c\in C}$ becomes a measurable partition of $\mathcal{X}$ such that each component $\mathcal{X}^c$ is invariant under the update channel $\bar{\rho}$, i.e., under the stochastic transformation defined by the average of actions $\rho_g \in \mathscr{K}(\mathcal{X})$ over $g \in \mathcal{G}$. However, to design our pose coordinate, we would like the whole, *unaveraged* MDP (π, ρ) to "decompose" into a family of MDPs restricted to each component $\mathcal{X}^c$. Indeed, we could then, similarly as in Section 3.1, design a notion of "minimal joining" of these restricted MDPs.

In general, the $\bar{\rho}$-invariance of each $\mathcal{X}^c$ does not imply its invariance under each ρ_g (e.g., if $\mathcal{G} = \mathcal{G}_0 \sqcup \mathcal{G}_1$ and $\pi(\mathcal{G}_0|x) = 0$ for all $x \in \mathcal{X}$, then we can change ρ_g arbitrarily for $g \in \mathcal{G}_0$ without changing $\bar{\rho}$). Despite this fact, the next theorem shows that there exists a family of MDPs $(\pi^c, \rho^c)_{c\in C}$ restricted to the ergodic components that "decomposes" the original MDP (π, ρ), in the sense that given a stationary initial distribution, the process distribution of (π, ρ) can always be recovered by integrating those of the fammily $(\pi^c, \rho^c)_{c\in C}$.

Before stating the theorem, let us point out that as each $\mathcal{X}^c$ is measurable subset of the standard Borel space $\mathcal{X}$, from Theorem C.2.6, each $\mathcal{X}^c$ is itself a standard Borel space.

Theorem 3.4.1. *Let (π, ρ) be a standard Borel measurable MDP, such that Assumption 1 holds. Denote by $\overrightarrow{\pi\rho} \in \mathscr{K}(\mathcal{X}, \overrightarrow{\mathcal{G}\times\mathcal{X}})$ the corresponding process channel (see Definition 3.2.12). Let $(\mathcal{X}^c)_{c\in C}$ be the decomposition into ergodic components, $\kappa : \mathcal{X} \to C$ the corresponding projection, and $(\pi^c)_{c\in C}$ the family of restrictions of π to each $\mathcal{X}^c$. Then there exists a family $(\rho^c)_{c\in C}$, where $\rho^c \in \mathscr{K}(\mathcal{X}^c \times \mathcal{G}, \mathcal{X}^c)$ for all $c \in C$, such that:*[31]

(i) Denoting by $\overrightarrow{\pi^c\rho^c} \in \mathscr{K}(\mathcal{X}^c, \overrightarrow{\mathcal{G}\times\mathcal{X}^c})$ the corresponding process channel of each measurable MDP (π^c, ρ^c), the map

$$\begin{aligned} \mathcal{X} &\to \Delta_{\overrightarrow{\mathcal{X}\times\mathcal{G}}} \subseteq (\overrightarrow{\mathcal{X}\times\mathcal{G}})_{\mathsf{BL}} \\ x &\mapsto \epsilon^{\kappa(x)}\, \overrightarrow{\pi^{\kappa(x)}\rho^{\kappa(x)}} \end{aligned}$$

[31] Here, we identify distributions and channels to their restriction or extension to the relevant space (see Definitions C.2.8 and 3.2.2).

is Bochner integrable, and for all stationary $\mu_0 \in \Delta_{\mathcal{X}}$,

$$\mu_0 \overrightarrow{\pi\rho} = \int_{\mathcal{X}} \epsilon^{\kappa(x)} \overrightarrow{\pi^{\kappa(x)}\rho^{\kappa(x)}} d\mu_0(x). \tag{3.4.1}$$

In particular, for all $c \in \mathcal{C}$, *we have* $\epsilon^c \overrightarrow{\pi\rho} = \epsilon^c \overrightarrow{\pi^c \rho^c}$.

(ii) For all $c \in \mathcal{C}$, *denote by* $\overline{\rho^c} := \rho^c \circ (\mathrm{Id}_{\mathcal{X}^c} \rtimes \pi^c)$ *the update channel of each measurable MDP* (π^c, ρ^c), *and recall that* $\bar{\rho}^c$ *is the restriction of* $\bar{\rho} \in \mathscr{K}(\mathcal{X})$ *to* $\mathcal{X}^c$. *Then* $\overline{\rho^c} = \bar{\rho}^c$ *holds* ϵ^c*-a.e.. In particular,* $\epsilon^c \in \Delta_{\mathcal{X}^c}$ *is the unique stationary distribution w.r.t.* $\overline{\rho^c}$, *and it is ergodic w.r.t.* $\overline{\rho^c}$.

Moreover, an arbitrary family $(\tilde{\rho}^c)_{c\in\mathcal{C}}$ *satisfies points* (i) *and* (ii) *above if and only if for all* $c \in \mathcal{C}$, *we have* $\tilde{\rho}^c \in \mathscr{K}(\mathcal{X}^c \times \mathcal{G}, \mathcal{X}^c)$ *and* $\tilde{\rho}^c = \rho$ *holds* $\epsilon^c \pi^c$*-a.e..*

Proof. See Appendix C.5.1. □

Technical remark. As in (3.4.1), the term under the integral is constant on each $\mathcal{X}^c$, the integral can intuitively be understood as one over $\mathcal{C}$. However, at this stage, it would not be formally justified to write a Bochner integral over $\mathcal{C}$: indeed, we did not prove that, in general, the set $\mathcal{C}$ is itself a standard Borel space, so we cannot consider the Banach space "$\mathcal{C}_{\mathsf{BL}}$" (see, however, Section 3.3.4 for first steps in this direction).

Crucially, point (i) shows that for any stationary initial distribution, the process distribution $q = \mu_0 \overrightarrow{\pi\rho}$ decomposes as an integral, over μ_0, of the process distributions $\epsilon^{\kappa(x)} \overrightarrow{\pi^{\kappa(x)}\rho^{\kappa(x)}}$ describing the time evolutions of the corresponding stationary MDPs $(\epsilon^{c(x)}, \pi^{c(x)}, \rho^{c(x)})$ with state-space the corresponding ergodic components $\mathcal{X}^{c(x)}$. Point (ii) shows that each of the measurable MDPs (π^c, ρ^c) is *ergodic*, in the sense that it has a unique stationary distribution ϵ^c w.r.t. to the update channel $\bar{\rho}$, which is ergodic and coincides with the one from the state-space's ergodic decomposition. Moreover, any family $(\rho^c)_{c\in\mathcal{C}}$ satisfying point (i) defines the same family of stationary process distributions $(\epsilon^c \overrightarrow{\pi^c \rho^c})_{c\in\mathcal{C}}$: i.e., the decomposition is unique if we see MDPs in terms of process distributions.

Importantly, the last part of the statement means that each "restricted" transition channel ρ^c is obtained by modifying the MDP's transition channel $\rho \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X})$ on a set of state-action pairs (x, g) that has null probability w.r.t. to the joint distribution $\epsilon^c \pi^c \in \Delta_{\mathcal{X}\times\mathcal{G}}$ defined by the ergodic distribution ϵ^c and the restricted policy π^c, *in such a way that the ergodic component* $\mathcal{X}^c$ *becomes invariant under the modified version of* ρ.

In light of these properties, the family $(\epsilon^c, \pi^c, \rho^c)_{c\in\mathcal{C}}$ seems to deserve the name of *ergodic decomposition*:

Definition 3.4.2. Let (π, ρ) a standard Borel measurable MDP with state-space $\mathcal{X}$ and action space $\mathcal{G}$, satisfying Assumption 1. A family of standard Borel measured MDPs $(\epsilon^c, \pi^c, \rho^c)_{c\in\mathcal{C}}$, each with state-space $\mathcal{X}^c$ and action space $\mathcal{G}$, is called a *ergodic decomposition* of the measurable MDP (π, ρ) if it satisfies the conclusions of Theorem 3.4.1 — i.e., equivalently, if for all $c \in \mathcal{C}$, we have $\rho^c \in \mathscr{K}(\mathcal{X}^c \times \mathcal{G}, \mathcal{X}^c)$ and $\rho^c = \rho$ holds $\epsilon^c \pi^c$-a.e..

However, this decomposition is valid *only as long as one keeps using the same policy* π. This dependence on the policy is a crucial feature of our MDP ergodic decomposition, which has important implications for potential future work using this tool to study embodied agents.

Relevance to sensorimotor perception in embodied agents The policy-dependency of our MDP ergodic decomposition might be either a limitation or an advantage, depending on the use-case. It is a limitation in the sense that we would like to also have a "universal ergodic decomposition" of the transition channel ρ, i.e., one that holds whatever the choice of

actions. However, from an agent perspective, an ergodic decomposition that depends on a behaviourally relevant policy might identify, precisely, *behaviourally-relevant structure* in the environment (here seen as the state-space $\mathcal{X}$). In other words, to capture the structure of the agent-environment interaction, it should be natural to focus on what the agent *actually does*, rather than zooming out over everything it could do — e.g., this is necessary for any kind of structure involving closed-loop behaviour. E.g., from a sensorimotor perspective (seeing $\mathcal{X}$ as an environment fully observed by an agent taking actions $\mathcal{G}$), the set of all policy-dependent decompositions in ergodic components (equipped with their corresponding ergodic MDPs), over all possible policies, can be seen as one of the potentially fundamental structures of the agent's *sensorimotor habitat*, defined in (Buhrmann et al., 2013) as *"the set of all sensorimotor trajectories that can be generated by the closed-loop [agent-environment] system"*. Indeed, ergodic components capture, here, the features of the sensory space that remain invariant under a specific kind of behaviour. For policies that are in some sense behaviourally relevant, the corresponding decompositions in ergodic component could be seen as an aspect of the agent's *sensorimotor coordination*, defined in (Buhrmann et al., 2013) as *"SMCs described by co-dependencies between [sensors and motors states] that reliably contribute to functionality"*. We leave to future work more concrete investigations of how our MDP ergodic decomposition, or extensions of it, could contribute to these debates.

3.4.2 Application to actions of groups with stationary probability

Here, we show here that the ergodic decomposition described above encompasses the previously proposed, group-based notion of "class", at least for a large class of group that includes compact groups. I.e., we prove, in short, the following: if an action ρ of a group $\mathcal{G}$ on $\mathcal{X}$ has a (necessarily unique) stationary probability υ, then for the MDP with transitions ρ defined by the group action and policy π defined by υ, *the ergodic components $\mathcal{X}^c$ are exactly the orbits under the group action*. The "class" coordinate from the group-based class-pose decomposition framework is thus reframed into a more flexible MDP framework — where, in particular, we can now also deal with non-invertible and stochastic actions, which is closer to a realistic description of real-world agents.

As in previous sections, we consider a measurable MDP (π, ρ) with $\mathcal{X}$ and $\mathcal{G}$ standard Borel. However, here, we do *not* assume, directly, that $\mathcal{X} = \mathcal{X}_{\text{erg,inv}}$. Rather, it will be a consequence of settling ourselves in a group-theoretic setting.

Definition 3.4.3. A *measurable group* is a measurable space $(\mathcal{G}, \mathfrak{G})$ together with a group structure on $\mathcal{G}$ such that the group law and inverse map are measurable. The identity of the group is denoted by e.

Definition 3.4.4. Let $\mathcal{G}$ be a measurable group. A *group-stationary probability*[32] on $\mathcal{G}$ is a probability $\upsilon \in \Delta_{\mathcal{G}}$ such that $\upsilon(gF) = \upsilon(Fg) = \upsilon(F)$ for all $g \in \mathcal{G}$, $F \in \mathfrak{G}$. When this yields no ambiguity, we will refer to a group-stationary probability merely as a *stationary probability*.

Definition 3.4.5. Let $\mathcal{X}$ be a measurable space and $\mathcal{G}$ a measurable group. A *measurable action* of $\mathcal{G}$ on $\mathcal{X}$ is a measurable function

$$\begin{aligned} \mathcal{X} \times \mathcal{G} &\to \mathcal{X} \\ (x, g) &\mapsto g \cdot x \end{aligned}$$

such that for all $g, g' \in \mathcal{G}$, $x \in \mathcal{X}$, we have $(gg') \cdot x = g \cdot (g' \cdot x)$ and $e \cdot x = x$. We say that the transition channel ρ of a measurable MDP *defines a group action* if ρ is the deterministic channel defined by a measurable action. Moreover, for any distribution $\upsilon \in \Delta_{\mathcal{G}}$ on the group

[32] The usual terminology is *invariant* probability.

$\mathcal{G}$, the *independent policy* defined by υ is the channel $\pi_\upsilon \in \mathscr{K}(\mathcal{X}, \mathcal{G})$ such that $\pi_\upsilon(\cdot|x) := \upsilon$ for all $x \in \mathcal{X}$.

In short, the independent policy π_υ makes actions $g \in \mathcal{G}$ sampled from the group's stationary distribution υ, independently from the input state $x \in \mathcal{X}$.

Let $\mathcal{X}$, $\mathcal{G}$ standard Borel spaces such that $\mathcal{G}$ is a measurable group with a stationary probability υ, and $\rho \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X})$ a measurable action. We consider the MDP (π_υ, ρ). For each $x \in \mathcal{X}$, $F \in \mathfrak{X}$, we define

$$\mathcal{G}_{x \to F} := \{g \in \mathcal{G} : \ g \cdot x \in F\},$$

i.e., $\mathcal{G}_{x \to F}$ is made of the group elements g that send x to F. The update channel is then given by

$$\begin{aligned} \bar{\rho}(F|x) &= \int_{\mathcal{G}} \rho_g(F|x) d\pi(g|x) \\ &= \int_{\mathcal{G}} \delta_{g \cdot x \in F} d\upsilon(g) \\ &= \upsilon(\mathcal{G}_{x \to F}) \\ &=: \epsilon_x(F), \end{aligned} \tag{3.4.2}$$

where the last line *defines* the probability measure $\epsilon_x \in \Delta_{\mathcal{X}}$ (we will see below that this notation happens to be consistent with that of previous sections). Intuitively, "the more elements g send x to F, the more probable F is w.r.t ϵ_x".

Moreover, we denote by $[x]$ the orbit of a point $x \in \mathcal{X}$ under $\mathcal{G}$, i.e.,

$$[x] := \{g \cdot x, \ g \in \mathcal{G}\}.$$

Eventually, let us recall that $\bar{\rho}^{(n)} \in \mathscr{K}(\mathcal{X})$ denotes the n-th Césaro mean of the update channel $\bar{\rho}$ (see Definition 3.3.1).

Theorem 3.4.6. *Let $\mathcal{X}$, $\mathcal{G}$ standard Borel spaces such that $\mathcal{G}$ is a measurable group with a group-stationary probability $\upsilon \in \Delta_{\mathcal{G}}$, let $\rho \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X})$ a measurable action, and consider the measurable MDP (π_υ, ρ). Then, for ϵ_x defined in (3.4.2):*

(i) For $x \in \mathcal{X}$ and all $n \geq 1$, we have $\bar{\rho}^n \cdot \delta_x = \epsilon_x$; in particular, $\lim_{n \to \infty} \bar{\rho}^{(n)} \cdot \delta_x = \epsilon_x$ in $\mathcal{X}_{\mathsf{BL}}$.

(ii) $\mathcal{X} = \mathcal{X}_{\text{erg,inv}}$,

(iii) The partition in ergodic components $\{\mathcal{X}^c\}_{c \in C}$ coincides with the partition in orbits of $\mathcal{X}$ w.r.t. the action of $\mathcal{G}$

(iv) For all $x \in \mathcal{X}^c = [x]$, the probability ϵ_x is the unique $\bar{\rho}$-stationary probability such that $\epsilon_x(\mathcal{X}^c) = 1$, and it is ergodic w.r.t. $\bar{\rho}$.

(v) A probability μ on $\mathcal{X}$ is is $\bar{\rho}$-stationary if and only if it is ρ_g-stationary for all $g \in \mathcal{G}$.

(vi) The probability υ is the unique group-stationary probability on $\mathcal{G}$.

(vii) Denoting by ρ^c the restriction of the group action ρ to each orbit $\mathcal{X}^c$, the family of MDPs $(\epsilon^c, \pi^c_\upsilon, \rho^c)$ is an ergodic decomposition of the measurable MDP (π_υ, ρ).

Proof. See Appendix C.5.2. □

Point (i) shows that each limit of Césaro means ϵ_x is here reached after a single time-step: informally, this is because, here, single time-step actions include the whole group $\mathcal{G}$, so that their compositions into multiple time-step actions are already all contained in one time-step. Point (ii) shows that what was Assumption 1 in Section 3.4 is here a consequence of the group-theoretic setting. Crucially, point (iii) shows our claim from the beginning of this section: in short, under the above assumptions, the orbits are the ergodic components. Point (iv) shows that our notation ϵ_x from equation (3.4.2) is consistent with the notation from previous sections — compare with point (ii) in Theorem 3.3.7. Note that, of course, in the proof below that establishes this fact, ϵ_x refers to the probability defined in (3.4.2), and nothing else. Point (v) will be useful to prove that our "pose" coordinate, which will be defined in the next section, does coincide with previous group theory-based notions of pose. Point (vi) shows that for standard Borel measurable groups, a stationary probability υ is always unique. Point (vii) shows that the ergodic decomposition of MDPs exhibited in Section 3.4 is a measure-theoretic generalisation of the family of restrictions of a group action to its orbits (note that from point (vi), the independent policy π^c_υ is uniquely defined by $\mathcal{G}$, and thus ϵ^c is uniquely defined by $\mathcal{G}$ and ρ^c).

An important example of measurable groups with a (unique) stationary measure is that of *compact* topological groups. Indeed, it is well-known that a compact group has a (unique) stationary probability measure satisfying some standard regularity assumption, known as the *Haar probability measure*. Thus, the assumptions of Theorem 3.4.6 encompass all standard Borel compact groups — e.g., finite groups or compact Lie groups — acting measurably (in particular, continuously) on standard Borel spaces — e.g., countable spaces, Euclidean spaces or differentiable manifolds. These cases encompass the groups usually considered in the class-pose decomposition literature (Marchetti et al., 2023; Oizumi et al., 2025; Pérez Rey et al., 2023; Winter et al., 2022).

Note, though, that many groups do not have a stationary probability measure. This is for instance the case of $\mathbb{Z}$ seen as an additive group, which indexes time in the most standard setting of ergodic theory: invertible, stationary dynamical systems (see Definition 3.2.7).[33] However, for groups that do not have a stationary probability, we claim that orbits are *not* a good formalisation of the intuition of class. Let us recall, for instance, the example considered in Section 3.1.4: the group $\mathcal{G} = \{g^n, n \in \mathbb{Z}\}$ generated by a rotation g of irrational angle on the unit circle $\mathbb{S}^1$. We saw that in this case, there is an uncountable number of equally dense orbits, while we would expect a good notion of class to provide, in this case, a unique class given by the whole sate-space $\mathbb{S}^1$. Crucially, this is indeed what we obtain if we define classes as ergodic components: i.e., $\mathbb{S}^1$ is the unique ergodic component of the stationary dynamical system defined by g (Coudène, 2016) (where the stationary distribution is the Lebesgue measure on $\mathbb{S}^1$). More generally, orbits might create artificial distinctions between points that have the same asymptotic behaviour, while ergodic components capture this asymptotic behaviour by clustering points according to the limit of their Césaro means.

Moreover, from the perspective of modeling agents, let us recall that we are here interested in capturing the structure that arises from the agent's *behaviour*. Whatever form this structure may take — group-theoretic or otherwise — this means that our modeling should be dynamical at core. While it is worth noting that abstracting away the arrow of time can be a fruitful point of view in ergodic theory — which is now defined in many textbooks as the study of measurable group actions (Glasner, 2003; Kerr et al., 2016), this "atemporal" point of view is not necessarily adapted to the study of the structure of embodied agents' behaviour. Indeed, we regard the asymptotic notion of ergodic component as a building block to capture, in future work, structure arising in finite time, and where, e.g., the policy could also vary

[33] More commonly known as *invertible probability-preserving transformations*, and studied, e.g., in (Coudène, 2016).

along time. It thus seems at least premature — if not fundamentally limited — to maintain a point of view that disregard the dynamics' explicit time evolution.

This ergodic reformulation of the group-based notion of class could, however, be made more complete. In particular, a proper "unfolding along time" of the group structure should regard the space $\mathcal{G}$ of one time-step actions not as the whole group, but only as *generators* of the group: i.e., as a subset of group elements whose iterated multiplications can yield any other group element. Future work could thus aim at results similar to Theorem 3.4.6, but with $\mathcal{G}$ now made of only group generators.

Related work Our reformulation of group actions in a dynamical context, here disrete-time stationary MDPs, resonates with the *flow equivariance* framework (Keller, 2025; Keller et al., 2026; Lillemark et al., 2025). Indeed, this line of work proposes to consider group-equivariant neural networks where the group action is not made of spatial transformations of a static input (as, e.g., classic Lie groups), but is induced by the time evolution of spatiotemporal dynamics on both the "external" space and the "latent" space. Importantly, the time evolution of the "external" variable is proposed in (Keller et al., 2026) to be interpreted as the dynamics induced by the *movement* of either the agent or the environment — a perspective inspired from ecological theories of perception (Gibson, 2014) (see the subsection "Towards a confluence of algebraic, dynamical & informational approaches to SMCs?" in Section 1.2.3). Besides the fact that it does not consider class or pose variables, the main differences between the flow equivariance framework and our MDP reformulation of group actions, at the formal level, is that while the former considers a *specific kind* of group action (the flow of a differential equation), the latter is a *generalisation*, in discrete time, of the action of stationary groups to a potentially non-invertible, stochastic and closed-loop setting. In particular, our framework encompasses discrete-time stationary dynamical systems (choose a trivial action space $\mathcal{G}$), which includes the time-discretisation of measure-preserving flows of differential equations.

3.5 Minimal joinings and minimal class-pose parametrisation

In Sections 3.3.2 and 3.3.3, we saw that, given a standard Borel MDP (π, ρ) satisfying Assumption 1, we obtain a partition in ergodic components $(\mathcal{X}^c)_{c\in\mathcal{C}}$ of the state-space $\mathcal{X}$ w.r.t. the Markov chain $\bar{\rho}$ defined by the update channel. These ergodic components will correspond to our generalisation of the "'class" coordinate. But crucially, Section 3.4 showed that we can design an ergodic decomposition $(\epsilon^c, \pi^c, \rho^c)_{c\in\mathcal{C}}$ such that the process distributions of the global MDP (π, ρ) can always be obtained as an average of the resp. process distribution of each $(\epsilon^c, \pi^c, \rho^c)$. This ergodic decomposition of the measurable MDP will be the basis for the design of our generalised "pose" coordinate, which we now present.

Similarly as in Section 3.1.2, we want to formalise the intuition that this "pose" coordinate keeps track of the "simplest collective description" of all elements in the family of MDPs $(\epsilon^c, \pi^c, \rho^c)_{c\in\mathcal{C}}$. The "collective" part of this intuition suggests to consider an MDP defined on the Cartesian product of the ergodic components. This will be formalised by the notion of *joining* of stationary MDPs (see Definition 3.5.6 below). On the other hand, the "simplest" part of the intuition will be formalised by an algebraic notion of minimality: i.e., we consider minimal elements w.r.t. a *factor* pre-order relation among joinings of a given family of stationary MDPs (see Definition 3.5.10 and Proposition 3.5.11). The resulting object is a *minimal joining* of a family of MDPs (see Definition 3.5.12), which we prove always exists for a finite family of finite alphabet MDPs (see Theorem 3.5.13). Of course, we are here interested in the family $(\epsilon^c, \pi^c, \rho^c)_{c\in\mathcal{C}}$ of an MDP's ergodic decomposition, but the concept of minimal joining applies more broadly.

Our proposal yields a conceptual shift on class-pose decomposition. Indeed, while previous work considers a *change of coordinates from the state-space $\mathcal{X}$ to the product space $\mathcal{C} \times \mathcal{P}$*, here we only consider a *parametrisation from the product space $\mathcal{C} \times \mathcal{P}$ to the state-space $\mathcal{X}$*, which we require to be "as isomorphic as possible". As the term "decomposition" thus seems ill-adapted to this change of perspective, we dub our new notion *minimal class-pose parametrisation* (see Definition 3.5.15). We then prove that, for a countable number of ergodic components, a measure-theoretic version of the class-pose decomposition notion from Definition 3.1.1 is encompassed as an edge case of our novel definition (see Theorem 3.5.17).

3.5.1 Factors and isomorphisms for stationary MDPs

In this section, we will focus on stationary MDPs, and the initial distribution has a central place in our treatment. However, let us stress that we aim to apply these results to the *ergodic components* of a measurable MDP (π, ρ), each equipped with their resp. ergodic distributions. Importantly, the latter distributions are uniquely defined by the "global" MDP (π, ρ), without reference to a specific initial distribution on it. In other words, the measure-dependent content of this section, will, in the following sections, yield measure-independent results on an MDP (π, ρ).

Our aim is to adapt to stationary MDPs the notions of factor, joining, j-factor and minimal joining defined for group actions in Section 3.1.2. Here, we start with factors, from which will follow the remaining notions.

The definition makes heavy use of the hook-up notation (see Definition 3.2.3), along with the push-forward and tensor product notations (see Definitions 3.2.5 and 3.2.6). In particular, let $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ measurable spaces. We recall that for a probability distribution $\mu \in \Delta_{\mathcal{A}_1}$ and a channel $\gamma_{1,2} \in \mathscr{K}(\mathcal{A}_1, \mathcal{A}_2)$, then $\mu\gamma_{1,2} \in \Delta_{\mathcal{A}_1 \times \mathcal{A}_2}$ denotes the whole *joint* distribution defined by μ and $\gamma_{1,2}$. This notation is used in points (ii) and (iii) below. We can also iterate it: if now we also have $\gamma_{12,3} \in \mathscr{K}(\mathcal{A}_1 \times \mathcal{A}_2, \mathcal{A}_3)$, then $(\mu\gamma_{1,2})\gamma_{12,3}$ is a joint distribution on $\mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{A}_3$. This notation is used in point (iii) below.

Definition 3.5.1. Let (μ_0, π, ρ) and (μ_0', π', ρ') be stationary MDPs, with resp. state-spaces $(\mathcal{X}, \mathfrak{X})$, $(\mathcal{X}', \mathfrak{X}')$ and resp. action spaces $(\mathcal{G}, \mathfrak{G})$, $(\mathcal{G}', \mathfrak{G}')$. We say that (μ_0', π', ρ') is an *MDP factor*, or just *factor* for short, of (μ_0, π, ρ) if there exist measurable maps $\phi : \mathcal{X} \to \mathcal{X}'$, $\psi : \mathcal{G} \to \mathcal{G}'$, such that:

(i) In $\Delta_{\mathcal{X}'}$, we have $\phi \cdot \mu_0 = \mu_0'$,

(ii) In $\Delta_{\mathcal{X} \times \mathcal{G}'}$, we have $\mu_0(\pi' \circ \phi) = \mu_0(\psi \circ \pi)$,

(iii) In $\Delta_{\mathcal{X} \times \mathcal{G} \times \mathcal{X}'}$, we have $(\mu_0\pi)(\rho' \circ (\phi \otimes \psi)) = (\mu_0\pi)(\phi \circ \rho)$.

The maps ϕ and ψ are then called the *factor maps* from (μ_0, π, ρ) to (μ_0', π', ρ'). Moreover, (μ_0', π', ρ') is said *isomorphic* to (μ_0, π, ρ) if it is a factor of (μ_0, π, ρ) such that ϕ, resp. ψ, is a measured isomorphism from $(\mathcal{X}, \mu_0)$ to $(\mathcal{X}', \mu_0')$, resp. from $(\mathcal{G}, \pi \cdot \mu_0)$ to $(\mathcal{G}', \pi' \cdot \mu_0')$. A family of stationary MDPs $(\mu_0^c, \pi^c, \rho^c)_{c \in \mathcal{C}}$ is called *pairwise isomorphic* if for all $c, c' \in \mathcal{C}$, the MDP (μ_0^c, π^c, ρ^c) is isomorphic to the MDP $(\mu_0^{c'}, \pi^{c'}, \rho^{c'})$.

For clarity, let us unpack points (i) to (iii). Point (i) means that the push-forward of μ_0 by ϕ is μ_0', i.e., that ϕ is a measured morphism from $(\mathcal{X}, \mu_0)$ to $(\mathcal{X}', \mu_0')$ (see Definition 3.2.1). In point (ii), we have $\mu_0 \in \Delta_{\mathcal{X}}$, while $\pi \in \mathscr{K}(\mathcal{X}, \mathcal{G})$ and $\pi' \in \mathscr{K}(\mathcal{X}', \mathcal{G}')$, so that we have both $\pi' \circ \phi \in \mathscr{K}(\mathcal{X}, \mathcal{G}')$ and $\psi \circ \pi \in \mathscr{K}(\mathcal{X}, \mathcal{G}')$. Thus $\mu_0(\pi' \circ \phi) \in \Delta_{\mathcal{X} \times \mathcal{G}'}$ and $\mu_0(\psi \circ \pi) \in \Delta_{\mathcal{X} \times \mathcal{G}'}$. In point (iii), we have $\mu_0\pi \in \Delta_{\mathcal{X} \times \mathcal{G}}$, while $\rho \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X})$, $\rho' \in \mathscr{K}(\mathcal{X}' \times \mathcal{G}', \mathcal{X}')$ and the tensor product $\phi \otimes \psi$ is a measurable map from $\mathcal{X} \times \mathcal{G}$ to $\mathcal{X}' \times \mathcal{G}'$. Thus we have both $\phi \circ \rho \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X}')$ and $\rho' \circ (\phi \otimes \psi) \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X}')$; and eventually, we have both $(\mu_0\pi)(\phi \circ \rho) \in \Delta_{\mathcal{X} \times \mathcal{G} \times \mathcal{X}'}$ and $(\mu_0\pi)(\rho' \circ (\phi \otimes \psi)) \in \Delta_{\mathcal{X} \times \mathcal{G} \times \mathcal{X}'}$.

Moreover, if $\mathcal{X}$ and $\mathcal{G}$ are standard Borel, then point (ii) in Definition 3.5.1 is equivalent to the equality $\pi' \circ \phi = \psi \circ \pi$ holding μ_0-a.e., and point (iii) is equivalent to the equality $\rho' \circ (\phi \otimes \psi) = \phi \circ \rho$ holdinng $\mu_0\pi$-a.e. (this is a consequence of point (*i*) in Lemma C.3.1). However, the latter equivalences are not clear beyond the standard Borel case, and we do need, below, to consider factors on potentially uncountable products of ergodic components — which are not standard Borel (see Proposition C.2.5). On the other hand, points (ii) and (iii) are the conditions that we will need for our proofs to work. This is the reason why we choose these slightly unusual "commutation" relations.

More conceptually, the equalities in points (ii) and (iii) can be seen as the "measured version" of the resp. commutation relations $\pi' \circ \phi = \psi \circ \pi$ and $\rho' \circ (\phi \otimes \psi) = \phi \circ \rho$. The idea, here, is that we only require the commutations to occur under inputs provided by the distribution μ_0 on $\mathcal{X}$, resp. by the distribution $\mu_0\pi$ on $\mathcal{X} \times \mathcal{G}$. This focus on channels' input measures will yield, below, a notion of joining which is inherently measure-dependent — just like the original definition from ergodic theory (de la Rue, 2006). The dependence on measures will then be crucial when it comes to reformulating joinings in the language of information theory.

Let us propose a graphical representation of these "measured commutations". We want to use commutative diagrams, but our situation here differs from that in Section 3.1.2 in two respects. First, we are dealing with stochastic channels instead of deterministic maps: it thus natural to have probability simplices as vertices, and push-forwards of channels as arrows.[34] But more fundamentally, commutative diagrams usually do not represent a dependency on input distributions. Here, we work around this by conserving a "copy" of the input distribution along the commutative diagram. More precisely, for a measurable space $\mathcal{A}$, if we define the "copy map"

$$\begin{aligned} \mathsf{copy} : \mathcal{A} &\to \mathcal{A} \times \mathcal{A} \\ a &\mapsto (a, a) \end{aligned}$$

then for any measurable space $\mathcal{B}$ and channel $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$, we have $\mu\gamma = (\mathrm{Id}_{\mathcal{A}} \otimes \gamma) \cdot (\mathsf{copy} \cdot \mu)$. I.e., informally, once $\mu \in \mathcal{A}$ has been copied, if we "process" the second coordinate of $\mathcal{A} \times \mathcal{A}$ with γ while leaving the first coordinate unchanged, we keep track of the whole joint distribution $\mu\gamma$ instead of just the output's distribution $\gamma \cdot \mu$. Using this fact, the following characterisations can easily be obtained: point (ii) in Definition 3.5.1 is equivalent to the commutation of the diagram

$$\begin{array}{ccc} \{\mathsf{copy} \cdot \mu_0\} & \xrightarrow{\mathrm{Id}_{\mathcal{X}} \otimes \pi} & \Delta_{\mathcal{X} \times \mathcal{G}} \\ {\scriptstyle \mathrm{Id}_{\mathcal{X}} \otimes \phi} \downarrow & & \downarrow {\scriptstyle \mathrm{Id}_{\mathcal{X}} \otimes \psi} \\ \Delta_{\mathcal{X} \times \mathcal{X}'} & \xrightarrow{\mathrm{Id}_{\mathcal{X}} \otimes \pi'} & \Delta_{\mathcal{X} \times \mathcal{G}'} \end{array} \tag{3.5.1}$$

and point (iii) to the commutation of the diagram

$$\begin{array}{ccc} \{\mathsf{copy} \cdot \mu_0 \pi\} & \xrightarrow{\mathrm{Id}_{\mathcal{X} \times \mathcal{G}} \otimes \rho} & \Delta_{\mathcal{X} \times \mathcal{G} \times \mathcal{X}} \\ {\scriptstyle \mathrm{Id}_{\mathcal{X} \times \mathcal{G}} \otimes \phi \otimes \psi} \downarrow & & \downarrow {\scriptstyle \mathrm{Id}_{\mathcal{X} \times \mathcal{G}} \otimes \phi} \\ \Delta_{\mathcal{X} \times \mathcal{G} \times \mathcal{X}' \times \mathcal{G}'} & \xrightarrow{\mathrm{Id}_{\mathcal{X} \times \mathcal{G}} \otimes \rho'} & \Delta_{\mathcal{X} \times \mathcal{G} \times \mathcal{X}'} \end{array} \tag{3.5.2}$$

[34] Note that a channel $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$ is entirely determined by its push-forward: indeed, $\gamma \cdot \delta_a = \gamma(\cdot|a)$ for all $a \in \mathcal{A}$.

where $\{\mathsf{copy} \cdot \mu_0\}$ and $\{\mathsf{copy} \cdot \mu_0\pi\}$ are seen as subsets of resp. $\Delta_{\mathcal{X}\times\mathcal{X}}$ and $\Delta_{\mathcal{X}\times\mathcal{G}\times\mathcal{X}\times\mathcal{G}}$; and to alleviate notations, arrows are labeled by channels γ themselves instead of their push-forwards γ_*. We leave to future work the exploration of alternative graphical representations of MDP factors — such as, possibly, categorical probabilities' *string diagrams* (Fritz, 2020).

As mentioned above, Definition 3.5.1 does not assume standard Borel spaces, and we will indeed need this level of generality. If the factor happens to be standard Borel, though, the MDP factor relation has a simple characterisation.

Proposition 3.5.2. *Let (μ_0, π, ρ) and (μ'_0, π', ρ') be stationary MDPs, with resp. state-spaces $\mathcal{X}$, $\mathcal{X}'$ and resp. action spaces $\mathcal{G}$, $\mathcal{G}'$. Let $\phi : \mathcal{X} \to \mathcal{X}'$ and $\psi : \mathcal{G} \to \mathcal{G}'$ be measurable maps, and consider the following statements:*

(i) (μ'_0, π', ρ') is a factor of (μ_0, π, ρ) with factor maps (ϕ, ψ).

(ii) $\phi \otimes \psi \otimes \phi$ is a measured morphism from $(\mathcal{X}\times\mathcal{G}\times\mathcal{X}, \mu_0\pi\rho)$ to $(\mathcal{X}'\times\mathcal{G}'\times\mathcal{X}', \mu'_0\pi'\rho')$.

Then $(i) \Rightarrow (ii)$, and if morever $\mathcal{X}'$, $\mathcal{G}'$ are standard Borel, then $(i) \Leftrightarrow (ii)$.

Proof. See Appendix C.6.1. □

We can also specialise Definition 3.5.1 to a notion of factor for stationary Markov chains by choosing an MDP with $|\mathcal{G}| = 1$, i.e., with only one action corresponding to updating time. Explicitly:

Definition 3.5.3. Let (μ_0, τ) and (μ'_0, τ') be stationary Markov chains, with resp. state-spaces $(\mathcal{X}, \mathfrak{X})$, $(\mathcal{X}', \mathfrak{X}')$. We say that (μ'_0, τ') is an *MDP factor*, or just *factor* for short, of (μ_0, τ) if there exists a measurable map $\phi : \mathcal{X} \to \mathcal{X}'$ such that (using the hook-up notation from Definition 3.2.3):

(i) $\phi \cdot \mu_0 = \mu'_0$,

(ii) $\mu_0(\tau' \circ \phi) = \mu_0(\phi \circ \tau)$.

The map ϕ is then called the *factor map* from (μ_0, τ) to (μ'_0, τ'). Moreover, (μ'_0, τ') is said *isomorphic* to (μ_0, τ) if it is a factor of (μ_0, τ) such that ϕ is a measured isomorphism from $(\mathcal{X}, \mu_0)$ to $(\mathcal{X}', \mu'_0)$.

Similarly as above, point (ii) in Definition 3.5.3 is characterised by the commutation of the following diagram:

$$\begin{array}{ccc} \{\mathsf{copy} \cdot \mu_0\} & \xrightarrow{\mathrm{Id}_{\mathcal{X}}\otimes\tau} & \Delta_{\mathcal{X}\times\mathcal{X}} \\ {\scriptstyle \mathrm{Id}_{\mathcal{X}}\otimes\phi}\downarrow & & \downarrow{\scriptstyle \mathrm{Id}_{\mathcal{X}}\otimes\phi} \\ \Delta_{\mathcal{X}\times\mathcal{X}'} & \xrightarrow{\mathrm{Id}_{\mathcal{X}}\otimes\tau'} & \Delta_{\mathcal{X}\times\mathcal{X}'} \end{array} \tag{3.5.3}$$

and if $\mathcal{X}$ is standard Borel, then point (ii) can be replaced by the requirement that $\tau' \circ \phi = \phi \circ \tau$ holds μ_0-a.e.. Moreover, if the stationary Markov chain is deterministic, we obtain a notion of factor for stationary dynamical systems (see Definition 3.2.7) which, at least in the standard Borel case, is the same as the one usually considered in ergodic theory (Einsiedler et al., 2011).

Our constructions below will crucially rely on the fact that the relations that we just defined have a specific structure:

Proposition 3.5.4. *The relation defined on stationary MDPs by MDP factors is a pre-order: ie., it is reflexive and transitive. The relation defined by MDP isomorphisms is an equivalence relation: i.e., it is reflexive, symmetric, and transitive. In particular, the factor and isomorphism relations on stationary Markov chains are, resp., a pre-order and an equivalence relation.*

Proof. See Appendix C.6.1. □

Related work The notion of *MDP homomorphism* (Ravindran et al., 2002; van der Pol et al., 2020) bears some similarities with our definition of factor for stationary MDPs. The main difference is that in MDP homomorphisms, the factor maps (ϕ, ψ) must be induced by symmetries of the factored MDP, while the only thing that we require to be preserved in our notion of factor is the stationary MDP structure: e.g., the trivial stationary MDP made of one state and one action is, in the sense of Definition 3.5.1, an MDP factor of any stationary MDP. Moreover, MDP homomorphisms do not depend on any fixed policy or state-space distribution, but depend on a reward function. Eventually, in MDP homomorphisms, the factor map corresponding to actions can also depend on states, i.e., using the notations defined above, we have $\psi : \mathcal{X} \times \mathcal{G} \to \mathcal{G}'$ instead of $\psi : \mathcal{G} \to \mathcal{G}'$. In such a case, the pair (ϕ, ψ) can be seen as a special case of *bundle morphism* (Husemoller, 1994), which suggests an interesting direction for generalisations of the framework developed here to settings where the state-action space is only *locally* decomposable as the product of a state-space $\mathcal{X}$ with an action space $\mathcal{G}$ (as, e.g., in (Oizumi et al., 2025) for the case of group actions). Note that in the case of both dynamical systems and group actions, the notion of factor has been extensively studied in ergodic theory (de la Rue, 2023; Glasner, 2003). Moreover, previously proposed dynamical notions of factors that might prove particularly relevant to future work on the framework presented in this chapter include:

- (Pfante et al., 2014), which compares different notions of coarse-graining for Markov processes, one of which — *observational commutativity* — is a measure-independent version of our Definition 3.5.3 of factor for stationary Markov chains,
- (Pfante et al., 2015), which develops an operator-theoretic point of view on factors in dynamical systems,
- (Barnett et al., 2021), whose notion of *dynamical independence* can be seen as a generalisation of stationary Markov chain factors to arbitrary stochastic processes, where the factor map is now defined over the whole discrete time-range $\mathbb{N}$ instead of a single time-step,
- (Rosas et al., 2024), which develops and studies the relation between several factor-like notions for stochastic processes in discrete time — called *causal closure* and *informational closure* — or for deterministic automata — called *computational closure*; and brings into focus whole *lattices* of coarse-grainings of a given stochastic process.

3.5.2 Joinings and minimal joinings for stationary MDPs

Here, we are actually interested in the concept of MDP factor only insofar as it allows us to study *MDP joinings*. While factors are generalisations of divisors in arithmetics, joinings are generalisation of common multiples of a family of integers (de la Rue, 2006, 2023).

Motivation: joinings of stationary dynamical systems and sensorimotor perception

For didactic purposes, here we start with the usual notion of joining for dynamical systems, which is a standard concept in ergodic theory (de la Rue, 2006, 2023). We choose a slightly unusual formulation that will better mirror the generalisation to MDPs.

Definition 3.5.5. Let $(\mu_0^c, \tau^c)_{c\in C}$ be a family of stationary dynamical systems (see Definition 3.2.7), on resp. standard Borel state-spaces $(\mathcal{X}^c)_{c\in C}$. Then a *joining* of $(\mu_0^c, \tau^c)_{c\in C}$is a stationary dynamical system (ν_0, ξ), with state-space $\mathcal{P}$, such that (μ_0^c, τ^c) is a factor of (ν_0, ξ) for all $c \in C$; i.e., explicitly, for all $c \in C$, there exists a measurable map $\phi^c : \mathcal{P} \to \mathcal{X}^c$ such that:

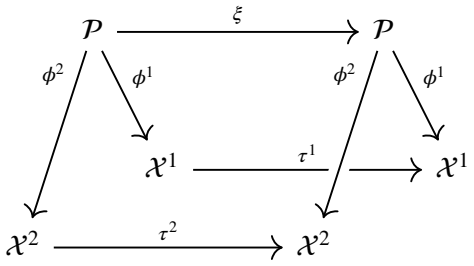
(i) $\phi^c \cdot \nu_0 = \mu_0^c$,

(ii) The equality $\tau^c \circ \phi^c = \phi^c \circ \xi$ holds ν_0-a.e..

A joining is called a *canonical joining* if its state-space $\mathcal{P}$ is the product space $\boldsymbol{\mathcal{X}}$ of the spaces $(\mathcal{X}^c)_{c\in C}$ (see Definition C.2.4), with, for all $c \in C$, the factor map ϕ^c given by the projection pr^c on the coordinate $\mathcal{X}^c$ of $\boldsymbol{\mathcal{X}}$.

Intuitively, a joining of (stationary) dynamical systems is a "common" (stationary) dynamical system that "contains" them both. Crucially, joinings thus allow the study of the *possible relationships* between dynamical systems whose relationship is *not defined a priori*. This object is interesting to the formalisation of sensorimotor theories of perception because it provides a concept of *abstraction arising from concrete sensorimotor interactions* — where, here, the concrete sensorimotor interactions correspond to the family of stationary dynamical systems $(\mu_0^c, \tau^c)_{c\in C}$, and the corresponding abstraction to their joining (ν_0, ξ).[35]

If $C = \{1, 2\}$ and if we disregard the requirement of ν_0-a.e. equality, the commutations in point (ii) can be summarised by the diagram



(3.5.4)

Let us now focus on the case of canonical joinings, still for $C = \{1, 2\}$. Then points (i) and (ii) can be reformulated as:

(i)' The marginal of $\nu_0 \in \Delta_{\mathcal{X}^1\times\mathcal{X}^2}$ on $\mathcal{X}^1$ is μ_0^1 and that on $\mathcal{X}^2$ is μ_0^2.

(ii)' The equalities $\tau^1 \circ \mathsf{pr}^1 = \mathsf{pr}^1 \circ \xi$ and $\tau^2 \circ \mathsf{pr}^2 = \mathsf{pr}^2 \circ \xi$ hold ν_0-a.e., which is equivalent[36] to $\xi = \tau^1 \otimes \tau^2$ holding ν_0-a.e.: i.e., up to a set of null probability, the transformation ξ of $\mathcal{X}^1 \times \mathcal{X}^2$ transforms the coordinate $\mathcal{X}^1$ with τ^1 and the coordinate $\mathcal{X}^2$ with τ^2.

In particular, given two stationary dynamical systems (μ_0^1, τ^1) and (μ_0^2, τ^2), a canonical joining is (on a set of ν_0-probability 1) uniquely determined by the joint distribution $\nu_0 \in \Delta_{\mathcal{X}^1\times\mathcal{X}^2}$ that is stationary under $\tau^1 \otimes \tau^2$ and whose marginals on $\mathcal{X}^1$ and $\mathcal{X}^2$ are resp. μ_0^1 and μ_0^2. Such a distribution is what is usually defined as "joining" in ergodic theory (see, e.g., Definition 1.3 in (de la Rue, 2006)).

[35] See Section 3.1.3 for a more detailed discussion of this sensorimotor intepretation (in the case of group actions, which is the same interpretation as for stationary dynamical systems).

[36] This can be easily verified, as ξ deterministic implies $\xi = (\xi^1, \xi^2)$ with $\xi^1 := \mathsf{pr}^1 \circ \xi$ and $\xi^2 := \mathsf{pr}^2 \circ \xi$.

Our alternative presentation has two motivations. Let us first motivate the explicit inclusion of the transformation ξ in the definition. This is because even for a canonical joining, point (ii)' above will not survive the generalisation to *stochastic* transformations $(\tau^c)_{c\in C}$: i.e., the joining's transformation $\xi : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ will not necessarily coincide (up to a null probability set) with the parallel processing $\bigotimes_{c\in C} \tau^c$ of each coordinate $\mathcal{X}^c$ by the corresponding τ^c (see Proposition 3.5.7 below). In particular, the joining's transformation ξ will not be uniquely defined (up to a null probability set) by the joining's distribution ν_0 and the transformations on coordinates $(\tau^c)_{c\in C}$, i.e., it will not be redundant to specify ξ.

Let us now motivate the choice of an unstructured state-space $\mathcal{P}$ for the joining, instead of the Cartesian product $\boldsymbol{\mathcal{X}}$ of the coordinates $\mathcal{X}^c$, as usually done in the theory of joinings (de la Rue, 2006, 2023). This is because we are actually interested in *minimal* joinings (defined below), where informally, the distribution ν_0 can be concentrated on a very "low-dimensional" region of the Cartesian product — similarly as in the polar coordinates example from Section 3.1.1, the *single* angular coordinate θ keeps track of the rotation of the *uncountable number* of circles of any radius. In such cases, "most" of the product space $\boldsymbol{\mathcal{X}}$, in which the joining's dynamics are embedded, will be included in a set of null probability, and thus be irrelevant. Moreover, this feature is important from a modeling persepective: indeed, this kind of "low-dimensional" joining (ν_0, ξ) will be interpreted as a *parsimonious abstraction* jointly describing the concrete sensorimotor dynamics of each system (μ_0^c, τ^c). It would thus be unnatural to define these "parsimonious" dynamics on an unnecessarily large state-space. Eventually, from the point of view of developing new methods of structure discovery, it is of course highly unproductive to work with unnecessarily large spaces. This is why we define joinings on a general state-space $\mathcal{P}$. However, from a formal perspective, it will sometimes be convenient to work in the product space $\boldsymbol{\mathcal{X}}$ — i.e., to work with what we call here canonical joinings.

We are now interested in studying joinings in settings that allow for distinct actions, instead of just a time update. One possibility is to consider (measure-theoretic) joinings of group actions: this is the topic of an already rich and still ongoing theory (Glasner, 2003).[37] Here, however, we want to have a notion of joining involving *non-invertible, stochastic and closed-loop actions*. This is the motivation for our MDP version of joinings below — which, to the best of our knowledge, has not been considered in previous literature.

Generalisation of joinings to stationary MDPs

Definition 3.5.6. A *joining* of a family of stationary MDPs $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ is a stationary MDP (ν_0, η, ξ),with state-space denoted by $\mathcal{P}$ and action-space denoted by $\mathcal{K}$, such that each (μ_0^c, π^c, ρ^c) if a factor of (ν_0, η, ξ). The corresponding factor maps $(\phi^c, \psi^c)_{c\in C}$ are the *marginalisation maps* of the joining. The set of all joinings of $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ is denoted by $\mathsf{Join}((\mu_0^c, \pi^c, \rho^c)_{c\in C})$. If the state-space $\mathcal{P}$ and the action space $\mathcal{K}$ are both standard Borel, then we call (ν_0, η, ξ) a *standard Borel joining*. A *canonical joining* is a joining such that $\mathcal{P} := \boldsymbol{\mathcal{X}}$ and $\mathcal{K} := \boldsymbol{\mathcal{G}}$, where $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{G}}$ are the product measurable spaces of resp. $(\mathcal{X}^c)_{c\in C}$ and $(\mathcal{G}^c)_{c\in C}$, and with, for all $c \in C$, the marginalisation maps ϕ^c, resp. ψ^c given by the projection on the coordinate $\mathcal{X}^c$ in $\boldsymbol{\mathcal{X}}$, resp. the projection on the coordinate $\mathcal{G}^c$ in $\boldsymbol{\mathcal{G}}$. Canonical joinings will often be denoted by $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$ rather than (ν_0, ξ, η).

As stationary dynamical systems (which we defined as always deterministic in Definition 3.2.7) can be seen as stationary MDPs with a trivial action space and deterministic transition channel (see Section 3.2.7), it can be easily verified that joinings of a family of standard Borel stationary MDPs as in Definition 3.5.6 generalise joinings of standard Borel stationary

[37] The notion of joining of group actions from (Glasner, 2003) can be seen as the measure-theoretic version of that in Section 3.1.2, which only considered a set-theoretic structure and was presented for didactic purposes.

dynamical systems as in Definition 3.5.5. Moreover, similarly as for dynamical systems, a joining of stationary MDPs is, intuitively, a "common" stationary MDP that "contains" all of them, and thus allows for studying the possible relationships between these MDPs. In particular, it is an interesting candidate for modeling the emergence of abstractions from concrete, closed-loop sensorimotor interactions.[38]

Importantly, while for deterministic MDPs, a canonical joining's policy and transition channel must essentially be the parallel processing defined by each coordinate's policy and transition channel, this is not true anymore in the stochastic case:

Proposition 3.5.7. *Let $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$ be a canonical joining of stationary standard Borel MDPs $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$. Define the tensor products (see Definition 3.2.6)*

$$\begin{aligned} \boldsymbol{\pi}^{\otimes} &:= \bigotimes_{c\in C} \pi^c \in \mathscr{K}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{G}}), \\ \boldsymbol{\rho}^{\otimes} &:= \bigotimes_{c\in C} \rho^c \in \mathscr{K}(\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{X}}). \end{aligned} \tag{3.5.5}$$

Then:

- *If π^c is deterministic for all $c \in C$, then $\boldsymbol{\pi} = \boldsymbol{\pi}^{\otimes}$ holds $\boldsymbol{\mu}_0$-a.e.; but for a general family $(\pi^c)_{c\in C}$, the latter equality might not hold.*
- *If ρ^c is deterministic for all $c \in C$, then $\boldsymbol{\rho} = \boldsymbol{\rho}^{\otimes}$ holds $\boldsymbol{\mu}_0\boldsymbol{\pi}$-a.e.; but for a general family $(\rho^c)_{c\in C}$, the latter equality might not hold.*

Proof. See Appendix C.6.2, where, in particular, we provide a concrete counter-example in the stochastic case. □

Intuitively, Proposition 3.5.7 is a consequence of the fact that for deterministic channels, knowing the value of each separate coordinate of the channel's output wholly determines it, while for stochastic channels, given the input, the marginal distribution of each output coordinate does not describe the correlations across output coordinates. Proposition 3.5.7 is important because it clarifies a crucial difference between joinings of deterministic transformations and joinings of stochastic transformations — where the latter have, to our knowledge, not been previously considered, be it for dynamical systems, group actions or MDPs. I.e., here, for stochastic policies $(\pi^c)_{c\in C}$, resp. stochastic transition channels $(\rho^c)_{c\in C}$, in the family of MDPs to be joined, the joining's policy η, resp. transition channel ξ, is *not* uniquely defined by the joining's state-space distribution ν and the "marginal" policies π^c, resp. "marginal" transition channels ρ^c: rather, η and ξ are part of the *choice* that defines a joining.

Minimal joinings of stationary MDPs

We now turn to the question: *how "tightly" can a family of MDPs be joined*? Can they all be seen as essentially the same MDP; are they, on the contrary, inherently "incompatible"; or is it something in between? We start by focusing on the edge cases corresponding to the first two options, formalised resp. with the notions of *isomorphism* and *disjointness* — which we adapt from similar notions for joinings of dynamical systems (de la Rue, 2006, 2023) or group actions (Glasner, 2003).

Definition 3.5.8. Let $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ be a family of stationary MDPs. The *product joining* is the canonical joining $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho}) := (\boldsymbol{\mu}_0^{\otimes}, \boldsymbol{\pi}^{\otimes}, \boldsymbol{\rho}^{\otimes})$, where $\boldsymbol{\pi}^{\otimes}$ and $\boldsymbol{\rho}^{\otimes}$ are defined in (3.5.5)

[38] See Section 3.1.3 for a more detailed discussion of this sensorimotor intepretation (in the case of group actions, which is the same interpretation as for stationary MDPs).

and[39]

$$\boldsymbol{\mu}_0^{\otimes} := \bigotimes_{c\in C} \mu_0^c \in \Delta_{\mathcal{X}}.$$

On the other hand, an *isomorphic joining* is a joining (ν_0, ξ, η) which is MDP isomorphic to (μ_0^c, π^c, ρ^c) for all $c \in C$. The family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ is called *disjoint* if, up to measured MDP isomorphism, its only joining is the product joining; and it is called *isomorphic* if it has an isomorphic joining.

At one extreme, the product joining is the one that "makes all state-space coordinates independent and runs the MDPs in parallel" (with probability one). In this sense, it does not capture any potentially common feature of the dynamics among these MDPs. Thus, disjoint families of MDPs are those whose respective dynamics "do not share anything in common" — the MDP version of relatively prime numbers. Let us mention though, for completeness, that even in the case of deterministic dynamics, the analogy with arithmetics can be subtle (de la Rue, 2006, 2023). Indeed, on the one hand, two integers are relatively prime, equivalently, if their only common factor is one, or if their least common multiple is their product. But it turns out that two stationary dynamical systems with no non-trivial common factor might still be non-disjoint (de la Rue, 2023; Rudolph, 1979).

At the other extreme, an isomorphic joining "makes state-space coordinates depend bijectively on one another and runs the MDPs in sync" (with probability one) by making each MDP in the family isomorphic to the same MDP (ν_0, η, ξ). This can also be seen as the fact that in an isomorphic family, any element (μ_0^c, π^c, ρ^c) can be used to "simulate" any of the other ones. More precisely:

Proposition 3.5.9. *The following holds:*

(i) A family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ of stationary MDPs is isomorphic if and only if it is pairwise isomorphic, i.e., if for all $c, c' \in C$, we have a stationary MDP isomorphism between (μ_0^c, π^c, ρ^c) and $(\mu_0^{c'}, \pi^{c'}, \rho^{c'})$.

(ii) In an isomorphic family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ of stationary MDPs, any element (μ_0^c, π^c, ρ^c) provides an isomorphic joining of the whole family; in particular, the isomorphic joining can always be chosen standard Borel if at least one MDP (μ_0^c, π^c, ρ^c) is standard Borel.

Proof. This is straightforward if we use the fact that stationary MDP isomorphism is an equivalence relation (see Proposition 3.5.4). □

Here, however, we want to design a formalism that can "optimally join" *arbitrary* families of MDP, beyond the edge cases of isomorphic or disjoint families. Does there exist joinings that capture as much as possible the common aspects across the dynamics of the family's diversity of MDPs, thus "simultaneously simulating" them into a common MDP that avoids unnecessary complexity by, as much as possible, "using the same dynamics to run different MDPs"? This intuition can be formalised in two different directions. Our IB-oriented approach from the previous chapter suggests to design a relevant notion of "parsimony" that would be formulated in information-theoretic terms. Alternatively, we can rather formalise the intuition of making the joining "as isomorphic as possible", through a well chosen preorder, similarly as in Definition 3.1.4. Here, we start with the latter, algebraic notion. The information-theoretic point of view will be developed — for the finite case — in Section 3.6.2, where we will show that these two languages turn out to be equivalent.

[39] See Definition 3.2.6. The stationarity of $\boldsymbol{\mu}_0^{\otimes}$ w.r.t. $\overline{\boldsymbol{\rho}^{\otimes}}$ comes from that of μ^c w.r.t. $\overline{\rho^c}$ for each $c \in C$.

Similarly as in Section 3.1.2, the aim of the pre-order will here be to compare different joinings of a given family of stationary MDPs. While the notion of MDP factor from Definition 3.5.1 does provide a pre-order, this relation does not take into account the marginalisation maps of the respective joinings. Doing so leads to the following relation:

Definition 3.5.10. Let (ν_0, η, ξ) and (ν_0', η', ξ') be two joinings of a family of stationary MDPs $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$, with resp. marginalisation maps $(\phi^c, \psi^c)_{c\in C}$ and $((\phi')^c, (\psi')^c)_{c\in C}$. We say that (ν_0', η', ξ') is an *MDP joining factor* — or *j-factor* for short — of (ν_0, η, ξ) if (ν_0', η', ξ') is an MDP factor of (ν_0, η, ξ) with factor maps (Φ, Ψ) such that, using the hook-up notation, we have for all $c \in C$:

$$(\nu_0\eta)\Big(\Big((\phi')^c \otimes (\psi')^c\Big)\circ\Big(\Phi \otimes \Psi\Big)\Big) = (\nu_0\eta)(\phi^c \otimes \psi^c). \tag{3.5.6}$$

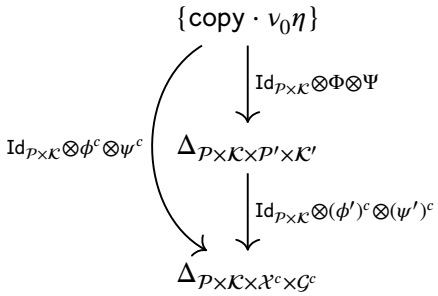
We then write

$$(\nu_0', \eta', \xi') \lesssim (\nu_0, \eta, \xi).$$

In short, (ν_0', η', ξ') is a j-factor of (ν_0, η, ξ) if it is an MDP factor of (ν_0, η, ξ) such that, with probability one, the corresponding factor maps (Φ, Ψ) resp. transform (by composition at the input) the marginalisation maps $(\phi^c, \psi^c)_{c\in C}$ of the latter joining into the marginalisation maps $((\phi')^c, (\psi')^c)_{c\in C}$ of the former joining. Similarly as for the MDP factor relation, equation (3.5.6) can be seen as the "measured version" of the commutation relations

$$(\phi')^c \circ \Phi = \phi^c,$$
$$(\psi')^c \circ \Psi = \psi^c,$$

and it is equivalent to the commutativity of the diagram



(3.5.7)

where $\{\mathsf{copy} \cdot \nu_0\eta\}$ is seen as a subset of $\Delta_{\mathcal{P}\times\mathcal{K}\times\mathcal{P}\times\mathcal{K}}$, and to alleviate notations, the arrows are labeled by the channels instead of their push-forward. Importantly, this new relation between joinings is still a pre-order:

Proposition 3.5.11. *The relation defined on the joinings of a given family of stationary MDPs $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ by j-factors is a pre-order: i.e., it is reflexive and transitive.*

Proof. See Appendix C.6.2. □

We are eventually ready to introduce the **central concept of this chapter**:

Definition 3.5.12. A *minimal joining* of a given family of stationary MDPs $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ is a joining which is minimal for the j-factor pre-order $\lesssim$, i.e., which is a j-factor of any other joining.

Minimal joinings can be seen as the stationary MDP version of integers' least common multiples. Surprisingly, we could not find, despite extensive search in the literature, a notion

of minimal joining in the sense of Definition 3.5.12, even for deterministic dynamics.[40] Thus, to the best of our knowledge, this is the first time this notion is introduced. However, we are not experts in the theory of joinings, which has been developed in a vast and specialised literature — see (de la Rue, 2023) for a recent review and (Glasner, 2003) for a monograph.

Note that if a family of MDPs is isomorphic, then minimal joinings coincide with the isomorphic joinings (this is a direct consequence of the fact that the factor relation is a pre-order and the isomorphism one an equivalence relation). Definition 3.5.12 thus provides a generalisation of the highly non-generic case of isomorphic joinings. But how generic are minimal joinings themselves? Does there always exist minimal joinings of a family of (stationary, standard Borel) MDPs? Below, we present the main result of this section, which states that the answer is yes in the finite case. We leave to future work the investigation of the existence of minimal joinings in the non-finite case.

Theorem 3.5.13. *For a finite family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ of finite-alphabet, stationary MDPs, there always exists at least one minimal joining.*

Proof. See Appendix C.6.2. □

This theorem should be contrasted with the fact that isomorphic joinings of a given family do *not* always exists — otherwise, e.g., all stationary dynamical systems would be isomorphic, which is of course not the case, even in the finite case.

As mentioned in the above discussion of joinings of dynamical systems, while in general we want to be able to work with joinings defined on an arbitrary state-space $\mathcal{P}$, focusing on canonical joinings will sometimes deliver interesting insights or be helpful for proofs. For that purpose, it is important to understand how general (minimal) joinings relate to canonical ones:

Proposition 3.5.14. *Let $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ be a countable family of standard Borel stationary MDPs. Then for any joining (ν_0, η, ξ) of $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$, there exists a canonical joining of $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ which is a j-factor of (ν_0, η, ξ). In particular, the family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ has a minimal joining if and only if it has a canonical minimal joining.*

Proof. See Appendix C.6.2. □

The assumption of C countable in Proposition 3.5.14 is an important limitation — as we are here interested in joinings of MDP ergodic components, which are usually in uncountable number for continuous spaces: see, e.g., the polar coordinates examples from Section 3.1.1, where from Theorem 3.4.6, the orbits coincide with ergodic components. We leave to future work its generalisation beyond this countability assumption.

Technical remark. Let us recall that if C is uncountable (and if all MDPs in the family are non-trivial), then the spaces $\mathcal{X}$ and $\mathcal{G}$ are *uncountable products* of standard Borel spaces, which are never standard Borel (see Proposition C.2.5). This suggests that even if a minimal joining exists, there might not exist one living on standard Borel spaces $\mathcal{P}$ and $\mathcal{K}$. This is a problem as, for instance, the ergodic theory results quoted in Section 3.3.2 require a standard Borel state-space — which means that they might be unusable to do ergodic theory on the minimal joining's spaces $\mathcal{P}$ and $\mathcal{K}$. Moreover, our long-term aim is to develop an information-theoretic treatment of minimal joinings in continuous spaces. While up to the level of generality of standard Borel spaces, information-theoretic tools are quite mature and based on traditional measure theory (Gray, 2011), to the best of our knowledge, more general frameworks are still under development. In particular, some ongoing category-theoretic approaches to information (Perrone, 2024) are closely connected to efforts at extending measure and ergodic theory to

[40] There exists a notion of *minimal self-joining* (de la Rue, 2023), which is not directly related to Definition 3.5.12.

address problems arising from uncountability (Fritz et al., 2025; Jamneshan et al., 2023; Moss et al., 2023), which is known to touch upon deep limitations of traditional measure theory (König, 2012).

Related work The notion of isomorphic joining bears some similarities with the "stitching" of object manifolds developed in (Keurti et al., 2024). The latter work is set in the framework of group actions, where once the group actions describing the changes of pose in space of distinct rigid objects have been learned, the problem is to find linear transformations that allow one to "switch", in an equivariant way, from one group action to the other. In our understanding, the mathematical structure underlying this procedure is very close to an isomorphic joining of the family of group actions defined by each object, where these group actions are seen as transition channels of MDPs as in Section 3.4.2. Moreover, our notion of minimal joining resonates with the *Differential Heterogenesis* framework (Sarti et al., 2022), to the extent that the latter studies the emergence of novel dynamics on a "lifted" state-space through the combination of distinct dynamics on distinct state-spaces — although with different tools, rooted in sub-Riemannan geometry.

3.5.3 Minimal class-pose parametrisation

We are eventually ready to present our measure-theoretic generalisation of minimal class-pose parametrisation.

Definition 3.5.15. Let (π, ρ) a standard Borel MDP with state-space $\mathcal{X}$ and action space $\mathcal{G}$, satisfying Assumption 1. A *minimal class-pose parametrisation of* $\mathcal{X}$ *w.r.t.* (π, ρ), is a stationary MDP (ν_0, η, ξ) with state-space $\mathcal{P}$ and action space $\mathcal{K}$, together with measurable maps $\phi : C \times \mathcal{P} \to \mathcal{X}$, $\psi : C \times \mathcal{K} \to \mathcal{G}$, with C a measurable space, such that:

- (ν_0, η, ξ) is a minimal joining of a decomposition into ergodic components $(\epsilon^c, \pi^c, \rho^c)_{c\in C}$ of the MDP (π, ρ),
- Writing $\phi^c := \phi(c, \cdot)$ and $\psi^c := \psi(c, \cdot)$, the family $(\phi^c, \psi^c)_{c\in C}$ provides the minimal joining's marginalisation maps.

A minimal class-pose parametrisation is called *standard Borel* if the spaces C, $\mathcal{P}$ and $\mathcal{K}$ are all standard Borel, and it is called an *isomorphic class-pose parametrisation* if the minimal joining is isomorphic.

Note that there is here a subtle interplay between notions that depend on a measure and notions that do not. On the one hand, we do *not* fix any arbitrary initial distribution on $\mathcal{X}$: we just start with the policy π and the transition channel ρ. However, the family of ergodic probabilities $(\epsilon^c)_{c\in C}$, which is uniquely defined by (π, ρ), is instrumental in defining the minimal joining (ν_0, η, ξ) (see Section 3.5.2).

From Theorem 3.4.1, a measurable MDP (π, ρ) always has (under Assumption 1) a decomposition into ergodic components, and from Theorem 3.5.13, a minimal joining always exists in the finite case. Thus finite families of finite alphabet MDPs always have (under Assumption 1) a minimal class-pose parametrisation. However, at this stage, it is not clear whether arbitrary standard Borel MDPs (π, ρ) have a minimal class-pose parametrisation. One problem is the existence of minimal joinings, which we did not prove in the non-finite case (see Section 3.5.2). Another problem is the measurability of the maps ϕ and ψ:

Technical remark. If the push-forward $\bar{\rho}_* : \mathcal{X}_{\mathsf{BL}} \to \mathcal{X}_{\mathsf{BL}}$ is not continuous and C is uncountable, it is at this stage not even clear which σ-algebra the space C should be equipped with (see Section 3.3.4). Moreover, even if we did, we might run into new difficulties when C is uncountable. If for instance we consider canonical joinings, i.e., joinings defined on the product spaces $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{G}}$ of resp. $(\mathcal{X}^c)_{c\in C}$ and $(\mathcal{G}^c)_{c\in C}$ (see Definition 3.5.6), then maps ϕ and ψ

that are measurables w.r.t. the resp. product σ-algebras must depend on a countable number of coordinates (see Exercise 2.4.1 in (Tao, 2011)), i.e., on a countable set $\tilde{C} \subseteq C$ of ergodic components. In the latter situation, if C is uncountable, this means that the minimal joining of ergodic components does not depend on an uncountable set $C \setminus \tilde{C}$ of ergodic components, which seems problematic. We leave to future work the investigation of these subtleties which, again, might stumble upon deep limitations of traditional measure theory (Jamneshan et al., 2023; König, 2012).

Similarly as in the introduction, Definition 3.5.15 proposes a shift of perspective w.r.t. the class-pose decomposition literature: in a nutshell, we propose to "reverse the direction of the vertical arrows in the commutative diagrams". Indeed, previous work on class-pose decomposition seeks to exhibit a *change of coordinate from the state-space $\mathcal{X}$ to a product space $C \times \mathcal{P}$* (see Definition 3.1.1). On the contrary, here, we are looking at a *parametrisation of the state-space $\mathcal{X}$ by the product space $C \times \mathcal{P}$*. I.e., we have a map $\phi : C \times \mathcal{P} \to \mathcal{X}$ which is "measure-theoretically surjective", in the sense that $(\mathcal{X}^c)_{c \in C}$ is a partition of $\mathcal{X}$ and for all $c \in C$, the restriction ϕ^c is a measured morphism from $(\mathcal{P}, \nu)$ to $(\mathcal{X}^c, \epsilon^c)$. Crucially, this "measure-theoretic surjection" might *not* be "measure-theoretically injective": i.e., each ϕ^c might not be a measured isomorphism. Moreover, as we are considering potentially stochastic and closed-loop actions, it becomes important to include an action space $\mathcal{K}$ in the parametrisation, with a corresponding map $\psi : C \times \mathcal{K} \to \mathcal{G}$ that satisfies similar properties as ϕ (i.e., for each $c \in C$ the map ψ^c is a measured morphism but might not be a measured isomorphism). However, we require this parametrisation to be "as isomorphic as possible", in the sense that the parametrisation is a factor of any similar parametrisation (see the Definition 3.5.12 of minimal joining). Isomorphic class-pose parametrisations correspond to the edge case when this minimal parametrisation happens to be "fully isomorphic"— in the measure-theoretic sense that for all $c \in C$, the marginalisation maps ϕ^c and ψ^c are both stationary MDP isomorphisms (see Definition 3.5.1).

Before turning to an information-theoretic reformulation of these new notions of class and pose, we show in the next section that the previously considered notion of class-pose decomposition can be captured as a special case of our framework.

3.5.4 Application to group-theoretic class-pose decomposition

Here, we show that, when there is a countable number of orbits, our framework of *minimal* class-pose *parametrisation* is a generalisation of (a measure-theoretic version of) class-pose *decomposition*.

We settle ourselves again in the framework of Section 3.4.2: i.e., we fix a standard Borel space $\mathcal{G}$ with a measurable group structure, and assume that $\mathcal{G}$ has a group-stationary probability υ (see Definitions 3.4.3 and 3.4.4) — which is necessarily unique (see point (vii) in Theorem 3.4.6). Then, for any measurable action ρ of $\mathcal{G}$ on some standard Borel space $\mathcal{X}$, with corresponding independent policy $\pi_\upsilon \in \mathscr{K}(\mathcal{X}, \mathcal{G})$, Theorem 3.4.6 applies, and we denote by:

- $(\mathcal{X}^c)_{c \in C}$ the partition in orbits (which coincide here with the ergodic components),
- $(\rho^c)_{c \in C}$ the restrictions of ρ to each orbit $\mathcal{X}^c$,
- $(\pi_\upsilon^c)_{c \in C}$ the restrictions of the independent policy to each orbit $\mathcal{X}^c$,
- $(\epsilon^c)_{c \in C}$ the corresponding family of ergodic measures.

From points (iv) and (v) in Theorem 3.4.6, each ϵ^c is $\bar{\rho}$-stationary and also ρ_g-stationary for all $g \in \mathcal{G}$.

We consider the following measure-theoretic version of class-pose decomposition (compare with the set-theoretic version in Definition 3.1.1):

Definition 3.5.16. A *(measured) class-pose decomposition* of $\mathcal{X}$ w.r.t. ρ is a tuple (κ, θ, ξ), with $\kappa : \mathcal{X} \to C$, $\theta : \mathcal{X} \to \mathcal{P}$ measurable and a measurable group action $\xi : \mathcal{P} \times \mathcal{G} \to \mathcal{P}$, such that:

(*i*) κ is the projection on orbits,

(*ii*) There exists $\tilde{\mathcal{G}} \subseteq \mathcal{G}$ with $\upsilon(\tilde{\mathcal{G}}) = 1$ such that for all $c \in C$ and all $g \in \tilde{\mathcal{G}}$, the restriction θ^c of θ to $\mathcal{X}^c$ is a stationary Markov chain isomorphism from (ϵ^c, ρ^c_g) to (ν_0, ξ_g), for some stationary $\nu_0 \in \Delta_{\mathcal{P}}$.

In short, Definition (3.5.16) is a varation of our previous set-theoretic Definition (3.1.1) that adds measurability requirements, and allows for the commutation relations to only hold on a full-probability subset of group elements (where the probability is the group's unique stationary probability). Note that in point (*ii*), the full probability set $\tilde{\mathcal{G}}$ does not depend on the orbit $c \in C$. This notion of class-pose decomposition relates to isomorphic class-pose parametrisations in the following way:

Theorem 3.5.17. *Assume that there is a countable number of orbits. Then the following are equivalent:*

(i) The group action has a measured class-pose decomposition (κ, θ, ξ) with state-space $\mathcal{P}$.

(ii) The MDP (π_υ, ρ) defined by the group action ρ has an isomorphic class-pose parametrisation whose isomorphic minimal joining is of the form $(\nu_0, \eta_\upsilon, \xi)$, with state-space $\mathcal{P}$ and action space $\mathcal{K} := \mathcal{G}$, where $\eta_\upsilon \in \mathscr{K}(\mathcal{P}, \mathcal{G})$ is the independent policy, and with marginalisation maps (ϕ, ψ) such that for all $c \in C$, the map $\psi^c : \mathcal{K} = \mathcal{G} \to \mathcal{G}$ is trivial: i.e., it is the identity map $\mathtt{Id}_{\mathcal{G}}$.

Moreover, if any of the above holds, then the class-pose decomposition from (i) and the isometric class-pose parametrisation from (ii) can be chosen such that:

- *$\mathcal{P}$ is the same space in points (i) and (ii),*
- *ξ is the same transition channel in points (i) and (ii),*
- *For all $c \in C$, the restriction θ^c of θ of point (i) — which is a measured isomorphism from $(\mathcal{X}^c, \epsilon^c)$ to $(\mathcal{P}, \nu_0)$ — and the restriction ϕ^c of ϕ of point (ii) — which is a measured isomorphism from $(\mathcal{P}, \nu_0)$ to $(\mathcal{X}^c, \epsilon^c)$ — are mod 0 inverses of each other.*

Proof. See Appendix C.6.3. □

In short, for group actions satisfying the assumptions above: an isomorphic class-pose parametrisation is obtained by "measure-theoretically inverting", on each class $c \in C$, the restriction θ^c of a measured class-pose decomposition's "pose" channel θ; and conversely, a measured class-pose decomposition is obtained from an isomorphic class-pose parametrisation with trivial action space marginalisation maps $(\psi^c)_{c \in C} = (\mathtt{Id}_{\mathcal{G}})_{c \in C}$ by "measure-theoretically inverting", on each class $c \in C$, the restriction ϕ^c of the projection ϕ.

Note that our assumption of a countable number of orbits is a major limitation of the result, as it is not expeceted to hold in most interesting scenarios for the continuous case (e.g., even for the simple example of polar coordinates, there is an uncountable number of orbits). We leave to future work potential generalisations of Theorem 3.5.17 that would drop this countability assumption.

However, the result stills roots our novel framework in the pre-existing literature on class-pose decomposition, and shows explicitly in which sense the former is a generalisation of the latter.

3.6 Information-theoretic characterisation and softening

We now turn to information-theoretic reformulations — in the finite case — of both the decomposition into ergodic components and minimal joinings. In Section 3.6.1, we leverage the fact that the projection on ergodic components can be seen as a "mean-asymptotic minimal sufficient statistic" (see Section 3.3.5) to characterise it as the solution, for maximal trade-off parameter, to an instance of the DIB problem from Chapter 2. The problem consists in compressing the state-space $\mathcal{X}$ in a way that preserves the mutual information between the initial state and the time-average of the resulting trajectory. Then, in Section 3.6.2, we show that minimal joinings coincide with *minimum entropy joinings* — or equivalently, *maximum multi-information joinings*. While at this stage, we do not information-theoretically characterise the property of being a joining itself — and thus do not obtain a full characterisaion of minimal joinings as solutions to a generalised IB problem — this results exhibits an equivalence between an algebraic and an information-theoretic notion of minimality. It also lays the groundwork for a fuller information-theoretic reformulation of minimal joinings — and thus, once combined with the previous reformulation of ergodic components, for one of minimal class-pose parametrisation.

3.6.1 Ergodic components as mean-asymptotic information-preserving compression

Our aim is here to derive an IB-like problem such that the projection on ergodic components would be the essentially unique solution for maximal trade-off parameter. However, defining the quantity that will be "preserved" by the compression requires first setting up some background.

Framework

We assume that $\mathcal{X}$ is finite, and we fix a Markov chain $\tau \in \mathscr{K}(\mathcal{X})$. As in Section 3.4, we adopt Assumption 1, i.e., that $\mathcal{X} = \mathcal{X}_{\mathrm{erg}} = \mathcal{X}_{\mathrm{erg,inv}}$ (where $\mathcal{X}_{\mathrm{erg}}$ and $\mathcal{X}_{\mathrm{erg,inv}}$ are defined, resp., in equations (3.3.1) and (3.3.3)). For readers familiar with countable Markov chain theory: this assumption holds, e.g., if all points are positive recurrent, i.e., if there are no transient points (see Remark 3.3.3).

Here, the decomposition into ergodic components $(\mathcal{X}^c)_{c\in C}$ is thus a partition of the whole state-space $\mathcal{X}$ (see Proposiition 3.3.9). We denote by $\mathsf{pr} : \mathcal{X} \to C$ the corresponding projection on ergodic components, and by $(\epsilon^c)_{c\in C}$ the corresponding family of ergodic measures (see point (iv) in Theorem 3.3.4).

Let us now fix a full-support, stationary distribution $\mu_0 \in \Delta_{\mathcal{X}}$. As before, we denote the process distribution of the Markov chain (μ_0, τ) by

$$q := q(\overrightarrow{X}) \in \Delta_{\mathcal{X}^{\mathbb{N}}}.$$

However, now, we also consider a "bottleneck" countable space $\mathcal{T}$ and for each "bottleneck" channel $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$, we denote by

$$q_\kappa := q_\kappa(\overrightarrow{X,T}) \in \Delta_{(\mathcal{X}\times\mathcal{T})^{\mathbb{N}}}$$

the joint process distribution defined through the Bayesian network

$$\begin{array}{ccccccc} X_0 & \xrightarrow{\tau} & X_1 & \xrightarrow{\tau} & X_2 & \cdots\cdots & \dots \\ \downarrow{\scriptstyle \kappa} & & \downarrow{\scriptstyle \kappa} & & \downarrow{\scriptstyle \kappa} & & \dots \\ T_0 & & T_1 & & T_2 & & \dots \end{array} \tag{3.6.1}$$

I.e., explicitly, $q_\kappa \in \Delta_{(\mathcal{X}\times\mathcal{T})^{\mathbb{N}}}$ is the unique process distribution such that for all $n \in \mathbb{N}$, using the hook-up and tensor product notations (see Definitions 3.2.3 and 3.2.6),

$$q_\kappa(X_0, \dots, X_n, T_0, \dots, T_n) := q(X_0, \dots, X_n) \bigotimes_{i=0}^{n} \kappa, \tag{3.6.2}$$

where the existence and uniqueness are ensured by the Kolmogorov extension theorem (see Theorem C.2.10). Let us now define, for all $n \in \mathbb{N}$, the n-th Césaro mean of the joint distribution between the initial point and an iterated point:

$$\bar{q}^n(X, X') := \frac{1}{n} \sum_{i=0}^{n-1} q(X_0, X_i) \in \Delta_{\mathcal{X}\times\mathcal{X}}.$$

From Corollary 3.3.13, these Césaro means converge (symbol-wise) to a distribution $\bar{q} = \bar{q}(X, X') \in \Delta_{\mathcal{X}\times\mathcal{X}}$, and, as in Section 3.3.5,

$$\begin{aligned} \bar{q}(x, x') &:= \sum_{c\in C} (\mathsf{pr} \cdot \mu_0)(c) \epsilon^c(x) \epsilon^c(x') \\ &= \sum_{c\in C} \mu_0(\mathcal{X}^c) \epsilon^c(x) \epsilon^c(x') \qquad (3.6.3) \\ &= \sum_{c\in C} \mu_0(\mathcal{X}^c) \epsilon^c(x) \epsilon^c(x') \delta_{x,x'\in\mathcal{X}^c}. \qquad (3.6.4) \end{aligned}$$

Note that, from the definition of $\bar{q}(X, X')$ as the symbol-wise limit of the $\bar{q}^n$, we have $\bar{q}(X') = \bar{q}(X) = \bar{q}(X_0)$, with these distributions being full-support, while $\bar{q}(X'|X)$ is the symbol-wise limit of $\frac{1}{n}\sum_{i=0}^{n-1} q(X_i|X_0)$: in this sense, $\bar{q}(X, X')$ is the joint distribution between the initial point and the time-average of the resulting trajectory. Line (3.6.3) then means, informally, that this mean-asymptotic distribution $\bar{q}$ happens to be the average of the independent product of ergodic distributions $(\epsilon^c \otimes \epsilon^c)_{c\in C}$, with resp. weight given by the probability of the ergodic component under μ. Line (3.6.4) emphasizes that each ergodic distribution ϵ^c is concentrated on the ergodic component $\mathcal{X}^c$ (see point (i) in Theorem 3.3.7). These equations also imply that the projection on ergodic components pr defines a minimal sufficient statistic of X w.r.t. X', and of X' w.r.t. X (see Theorem 3.3.18).

Let us construct an asymptotic joint distribution on $\mathcal{X} \times \mathcal{X} \times \mathcal{T} \times \mathcal{T}$ that extends the asymptotic joint distribution $\bar{q}(X, X')$ to the bottleneck space. From the definition (3.6.2) of q_κ, we have

$$q_\kappa(X_0, X_n, T_0, T_n) = q(X_0, X_n)(\kappa \otimes \kappa),$$

so that, by linearity of the hook-up operation,

$$\begin{aligned}\bar{q}^n_\kappa(X, X', T, T') &:= \frac{1}{n}\sum_{i=0}^{n-1} q_\kappa(X_0, X_i, T_0, T_i)\\ &= \left(\frac{1}{n}\sum_{i=0}^{n-1} q(X_0, X_i)\right)(\kappa \otimes \kappa)\\ &= \bar{q}^n(X, X')(\kappa \otimes \kappa).\end{aligned}$$

Moreover, in our current countable setting, it is straightforward to verify that the map

$$\begin{aligned}\Delta_{\mathcal{X}\times\mathcal{X}} &\to \Delta_{\mathcal{X}\times\mathcal{X}\times\mathcal{T}\times\mathcal{T}}\\ q(X, X') &\mapsto q(X, X')(\kappa \otimes \kappa),\end{aligned}$$

that hooks-up a joint distribution $q(X, X')$ with the channel $\kappa \otimes \kappa$, is continuous. Thus the (symbol-wise) convergence in of $\bar{q}^n(X, X')$ to $\bar{q}(X, X')$ implies the (symbol-wise) convergence of the sequence

$$\left(\bar{q}^n(X, X', T, T')\right)_{n\in\mathbb{N}} = \left(\bar{q}^n(X, X')(\kappa \otimes \kappa)\right)_{n\in\mathbb{N}},$$

and

$$\begin{aligned}\bar{q}_\kappa(X, X', T, T') &:= \lim_{n\to\infty} \bar{q}^n_\kappa(X, X', T, T')\\ &= \bar{q}(X, X')(\kappa \otimes \kappa)\end{aligned} \tag{3.6.5}$$

Note that the second equality in equation (3.6.5) implies that, for all $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$, the distribution $\bar{q}_\kappa(X, X', T, T')$ satisfies the Markov chain

$$T - X - X' - T'. \tag{3.6.6}$$

Let us denote by $\bar{I}^n(X; X')$, $\bar{I}(X; X')$, $\bar{I}(X; X')$, $\bar{I}^n_\kappa(T; T')$, and $\bar{I}_\kappa(T; T')$ the mutual informations defined, resp. by the joint distribution $\bar{q}^n(X, X')$, $\bar{q}(X, X')$, $\bar{q}^n_\kappa(T, T')$, and $\bar{q}_\kappa(T, T')$. The Markov chain (3.6.6) implies, using the data-processing inequality (Cover et al., 2009), that

$$\bar{I}_\kappa(T; T') \leq \bar{I}(X; X'). \tag{3.6.7}$$

Moreover, as the mutual information is continuous w.r.t the underlying joint distribution,

$$\begin{aligned}\lim_{n\to\infty} \bar{I}^n(X; X') &= \bar{I}(X; X'),\\ \lim_{n\to\infty} \bar{I}^n_\kappa(T; T') &= \bar{I}_\kappa(T; T').\end{aligned} \tag{3.6.8}$$

Note also that, as $q(X_0) = \mu_0$ is assumed stationary, so is the joint distribution $q_\kappa(X_0, T_0) = \mu_0\kappa$, which implies that, for all $n \in \mathbb{N}$,

$$\bar{q}_\kappa(X, T) = \bar{q}_\kappa(X', T') = q_\kappa(X_n, T_n).$$

Thus, denoting also by $\bar{I}_\kappa(X; T)$ and $\bar{I}_\kappa(X'; T')$ the mutual informations defined by the resp. distributions $\bar{q}_\kappa(X, T)$ and $\bar{q}_\kappa(X', T')$, we have, for all $n \in \mathbb{N}$,

$$\bar{I}_\kappa(X; T) = \bar{I}_\kappa(X'; T') = I_\kappa(X_n; T_n). \tag{3.6.9}$$

IB formulation of the decomposition into ergodic components

We are now ready to define the IB problem that will provide the information-theoretic characterisation, and softening, of the decomposition into ergodic component. Namely, for all $0 \leq \lambda \leq \Lambda := \bar{I}(X;X')$, we consider

$$\underset{\substack{\kappa \in \mathscr{K}(\mathcal{X},\mathcal{T}) \\ \bar{I}_\kappa(T;T') \geq \lambda}}{\arg\min}\ \bar{I}_\kappa(X;T), \tag{3.6.10}$$

where we recall that the mutual informations $\bar{I}_\kappa(X;T)$ and $\bar{I}_\kappa(T;T')$ are defined according to the joint law $\bar{q}(X,X',T,T')$, which itself is uniquely defined by the initial distribution $\mu_0 \in \Delta_{\mathcal{X}}$, the Markov chain $\tau \in \mathscr{K}(\mathcal{X})$, and the compression channel — see above, in particular equations (3.6.4) and (3.6.5). Given how we constructed the distribution $\bar{q}(X,X',T,T')$, this optimisation problem implements, intuitively, a trade-off between compression and the preservation of the *mutual information between the initial state and the time-average of the resulting trajectory* — which may be called, in short, the *mean-asymptotic mutual information*. Before presenting the main result of this section, let us describe problem (3.6.10) in more details.

From equation (3.6.7), the problem (3.6.10) does not have any solution for parameter $\lambda > \Lambda := \bar{I}(X,X')$. Moreover, using equation (3.6.3), it is easy to verify that the ergodicity of the stationary Markov chain (μ_0,τ) is equivalent to the condition $\bar{I}(X,X') = 0$: in particular, the problem (3.6.10) is only non-trivial for non-ergodic Markov chains. Note also that, from equation (3.6.9), the target function $\bar{I}_\kappa(X;T)$ here actually depends only on the single time-step distribution $q(X_n,T_n) = \mu\kappa$ (for any fixed $n \in \mathbb{N}$, e.g., $n = 0$); while from the construction above, the constraint function $\bar{I}_\kappa(T;T')$ depends on the distribution $q(\overrightarrow{T})$ of the whole bottleneck process — but it can be computed as an asymptotic limit of mututal informations, each depending on a finite number of time-steps (see equation (3.6.8)).

The problem (3.6.10) is very similar to a Symmetric IB (Slonim et al., 2006) with source the joint distribution $\bar{q}(X,X')$: the only difference is that here, we require the channels compressing resp. X and X' to be the same channel κ. It can also be easily verified that (3.6.10) is a Divergence IB (see Section 2.3.1 in Chapter 2) with data space $\mathcal{A} = \mathcal{X} \times \mathcal{X}$, distribution $\mu = \bar{q}(X,X')$, exponential family $\mathcal{E} = \Delta_{\mathcal{X}} \otimes \Delta_{\mathcal{X}}$ made of distributions making X and X' independent, and set of channel shape constraints requiring X and X' to be compressed separately but with the same channel, i.e.,

$$\mathscr{K}_{\mathsf{shape}} = \left\{ \kappa_{\mathcal{X}\times\mathcal{X}} \in \mathscr{K}(\mathcal{X}\times\mathcal{X},\mathcal{T}) \ :\ \exists \kappa_{\mathcal{X}} \in \mathscr{K}(\mathcal{X},\mathcal{T}) \ :\ \kappa_{\mathcal{X}\times\mathcal{X}} = \kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{X}} \right\}.$$

This makes the tools developed for the Divergence IB applicable to our new problem. Indeed, using Theorem B.1.3 in Appendix B.1, we obtain the following:

Theorem 3.6.1. *Let us recall that* $\mathsf{pr} : \mathcal{X} \to \mathcal{C}$ *denotes the projection on ergodic components w.r.t. the Markov chain* τ*. The following holds:*

(i) For $\lambda = \Lambda$*, a channel* $\kappa \in \mathscr{K}(\mathcal{X},\mathcal{T})$ *is a solution to* (3.6.10) *if and only if* $\kappa = \iota \circ \mathsf{pr}$*, for a congruent*[41] *channel* $\iota \in \mathscr{K}_{\mathsf{cong}}(\mathcal{C},\mathcal{T})$*.*

(ii) For all $0 \leq \lambda \leq \Lambda$*, all solutions* $\kappa \in \mathscr{K}(\mathcal{X},\mathcal{T})$ *to* (3.6.10) *satisfy* $\kappa = \gamma \circ \mathsf{pr}$*, for some* $\gamma \in \mathscr{K}(\mathcal{X},\mathcal{T})$*.*

Proof. See Appendix C.7.1. □

Crucially, point (i) shows that for maximal trade-off parameter $\lambda = \Lambda$, the solutions to the problem (3.6.10) coincide with the projection on ergodic components — up to congruent

[41] See Definition 2.2.2.

channels which, from the point of view of informational compression, are trivial transformations. In other words, we have reformulated (for finite Markov chains) the projection on ergodic components as an optimal compression under the constraint of preserving the mutual information between the initial state and the time-average of the resulting trajectory — or, in short, the process' *mean-asymptotic mutual information*. **This is one of the most important results of this chapter, and of this thesis as a whole**. We will comment further on its interpretation once we will have specialised it to group actions (see Corollary 3.6.2 below). For now, let us add some technical remarks.

Point (*ii*) in Theorem 3.6.1 shows that for *any* trade-off parameter λ, the solutions to (3.6.10) can be factorised by the projection pr. In other words, the coarse-grainings κ defined by the generalised IB problem (3.6.10) are always obtained by "post-processing", with some channel γ, the projection on ergodic components. However, it is not clear whether for $\lambda < \Lambda$, the post-processing γ captures any interesting structure. Indeed, from equation (3.6.4), we actually have $\mathsf{pr}(X) = \mathsf{pr}(X')$; in particular, $\mathsf{pr}(X)$ and $\mathsf{pr}(X')$ depend deterministically on one another. But this, combined with the fact that bottleneck channels κ can only "see" the source variable X, resp. X', through the lens of its projection on ergodic component $\mathsf{pr}(X)$, resp. $\mathsf{pr}(X')$, makes the problem (3.6.10) worryingly similar to the case, in the classic IB, where the relevancy variable is a deterministic function of the source variable (Tishby et al., 2000). For the latter, it is known that when $0 < \lambda < \Lambda$, the bottleneck solutions do not capture any interesting structure beyond that captured by the solution for maximal parameter $\lambda = \Lambda$ — more precisely, this is at least true for the Lagrangian version of the IB (Kolchinsky et al., 2019).

For that reason, it might be interesting to consider the finite-time counterpart of (3.6.10):

$$\underset{\substack{\kappa\in\mathscr{K}(\mathcal{X},\mathcal{T})\\ \bar{I}^n_\kappa(T;T')\geq\lambda}}{\arg\min}\ \bar{I}(X_0;T_0) \tag{3.6.11}$$

Indeed, when $\bar{q}^n_\kappa$ has not yet fully converged to $\bar{q}_\kappa$, this means, intuitively, that the trajectory's mean distribution has not fully "diffused" accross the ergodic components. Moreover, starting from a given point x_0, it is reasonable to expect that the trajectory's mean distribution will "diffuse" quicker to the points with which x_0 "communicates strongly",[42] i.e., points that are reached with relatively higher probability in a given number of time-steps. Information about this variability should then be carried by the distribution $\bar{q}^n_\kappa$ if $n \in \mathbb{N}$ is in an appropriate range. Intuitively, this uneven "diffusion structure" might then be captured by solutions to (3.6.11) for $0 < \lambda < \Lambda$, where the coarse-graining κ would first (stochastically) cluster "strongly communicating" points for large λ, and then cluster more and more "weakly communicating" points for decreasing λ.

Of course, the latter arguments are, at this stage, only heuristic conjectures. But they suggest that replacing the asymptotic mutual information $\bar{I}_\kappa(T;T')$ by its finite time-step counterpart $\bar{I}^n_\kappa(T;T')$ might be of interest in its own right. In any case, this focus on $\bar{I}^n_\kappa(T;T')$ will probably be necessary for algorithms computing or approximating solutions to (3.6.10). We leave to future work the design and implementation of such algorithms, and to investigate the scientific relevance of problem (3.6.11) for a wide range of values of $n \in \mathbb{N}$ and $0 \leq \lambda \leq \Lambda$.

Related work The quantity $\bar{I}^n_\kappa(X, X')$ above is reminiscent of the *persistent mutual information* (PMI) from (Ball et al., 2010), which is defined as the mutual information between the past of a stochastic process (up to time 0) and its future starting from a given time $n > 0$

[42] This informal term is inspired by the usual terminology of *communicating classes*, i.e., classes of points that can be reached, with positive probability, in finite time from one another.

that might be large. The differences with our quantity $\bar{I}^n_\kappa(X, X')$ are that (*i*) PMI considers double-sided processes, and (*ii*) the *average up to* time n considered in $\bar{I}^n_\kappa(X, X')$ is replaced in PMI by the *future from* time n. Moreover, PMI is used as a quantification of emergence in (Ball et al., 2010), while here we use $\bar{I}^n_\kappa(X, X')$ as a (partially or fully) preserved quantity in optimal compressions softening the notion of decomposition into ergodic components.

Application to the IB formulation of the projection on orbits

We can choose the Markov chain τ above to be the update channel $\bar{\rho}$ of a finite alphabet MDP (π, ρ). Consider, e.g., a deterministic transition channel $\rho \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X})$ corresponding to the action of a finite group $\mathcal{G}$ on a finite state-space $\mathcal{X}$. It is easy to verify that, because the action of the group on itself is transitive, the only group-stationary distribution on $\mathcal{G}$ is the uniform distribution $\upsilon \in \Delta_{\mathcal{G}}$ (see Definition 3.4.4). Denoting by π_υ the corresponding independent policy (see Definition 3.4.4 again), we thus obtain a measurable MDP (π_υ, ρ) whose update channel is

$$\bar{\rho} := \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \rho_g \in \mathscr{K}(\mathcal{X}). \tag{3.6.12}$$

It is easy to verify that a distribution $\mu_0 \in \Delta_{\mathcal{X}}$ is $\bar{\rho}$-stationary if and only if it is constant on each orbit of the group action — e.g., the uniform distribution on $\mathcal{X}$ is always $\bar{\rho}$-stationary, even though it is not the only one when the action of $\mathcal{G}$ on $\mathcal{X}$ is not transitive.

We are now ready to state our information-theoretic characterisation of a group action's projection on orbits:

Corollary 3.6.2. *Let $\rho \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X})$ be an action of a finite group $\mathcal{G}$ on a finite state-space $\mathcal{X}$, and $\bar{\rho}$ defined as in (3.6.12). Let $\mu \in \Delta_{\mathcal{X}}$ be full-support and $\bar{\rho}$-stationary, let $q(X, X') := \mu\bar{\rho}$ and for all $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$ with $\mathcal{T}$ countable,*

$$q_\kappa(X, X', T, T') := q(X, X')(\kappa \otimes \kappa) \in \Delta_{\mathcal{X}\times\mathcal{X}\times\mathcal{T}\times\mathcal{T}},$$

Then a channel κ is a solution of

$$\underset{\substack{\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T}) \\ I_\kappa(T;T') = I(X;X')}}{\arg\min} \quad I_\kappa(X;T) \tag{3.6.13}$$

if and only if $\kappa = \iota \circ \mathsf{pr}$, where $\iota \in \mathscr{K}_{\mathsf{cong}}(C, \mathcal{T})$ is a congruent channel, and $\mathsf{pr} \in \mathscr{K}(\mathcal{X}, C)$ is the projection on orbits w.r.t. the action ρ.

Proof. See Appendix C.7.2. □

Note that, in (3.6.13), the asymptotic mutual information $\bar{I}_\kappa(T;T')$ from (3.6.10) becomes a single time-step mutual information. This is because here, we assumed that the *whole* group action is contained in the transition channel ρ, so that single time-step actions already exhaust all possible multiple time-step actions: i.e., for any $n \in \mathbb{N}$ and $g_0, \dots, g_n \in \mathcal{G}$, there exists some $g \in \mathcal{G}$ such that $\rho_{g_0} \circ \dots \circ \rho_{g_n} = \rho_g$. As actions g are sampled uniformly across the group $\mathcal{G}$, Corollary 3.6.2 implies, intuitively, that the "information" that two consecutive states of the trajectory carry about each other is the orbit to which both belong: i.e., it is precisely the "information" that is invariant under the action of any element $g \in G$. This result thus formalises the intuition that the projection on orbits is the *coarsest compression preserving all the information that is invariant under the group action*.

This result is a crucial step forward concerning the limitations that we identified at the end of Chapter 2 (see Section 2.5.3 there). I.e., Corollary 3.6.2 eventually characterises in an

information-theoretic language what it means, for a compression channel, to be the projection on orbits of a given group action (up to composition by congruent channels at the output). Crucially, this result both yields a natural softening of the notion of projection on orbits — through solutions to (3.6.13) with $\lambda < \Lambda := I(X; X')$ — and generalises seamlessly to the case of stochastic actions — through the more general problem (3.6.10), and the corresponding Theorem 3.6.1. Of course, the transformation-based bottleneck framework outlined in Section 2.5.3 of Chapter 2 still deserves a more explicit presentation, and a proof that it does achieves the aims stated there. Moreover, this would likely require moving beyond the stationarity assumption that we make in the current chapter. While these tasks are beyond the scope of this thesis, our now completed toolbox yields a strong basis to carry them out in future work.

More generally, beyond the specific work on channel equivariance from Section 2.5.3, Theorem 3.6.1 and Corollary 3.6.2 above are a **stepping stone in the formalisation of the intuition of "duality" between information parsimony and symmetry that underlies this whole thesis**. Indeed, in the edge case $\lambda := \Lambda$ the trade-off in the problem (3.6.10) makes the coarse-graining implemented by its solution κ *preserve precisely what is being left invariant by the dynamics* of the Markov chain τ on $\mathcal{X}$ — which boils down to the partition in orbits if $\tau := \bar{\rho}$ is the average of a group action over its stationary probability, as in Corollary 3.6.2. Then, for $\lambda < \Lambda$, the state-space is further coarse-grained, while still preserving along the way as much information as possible about the features left invariant by τ. I.e., in short, we exhibited an informational trade-off that *binds* a compression channel and a transformation channel, in a way that requires the coarse-graining implemented by the compression channel to capture the (exact or soft) invariants of the transformation channel.

We thus expect Theorem 3.6.1 to play, in future work, the role of a bridge between the realm of information parsimony — mostly "populated" by compression channels, i.e., (potentially stochastic) maps from a given space to a simpler space — and the realm of symmetries — mostly "populated" by families of transformations, i.e., (potentially stochastic) maps from a space to itself.

3.6.2 Information-theoretic characterisation of minimal joinings

We now turn to an information-theoretic formulation of minimal joinings of a finite family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ of finite alphabet stationary MDPs. We will mostly focus on canonical joinings, as Proposition 3.5.14 implies that for finite alphabets, this yields no loss of generality.

Let us recall that the definition of minimal joinings is motivated by the intuition of an MDP collectively describing the MDPs in a given family, but in a way that captures as much as possible the common dynamics across the MDPs in the family (see Section 3.5.2). An alternative direction is thus to search for a joining that makes each MDP in the family *depend as much as possible on the other ones*. Information theory provides a principled way to formalise this intuition, through the notion of *mutli-information*.

Distribution and channel multi-information

Definition 3.6.3. Let $(\mathcal{A}^c)_{c\in C}$ be a finite family of finite sets, $\boldsymbol{\mathcal{A}}$ their Cartesian product, $\boldsymbol{q} := \boldsymbol{q}(A_1, \dots, A_n) \in \Delta_{\boldsymbol{\mathcal{A}}}$ and for q_i the marginal of $\boldsymbol{q}$ on the coordinate $\mathcal{A}^c$, define $\boldsymbol{q}^{\otimes} := \bigotimes_{c\in C} q^c$. Then, denoting by D the Kullback-Leibler (KL) divergence, the *multi-information of the distribution* $\boldsymbol{q}$ is defined as

$$\boldsymbol{I}(\boldsymbol{q}) := \boldsymbol{I}\left((A^c)_{c\in C}\right) := D(\boldsymbol{q}||\boldsymbol{q}^{\otimes}).$$

It can be verified that the multi-information $\boldsymbol{I}(\boldsymbol{q})$ minimises the KL divergence between $\boldsymbol{q}$ and distributions on the exponential family of independent distributions (Amari, 2001). In

this sense, multi-information measures "how far from making the marginals q^c independent" is $\boldsymbol{q}$: i.e., the "spatial interdependence" of the distinct coordinates. Multi-information can be characterised in terms of entropies (Amari, 2001):

Proposition 3.6.4. *With the same notations as in Definition 3.6.3, we have*

$$\boldsymbol{I}(\boldsymbol{q}) = \sum_{c\in C} H(q^c) - H(\boldsymbol{q}).$$

Also relevant to us is the following extension of multi-information to the dynamical setting, proposed in (Ay, 2015; Ay et al., 2003). It relies on an extension the KL divergence D to channels:

Definition 3.6.5. Let $\mathcal{A}, \mathcal{B}$ finite sets, $q \in \Delta_{\mathcal{A}}$ and $\gamma, \gamma' \in \mathscr{K}(\mathcal{A}, \mathcal{B})$. Then, for D the KL divergence, we define

$$D_q(\gamma||\gamma') := \sum_{a\in\mathcal{A}} q(a) D(\gamma(\cdot|a)||\gamma'(\cdot|a)).$$

Definition 3.6.6. Let $(\mathcal{A}^c)_{c\in C}$, $(\mathcal{B}^c)_{c\in C}$ be two finite family of finite sets, $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$ the corresponding Cartesian products, $\boldsymbol{q} := \boldsymbol{q}\big((A^c)_{c\in C}\big) \in \Delta_{\boldsymbol{\mathcal{A}}}$ with marginal q^c on each $\mathcal{A}^c$, and a channel $\boldsymbol{\gamma} \in \mathscr{K}(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}})$. For all $c \in C$, a *marginal of the channel* $\boldsymbol{\gamma}$ *w.r.t. the distribution* $\boldsymbol{q}$ *on the coordinate* $\mathcal{A}^c \times \mathcal{B}^c$ is a channel $\gamma^c \in \mathscr{K}(\mathcal{A}^c, \mathcal{B}^c)$ such that $q^c\gamma^c$ is a marginal of the distribution $\boldsymbol{q}\boldsymbol{\gamma}$ on the coordinate $\mathcal{A}^c \times \mathcal{B}^c$. Let us now fix a family $(\gamma^c)_{c\in C}$ of marginals of $\boldsymbol{\gamma}$ w.r.t. $\boldsymbol{q}$ on each coordinate $\mathcal{A}^c \times \mathcal{B}^c$, and define $\boldsymbol{\gamma}^{\otimes} := \bigotimes_{c\in C} \gamma^c$. The *multi-information of the channel* $\boldsymbol{\gamma}$ *w.r.t. the distribution* $\boldsymbol{q}$ is defined as[43]

$$\boldsymbol{I}_{\boldsymbol{q}}(\boldsymbol{\gamma}) := D_{\boldsymbol{q}}(\boldsymbol{\gamma}||\boldsymbol{\gamma}^{\otimes}).$$

It can be verified that the quantity $\boldsymbol{I}_{\boldsymbol{q}}(\boldsymbol{\gamma})$ is the same for all choice of channel marginals γ^c (which might not be unique if $\boldsymbol{q}$ is not full-support), and that it minimises $D_{\boldsymbol{q}}(\boldsymbol{\gamma}||\boldsymbol{\gamma}')$ among all *split* channels γ' (Ay, 2015), i.e., channels with a tensor product form $\boldsymbol{\gamma}' := \bigotimes_{c\in C}(\gamma')^c$. The quantity can thus be understood as measuring "how far from being parallel" is the processing performed by the channel $\boldsymbol{\gamma}$: i.e., "how jointly" are distinct coordinates processed.

The following relation between distribution multi-information and channel multi-information can be verified with a straightforward computation:[44]

Proposition 3.6.7. *With the same notations as in Definition 3.6.6, we have, using the hook-up notation,*

$$\boldsymbol{I}(\boldsymbol{q}\boldsymbol{\gamma}) = \boldsymbol{I}(\boldsymbol{q}) + \boldsymbol{I}_{\boldsymbol{q}}(\boldsymbol{\gamma}).$$

For instance, if $\boldsymbol{\mathcal{A}} := \boldsymbol{\mathcal{B}} := \boldsymbol{\mathcal{X}}$ and $\boldsymbol{\gamma} := \boldsymbol{\tau} \in \mathscr{K}(\boldsymbol{\mathcal{X}})$ is seen as Markov chain, Proposition 3.6.7 means that the "spatio-temporal" multi-information $\boldsymbol{I}(\boldsymbol{q}\boldsymbol{\gamma}) = \boldsymbol{I}\big((X_0^c, X_1^c)_{c\in C}\big)$ of two consecutive time-steps decomposes as the sum of the "spatial" stochastic interdependence $\boldsymbol{I}(\boldsymbol{q})$ and the "temporal" stochastic interaction $\boldsymbol{I}_{\boldsymbol{q}}(\boldsymbol{\gamma})$ (Ay et al., 2003).

Minimal joinings as maximum multi-information, i.e., minimum entropy joinings

Equiped with these concepts, let us come back to our information-theoretic reformulation of minimal joinings. Note that here, a stationary MDP (μ_0, π, ρ) is characterised by its one-time-step process distribution $q(X_0, G_0, X_1) := \mu_0\pi\rho$ (see Lemma C.5.1). This suggests to

[43] In (Ay et al., 2003), this quantity is defined for $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{B}}$ and called *stochastic interaction*, but for $\boldsymbol{\mathcal{A}} \neq \boldsymbol{\mathcal{B}}$ the term "interaction" is less adapted.

[44] We are not aware of a previous publication mentioning Proposition 3.6.7, even though we would not be surprised if this simple fact has already been noticed.

formalise our previous intuition of a joining "maximising the interdependence" between the MDPs of the family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ as a canonical joining $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$ maximising, among all canonical joinings, the multi-information

$$\boldsymbol{I}(\boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho}) = \boldsymbol{I}\big((X_0^c, G_0^c, X_1^c)_{c\in C}\big) = D\left(\boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho}\middle\|\bigotimes_{c\in C}\mu_0^c\pi^c\rho^c\right),$$

where $(X_0^c, G_0^c, X_1^c)_{c\in C}$ denotes the joint variable describing, for each coordinate $c \in C$, the corresponding state X_0^c at time 0, action G_0^c at time 0, and state X_1^c at time 1. Applying Proposition 3.6.7 twice shows that

$$\boldsymbol{I}(\boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho}) = \boldsymbol{I}(\boldsymbol{\mu}_0) + \boldsymbol{I}_{\boldsymbol{\mu}_0}(\boldsymbol{\pi}) + \boldsymbol{I}_{\boldsymbol{\mu}_0\boldsymbol{\pi}}(\boldsymbol{\rho}), \tag{3.6.14}$$

where the index $\boldsymbol{\mu}_0\boldsymbol{\pi} \in \Delta_{\mathcal{X}\times\mathcal{G}}$ of the last term uses the hook-up notation (see Definition 3.2.3). In other words, we are actually proposing to maximise three distinct quantities: the *spatial interdependence* (Ay et al., 2003) across the distinct MDPs' state-spaces; the amount to which the policy $\boldsymbol{\pi}$ *jointly selects* distinct action coordinates from distinct state coordinates; and the amount to which state coordinates are *jointly transformed* through joint actions.

Note that from Proposition 3.5.7, in the case of deterministic policies $(\pi^c)_{c\in C}$ and deterministic transition channels $(\rho^c)_{c\in C}$, we have $\boldsymbol{\mu}_0\boldsymbol{\pi} = \boldsymbol{\mu}_0\boldsymbol{\pi}^{\otimes}$ and $(\boldsymbol{\mu}_0\boldsymbol{\pi})\boldsymbol{\rho} = (\boldsymbol{\mu}_0\boldsymbol{\pi})\boldsymbol{\rho}^{\otimes}$, so that actually the quantity above becomes

$$\boldsymbol{I}(\boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho}) = \boldsymbol{I}(\boldsymbol{\mu}_0),$$

i.e., we are only maximising the spatial interdependency. But crucially, Proposition 3.5.7 also shows that if the policies $(\pi^c)_{c\in C}$ and transition channels $(\rho^c)_{c\in C}$ are stochastic, the terms $\boldsymbol{I}_{\boldsymbol{\mu}_0}(\boldsymbol{\pi})$ and $\boldsymbol{I}_{\boldsymbol{\mu}_0\boldsymbol{\pi}}(\boldsymbol{\rho})$ in equation (3.6.14) are in general non-trivial.

Now, Proposition 3.5.2 shows that for any canonical joining $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$, the marginal of $\boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho} \in \Delta_{\mathcal{X}\times\mathcal{G}\times\mathcal{X}}$ on the coordinate $c \in C$ is always $\mu_0^c\pi^c\rho^c \in \Delta_{\mathcal{X}^c\times\mathcal{G}^c\times\mathcal{X}^c}$. Therefore, from Proposition 3.6.4,

$$\boldsymbol{I}(\boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho}) = \sum_{c\in C} H(\mu_0^c\pi^c\rho^c) - H(\boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho}),$$

where $\sum_{c\in C} H(\mu_0^c\pi^c\rho^c)$ does not depend on the choice of the joining. This shows that a canonical joining maximises $\boldsymbol{I}(\boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho})$ among all joinings of the family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ if and only if it minimises the joint entropy $H(\boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho})$. We choose our terminology to reflect this latter point of view:

Definition 3.6.8. Let $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ be a finite family of finite-alphabet stationary MDPs. A *minimum entropy joining* of $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ is a joining $(\nu_0, \eta, \xi) \in \mathsf{Join}((\mu^c, \pi^c, \rho^c)_{c\in C})$ such that

$$H(\nu_0\pi\rho) = \min_{(\nu_0',\eta',\xi')\in\mathsf{Join}((\mu^c,\pi^c,\rho^c)_{c\in C})} H(\nu_0'\pi'\rho'), \tag{3.6.15}$$

i.e.,

$$H(P_0, K_0, P_1) = \min_{(\nu_0',\eta',\xi')\in\mathsf{Join}((\mu^c,\pi^c,\rho^c)_{c\in C})} H(P_0', K_0', P_1').$$

where (P_0, K_0, P_1) and (P_0', K_0', P_1') are the resp. marginals on the first time coordinates of the resp. state-action processes $\overrightarrow{(P, K)} := (P_0, K_0, P_1, K_1, \dots)$ and $\overrightarrow{(P', K')} := (P_0', K_0', P_1', K_1', \dots)$ defined by the resp. stationary MDPs (ν_0, η, ξ) and (ν_0', η', ξ') (see Definition 3.2.10).

Technical remark. The minimum is indeed always achieved in (3.6.15), as it can be verified that the set $\mathsf{Join}((\mu_0^c, \pi^c, \rho^c)_{c\in C})$ is compact, and the entropy is continuous w.r.t. the underlying distribution and thus achieves its minimum on a compact set.

As here, any joining can be factorised through a canonical joining (see Proposition 3.5.14), it can be easily verified that it would have been equivalent to restrict ourselves to canonical joinings and require the maximisation of multi-information. The latter point of view provides a conceptually important interpretation, and a link to information-theoretic approaches to complexity (Ay, 2015). But focusing on the minimum entropy perspective allows one to "forget" the product structure of the spaces $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{G}}$, thus potentially working with simpler ambient spaces. As entropy can intuitively be understood as a measure of "how spread" a distribution is, this perspective also highlights that we are here requiring the joining's process distribution to be "as concentrated as possible". In this sense, we are here dealing — again — with a notion of *information parsimony*. Let us point out that this kind of information parsimony does not fit the traditonal information-theoretic formalisation of Ockham razor's principle with the *maximum* entropy principle (Jaynes, 1957) — but this is just an example of how helpful mathematical formalisation can be to fine-tune the meaning of similar concepts to distinct contexts.

To the best of our knowledge, no notion of minimum entropy joining has been previously considered in the literature. However, minimum entropy joinings can be seen as a generalisation of *minimum entropy couplings* to a dynamical setting: a minimum entropy coupling of a finite family of distributions $(\mu_0^c)_{c\in C}$ on finite alphabets $(\mathcal{X}^c)_{c\in C}$ is a joint distribution $\boldsymbol{\mu}_0$ on the product space $\boldsymbol{\mathcal{X}}$ which minimises the entropy $H(\boldsymbol{\mu}_0)$ among all couplings of $(\mu_0^c)_{c\in C}$, i.e., all joint distributions whose marginal on each coordinate $\mathcal{X}^c$ is μ_0^c. It can be easily verified that a minimum entropy coupling is a minimum entropy joining for an MDP whose only action is the identity transformation. As minimum entropy couplings have been thoroughly studied in recent years (Bounoua et al., 2025; Cicalese et al., 2019; Li, 2021; Ma et al., 2025), this might be helpful for the study and computation of minimum entropy joinings.

Now, the reason why we are presenting minimum entropy joinings is because they happen to provide an information-theoretic characterisation of minimal joinings:

Theorem 3.6.9. *Let $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ be a finite family of finite-alphabet MDPs. Then:*

(i) For (ν_0, η, ξ), (ν_0', η', ξ') two joinings of $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$, we have, using the j-factor notation (see Definition 3.5.10),

$$(\nu_0', \eta', \xi') \lesssim (\nu_0, \eta, \xi) \quad \Rightarrow \quad H(\nu_0'\eta'\xi') \leq H(\nu_0\eta\xi). \tag{3.6.16}$$

(ii) A joining has minimum entropy if and only if it is a minimal joining.

(iii) All minimal joinings are isomorphic as stationary MDPs.

Proof. See Appendix C.7.3. □

Note that in equation (3.6.16), the comparison of entropies on the right-hand-side induces a relation on joinings of a given (finite) family of (finite alphabet) stationary MDPs. This relation is a *total pre-order*, i.e., it is not only transitive and reflexive as all pre-orders, but also *strongly connected* (which means that two joinings can always be compared). Point (*i*) can be seen as the fact that the latter total pre-order is a "coarse-graining" of the j-factor pre-order. Note that this "'coarse-graining" collapses the potentially rich pre-order structure on the potentially high-dimensional dimensional space of all joinings of a given family,[45] into a total order that only compares joinings through scalar values.

[45] It can be easily verified that the space of all joinings of a given family is stable under convex combinations. Thus it is a convex subset of an affine space, whose dimension is that of the vector space giving its direction.

Crucially, point (*ii*) shows that, despite this fact, *the total pre-order defined by the comparison of entropies has precisely the same minimal elements as those of the much richer j-factor pre-order* — i.e., mininal joinings. In other words, we obtained the **equivalence of an information-theoretic and an algebraic notion of minimality** — where "algebraic" here refers to the fact that the j-factor pre-order is, at core, based on commutation relations. Eventually, point (*iii*) shows that these common minimal elements can be seen, from the point of view of the stationary MDP structure, as an essentially unique minimal element. In particular, with Theorem 3.6.9, we obtain a new instance of *structure emerging through information parsimony requirements*.

Similarly as we did in Section 3.6.1 as well as in Chapter 2, Theorem 3.6.9 suggests that it is possible to obtain a characterisation of minimal joining with a full Information Bottleneck-like optimisation problem, whose solutions would be minimal joinings of a given family of stationary MDPs — where the minimisation of the entropy $H(\nu_0\eta\xi)$ would be traded-off against the preservation of one or several other well-chosen information-theoretic quantities. The latter would need to characterise the property that the MDP (ν_0, η, ξ) is a joining of the family of stationary MDPs $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ with family of marginalisation maps $(\phi^c, \psi^c)_{c\in C}$: i.e., from Proposition 3.5.2, that

(i) The distribution ν_0 is $\bar{\xi}$-stationary, where $\bar{\xi}$ is the update channel defined by the policy η and the transition channel ξ,

(ii) For all $c \in C$, we have $(\phi^c \otimes \psi^c \otimes \phi^c) \cdot \nu_0\eta\xi = \mu_0^c\pi^c\rho^c$.

For instance, the latter conditions can be achieved by requiring that

(i)' $D(\bar{\xi} \cdot \nu_0||\nu_0) = 0$,

(ii)' For all $c \in C$, we have $D((\phi^c \otimes \psi^c \otimes \phi^c) \cdot \nu_0\eta\xi||\mu_0^c\pi^c\rho^c) = 0$,

where D denotes here the Kullback-Leibler (KL) divergence. However, while conditions (i)' and (ii)' do provide an information-theoretic characterisation of the property of being a joining, it is not clear that a multi-objective optimisation problem involving the corresponding KL divergences, together with the entropy $H(\nu_0\eta\xi)$, would capture interesting structures once conditions (i)' and (ii)' are softened — i.e., along the full *Pareto front* defined by the multi-objective optimisation problem. Indeed, first, the KL divergence is not symmetric, and it is not clear how the choice done in (i)' and (ii)' would differ from the same quantities where the distributions inside the divergences are switched.

But more fundamentally, while the extensive literature on the Information Bottleneck (IB) framework, broadly understood, has clearly proven its relevance to the information parsimony-induced exploration of underlying structure in data (see Section 1.1.2), the divergences in (i)' and (ii)' seem to be of a different nature than those usually considered in the IB framework. Indeed, they correspond neither to mutual informations, nor even to a divergence from a (non-trivial) hierarchical model as in Chapter 2 — and switching the probabilities inside the resp. KL divergences does not seem to change that point. Another crucial difference with previously considered IB frameworks comes from the fact that we have generalised the class-pose decomposition framework by "reversing the arrows and breaking the bijectivity" in the commutation relations (see Sections 3.1.2 and 3.5.3). As a consequence, it would be the data defining the multi-objective optimisation problem — i.e., here, that defined by the MDPs $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ — that would be a "coarse-graining" of the solution to this optimisation problem — i.e., here, the minimal joining (ν_0, η, ξ), together with its marginalisation maps $(\phi^c, \psi^c)_{c\in C}$ — rather than the converse (as, e.g., in the classic IB framework, the bottleneck variable is a obtained by "coarse-graining" the source variable). In other words, the

picture that emerges here is one where **the "bottleneck" metaphor is in some sense turned on its head**.

These considerations suggests that a full information-theoretic trade-off characterising joinings could be starkly different from previously considered IB frameworks. This is, of course, not a problem in itself, but it makes it only more important to develop dedicated theoretical motivations, algorithmic tools, and data-based explorations. Such a framework might not necessarily involve the information-theoretic quantities from points (i)' and (ii)' above — which, at this stage, we only present for illustrative purposes.

We leave further investigation of these questions to future work. Note that solving them could open the way, once combined with our reformulation of ergodic components from Section 3.6.1, to the information-theoretic discovery of — "exact" and "soft" — minimal class-pose parametrisations (see Section 3.5).

3.7 Limitations

We mentioned, along the course of this chapter, a number of limitations to the content presented here. Here we summarise the most important ones, and outline a few others. Let us start with the most theoretical ones.

First, we faced several technical limitations in our measure-theoretic results on stationary MDPs. These limitations were often due to the fact that we want to be able to consider joinings of an *uncountable* number of MDPs. This requirement is important because, in the continuous case, there will typically be a uncountable number of ergodic components to be joined — e.g., even for the simple example of polar coordinates, there is an uncoutable number of orbits, i.e., of circles centered at the origin. However, this seems (at least in the case of canonical joinings) to require working, in general, with the product of an uncountable number of measurable spaces. The latter is problematic because (*i*) even if each coordinate is standard Borel, their uncountable product will never be standard Borel, and (*ii*) the usual product σ-algebra on the product space yields measurable functions that must depend on only a countable number of coordinates. Future work could try to address these limitations by either imposing stronger conditions on the considered MDPs (e.g., well-chosen notions of continuity), or by considering recent advances aiming precisely at solving issues arising from uncountability in measure and ergodic theory (Fritz et al., 2025; Jamneshan et al., 2023; Moss et al., 2023).

Attentive readers of the proofs might also have noticed that in the measure-theoretic setting, intuitions that are conceptually straightforward require a disproportionate amount of mathematical effort to make the technical machinery work. This suggests that, to further develop the formalism laid out here, we should properly *axiomatise* it: i.e., here, identify a relatively limited set of properties whose proof might require a substantial amount of work, but that, once proven, can be composed to obtain all the remaining statements of the theory without having to always dive back into measure-theoretic technicalities. While it does not seem unavoidable to call on this framework, *category theory* (Mac Lane, 1978) might be naturally suited for that purpose. Such a categorical formulation could also be instrumental for future generalisations beyond stationary MDPs. Indeed, it seems that many of the notions developed in Section 3.5 depend, at core, only on a given notion of *factor* — be it in the category of group actions, stationary MDPs or another one. Category theory-based proofs would have the advantage of being easily transferable from one category of systems to the other.

Whether or not using the help of a categorical language, such generalisations beyond stationary MDPs would first require to remove the stationarity assumption — e.g., by replacing it by *asymptotic mean stationarity* (Gray, 2011). However, for a full relevance to the modeling of concrete embodied agents, it will be necessary to move beyond the MDP framework itself: possible generalisations include, e.g., partially observable MDPs, input-output processes and

their ϵ-transducers (Barnett et al., 2015; Rosas et al., 2025) , or causal Bayesian network models of the perception-action loop (Ay et al., 2014; Polani et al., 2009).

While these abstract considerations are crucial for the long-term maturation of the theory outlined in this chapter, the concepts that we defined should be studied in more details on concrete examples — e.g., to start with, in the finite case. But to really ground this theory in data-based research — in particular sensorimotor perception — or even simple synthetic experiments, we need to develop algorithms to compute the mathematical objects defined here. Given the stochastic nature of these objects, it seems natural, at least in the finite case, to aim for information theory-based algorithms. This is part of our motivation for our information-theoretic characterisations of ergodic components and of minimal joinings. On the one hand, as we characterised ergodic components with an IB-like problem, it is reasonable to expect that one can derive a Blahut-Arimoto algorithm for the finite case; and that for extensions to the continuous case, one could adapt the algorithmic technology extensively developed, in the machine learning literature, for previous versions of the IB framework — see (Hu et al., 2024) for a recent review. As far as minimal joinings are concerned, as mentioned in Section 3.6.2, it seems natural to search for a full multi-objective optimisation problem that would characterise these mathematical objects. If this is possible, it would then be necessary to design algorithms to solve this problem, and it is not clear that methods from the IB framework would readily adapt here.

3.8 Discussion

3.8.1 Formal contribution

The framework developed in this chapter was motivated by the question of how adaptive systems identify structure in the interaction with their environment — in particular, sensorimotor theories of perception, which put a strong emphasis on how perceptually relevant structure is induced by the agent's own actions. This lead us to focus on previous work on class-pose decomposition which studies, in the language of group actions, a decomposition of a given state-space $\mathcal{X}$ into two action-induced coordinates $\mathcal{C}$ and $\mathcal{P}$: one that captures the features that are invariant under actions, and another that tracks the changes induced by actions, and only that.

Our main proposal is to generalise class-pose *decomposition* into *minimal* class-pose *parametrisation* in three directions: algebraic, dynamical, and information-theoretic. We started by presenting the algebraic aspect in the group-theoretic setting. From this perspective, it consists in moving from a state-space action ρ that is isomorphic to the class-pose action $\mathtt{Id}_{\mathcal{C}} \otimes \xi$, to ρ being only a *factor* of $\mathtt{Id}_{\mathcal{C}} \otimes \xi$: i.e., there exists a surjective (but not necessarily injective) map $\phi : \mathcal{C} \times \mathcal{P} \to \mathcal{X}$ that intertwines $\mathtt{Id}_{\mathcal{C}} \otimes \xi$ and ρ, in the sense that $\phi \circ (\mathtt{Id}_{\mathcal{C}} \otimes \xi) = \rho \circ \phi$. Crucially, we require this factor to be "as isomorphic as possible", in the sense of being minimal w.r.t. a pre-order based on the factor relation. The dynamical aspect of our generalisation consists in generalising these notions to potentially non-invertible, stochastic and closed-loop actions, using Markov Decision Processes (with no rewards and a fixed policy) on standard Borel spaces (which include, e.g., countable spaces, Euclidean spaces and differential manifolds). Classes are thus reframed as ergodic components under the actions of an MDP averaged over a given policy. These ergodic components yield a corresponding family of ergodic MDPs, whose *minimal joining* defines the novel pose coordinate. This ergodic-theoretic reformulation of classes and poses provides a bridge (in the case of finite state and action spaces) towards an information-theoretic formulation, where the projection on ergodic components (i.e., on classes) is characterised as an optimal compression preserving the process' *mean asymptotic mutual information*; while minimal joinings (i.e., poses and

their dynamics) are shown to coincide with *minimum entropy joinings* — or, equivalently, joinings that maximise the multi-information between the ergodic MDPs.

3.8.2 Conceptual contribution to sensorimotor theories of perception

The results summarised above are, at core, a mathematical contribution. The mathematical language is also a tool to pursue conceptual advances which, we believe, research on both structure learning and sensorimotor perception is in dire need of. Indeed, beyond the question of *how* to learn structure, these fields raise the foundational questions of *what do we mean by structure*, and *why should it be relevant to a real-world agent*, in particular in the case of biological agents.

Our focus on an information-theoretic reformulation is motivated by the latter "why" question, as it incorporates structure discovery into a tradition that uses information theory to formalise first principles of adaptive behaviour (Archer et al., 2022; Krakauer et al., 2020; Ortega et al., 2013; Salge et al., 2014; Tishby et al., 2011). Indeed, showing that specific structures, in the agent-environment interaction, emerge from specific forms of information parsimony — as we did in this chapter and the previous one — is a way of understanding why these structures would be relevant to embodied agents whose behaviour unfolds under stringent informational constraints. More precisely, our mathematical results contribute to formalising the intuition that real-world agents leverage structure (in the interaction with their environment) because, as information-constrained systems, *they cannot afford not to do so*. This is an instance of the idea that *meaning emerges from informational constraints* on a given agent.

As for the "what" question, here we focus on the structure of minimal class-pose parametrisations. Crucially, the latter yields a shift from understanding agents as *discovering* structure to understanding them as *creating* structure, in two distinct ways. First, the MDP formalism brings us much closer to a class-pose structure that is generated by the agent's own behaviour — as, here, the decomposition into ergodic components is defined by the action-dependent transition channel and policy. This approach can be understood as investigating the structure of the *sensorimotor coupling* of the agent to its environment (Aguilera et al., 2013; Buhrmann et al., 2013), and a generalisation of our framework to input-output processes — potentially through the notion of ϵ-transducer (Barnett et al., 2015; Rosas et al., 2025) — could go further in this direction. But a more fundamental notion of "structure creation" is also at play in our formalism: the fact that, as mentioned in Section 3.1.2, here the "direction of the arrows is reversed". I.e., it is the class-pose space which is projected onto the state-space, not the converse — and crucially, this projection is potentially non-isomorphic. If the class-pose space corresponds to an agent's internal dynamics and the state-space to an "outside" variable (the environment, or even just the agent's sensorimotor interface, as in ϵ-transducers (Barnett et al., 2015; Rosas et al., 2025)), it means that *the internal dynamics can in some respects be richer* than the "outside" variable's dynamics — and this holds despite the internal dynamics being "minimal" or "parsimonious". In other words, while our framework does understand poses as an abstraction emerging from the state space's dynamics through the right kind of information parsimony, here this "pose" abstraction is *not* a coarse-graining of the state-space variable, but exists somehow at a "fictional", i.e., self-generated level — to the extent that the minimal joining is non-isomorphic. This "fiction", however, might be highly relevant to the agent's behaviour. Indeed, as mentioned in this chapter's Section 3.1.1 and in Section 1.2, from a sensorimotor perception perspective, classes can be interpreted as object-related SMCs defined by a given behaviour (i.e., policy). The dynamics of poses (including their potentially non-isomorphic, self-generated dimension) can thus be interpreted as a *parsimonious joint description* of the way how, under a given behaviour, the SMCs corresponding to *any* "perceptual object" (i.e., class) defined by this behaviour is explored by

the agent's actions (O'Regan et al., 2001). More generally, our novel notion of minimal joining could operationalise the intuition that embodied agents make sense of their sensorimotor history through *parsimonious fictions* induced by different episodes of this history that leave their trace, as much as possible, on the same brain dynamics.

Moreover, both aspects of this "structure creation" interpretation resonate strongly with the *inside-out* approach to neuronal dynamics — which essentially proposes that the brain does not passively process information, but creates it by inducing "perturbations" of its self-sustained dynamics through interaction with the environment (Buzsáki et al., 2019) (see Section 1.1.3). See also (Santos et al., 2022) for a minimal model along similar lines.

Our formalism also bears similarities with *Closed-Loop Perception* (CLP) theory (Ahissar et al., 2016, 2025). Indeed, let us recall that ergodic components are defined as equivalences classes of points that have the same limit of their resp. Césaro means (see equations (3.3.1) and (3.3.2)), which can be seen, intuitively, as the fact that asymptotically, their resp. trajectories are concentrated on the same attractor. Moreover, we are here proposing to interpret these ergodic components, i.e., classes, as object-related SMCs — which are a notion of sensorimotor percept. On the other hand, CLP theory proposes to define percepts as attractors in the sensori-neural-motor dynamics. While the MDP framework is of course limited to capture the latter dynamics in full generality, this suggests that generalisations of our framework could yield a notion of "class" that simultaneously captues CLP theory's and SMC theory's notions of percept.

Eventually, while we chose to keep using the "class" terminology from the "class-pose decomposition" framework, this does not have to be interpreted as agents performing "classification" in a representationalist sense — where classes would be the output of some agent's internal processing. Rather, recasting class as an MDP's ergodic component makes this mathematical object describe a specific attractor of the agent's *enacted behaviour* — similarly as in (Buhrmann et al., 2013), categorical perception is operationalised through the convergence to a dynamical attractor.[46]

3.8.3 Towards a broader research program

Overall, these ideas hint at a possible, broader program of understanding sensorimotor perception as some kind of "generalised arithmetics" (de la Rue, 2006, 2023) over the dynamics generated by the agent's own behaviour (Buhrmann et al., 2013; Olsson et al., 2006). Crucially, these dynamical "arithmetics" would not be understood as a representational processing by "computations" performed in the brain. Rather, they would describe the process of *attunement* (O'Regan et al., 2001) of the whole sensori-neural-motor loop along the agent's history. More precisely, development and learning would poise the dynamics of this loop close to appropriate trade-offs, for well-chosen multi-objective optimisation problems capturing the tension between conflicting norms underlying the agent's behaviour (e.g., information parsimony and other behaviourally relevant quantities). This would induce these dynamics to enact "soft" versions of "least common multiple"- or "greatest common divisor"-like properties — i.e., of what category theorists call *universal properties* (Mac Lane, 1978). Perceptual structure would then be defined as the enactment of these dynamical, norm-induced and "soft" universal properties.

Such a program would be complementary to previous formalisations of the emergence of an embodied agent's perceived world (Ay et al., 2015; Capdepuy et al., 2007) and of patterns from dynamics (Barnett et al., 2015; Rosas et al., 2025; Rosas et al., 2024; Rupe et al., 2022), whose underlying mathematical objects are mostly *coarse-grainings* or *factors* of a given

[46] While (Buhrmann et al., 2013) relies on a topological notion of attractor, ergodic components can be seen as a measure-theoretic one. Indeed, ergodic components identify points with the same Césaro mean, i.e., intuitively, points whose trajectories asymptotically "concentrate" of the same subset.

process (Pfante et al., 2014, 2015; Travers et al., 2025), by introducing to this picture *joinings* of distinct, but related processes (e.g., ergodic components of MDPs). Our formalism suggests that ergodic theory — and generalisations of it (Moss et al., 2023) — could here play an important role, through the theory of joinings (de la Rue, 2006, 2023; Glasner, 2003), but more generally because it both investigates structural properties of measurable group actions (Kerr et al., 2016) and has strong links with information theory (Billingsley, 1965; Gray, 2011). A generalised version of ergodic theory could thus provide a three-way bridge between dynamical, structural and informational perspectives on sensorimotor perception.

This is certainly an ambitious program. But our point, here, is to propose a new line of debate and collective inquiry. This effort would be best supported by a dialogue among diverse research communities, from dynamicists and applied category theorists to complex systems scientists, embodied intelligence researchers and enactivists of all backgrounds — not to mention psychologists and neuroscientists, even though the formal framework would need to mature before it can be helpful to experimentalists.

Chapter 4

Exact and Soft Successive Refinement of the Information Bottleneck

4.1 Introduction

4.1.1 Relation to previous chapters

In previous chapters, our approach has been to show that several group-theoretic notions of symmetry — e.g., channel invariances and equivariances, or minimal class-pose parametrisations under a group action — can be cast as edges cases of more general information-theoretic objects. In most cases, these took the form of solutions to IB-like problems that implement a trade-off between compression and preservation of a well-chosen information-theoretic quantity, say I_C. The trade-off is always parametrised by a scalar λ in a continuous range of values $[0, \Lambda]$, where the case $\lambda := \Lambda$ corresponds to the full preservation of I_C, and captures the group-theoretic structure under study. The case $0 \leq \lambda < \Lambda$ then *softens* this group-theoretic structure, and is thus expected to be more relevant to modeling or discovering real-world symmetries — e.g., in the sensorimotor interface of embodied agents. However, it must be acknowledged that until here, we barely scratched the surface of what this soft case actually looks like.

While exploring soft symmetries in realistic scenarios is beyond the scope of this thesis, we will now focus on a specific question relating to the soft case: that of the *relationship* between different bottleneck solutions, for varying λ. In particular, as reducing λ induces further compression, it might seem intuitive that for $\lambda_1 < \lambda_2$, a bottleneck channel κ_1 with parameter λ_1 should be a *post-processing*[1] of a bottleneck channel κ_2 with parameter λ_2: i.e., that there exists another channel $\gamma \in \mathscr{K}(\mathcal{T})$ such that

$$\kappa_1 = \gamma \circ \kappa_2. \tag{4.1.1}$$

For instance, Theorems 2.2.3 and 2.3.1 in Chapter 2 and Theorem 3.6.1 in Chapter 3 show that for resp. the classic IB, the Divergence IB with no shape constraints, and the IB characterising ergodic components, the relation (4.1.1) indeed holds if κ_2 is defined by the maximal trade-off parameter $\lambda := \Lambda$, and for any other bottleneck κ_1. In this chapter, we explore if, and if not "to which extent", relation (4.1.1) holds for the whole range of parameters $0 \leq \lambda_1 < \lambda_2 \leq \Lambda$, in the case of the classic IB problem.

Let me highlight the relevance of this work to the study of generalised symmetries considered in previous chapters. We saw that the invariances $\sigma \in \mathrm{Bij}(\mathcal{X})$ of a channel $p(Y|X)$ are characterised by the property $\kappa_* \circ \sigma = \kappa_*$, for κ_* a classic IB (with source X and relevancy Y) with parameter $\lambda := \Lambda$ requiring full information preservation. This lead us to define, for *any* $\lambda \in [0, \Lambda]$, soft invariances of "granularity" λ — or λ-*invariances* for short — as the transformations $\sigma \in \mathscr{K}(\mathcal{X})$ satisfying $\kappa \circ \sigma = \kappa$ for a bottleneck channel $\kappa \in \mathrm{IB}(\lambda)$ (see

[1] Also called *garbling* in (Rauh et al., 2017).

Section 2.2 in Chapter 2). Now, crucially, if $\lambda_1 < \lambda_2$ and the relation (4.1.1) does hold for $\kappa_1 \in \mathrm{IB}(\lambda_1)$ and $\kappa_2 \in \mathrm{IB}(\lambda_2)$, this yields, for all $\sigma \in \mathscr{K}(\mathcal{X})$,

$$\kappa_2 \circ \sigma = \kappa_2 \quad \Rightarrow \quad \kappa_1 \circ \sigma = \kappa_1,$$

i.e., that *all λ_2-invariances are λ_1-invariances*. If this would hold for any $\lambda_1 < \lambda_2$, we would thus obtain a semigroup of soft invariances that can only grow larger as the granularity λ decreases. Conversely, if the relation (4.1.1) does not hold for some bottlenecks κ_1, κ_2, then we could have invariances that "appear" at some granularities, but then "disappear" when coarsening further. Both these possibilities have interesting interpretations, but quite different ones.

However, the work presented here was originally published in (Charvin et al., 2023a) with a different interpretation of "incorporating new information along processing stages" (see below), which I decided to leave as such, as it is interesting in its own right. The link with (exact and soft) invariances is provided by Proposition 4.4.3 below, which makes clear that the concept studied here —- *successive refinement* — is formally equivalent to the relation (4.1.1) above.

4.1.2 Conceptualisation and Organisation Outline

In this chapter, it will be more convenient to work with the following, equivalent formulation of the Information Bottleneck (IB) problem (Gilad-Bachrach et al., 2003):

$$\underset{\substack{q(T|X)\,:\\ T-X-Y,\ I(X;T)\leq\lambda}}{\arg\max} \quad I(Y;T). \tag{4.1.2}$$

Let us recall that it formalises the problem, for an information-processing system, of extracting relevant information about a target variable Y within a correlated source variable X, under constraints on the cost of the information processing needed to do so—yielding a compressed variable T. The trade-off parameter λ thus controls the bound on the permitted information-processing cost — i.e., intuitively, the resulting coarse-graining's granularity. The Markov chain condition $T - X - Y$ ensures that any information that the bottleneck T extracts about the relevancy variable Y can only come from the source X. Moreover, the solutions to (4.1.2) for varying λ trace the so-called *information curve*, i.e., the λ-parameterised curve

$$\big(I_\lambda(X;T), I_\lambda(Y;T)\big)_{\lambda\geq 0} \subseteq \mathbb{R}^2, \tag{4.1.3}$$

where $I_\lambda(X;T)$ and $I_\lambda(X;T)$ are defined by a bottleneck T of parameter λ (see the black curve in the first figure in Section 4.2 below). This curve indicates the informationally optimal bounds on the feasible trade-offs between relevancy $I(Y;T)$ and complexity $I(X;T)$ of the bottleneck T. In this sense, the IB method provides a fundamental understanding of the *informationally optimal limits* of information-processing systems.

However, one aspect of the IB framework conflicts with a crucial feature of real-world systems: the informationally optimal limits that it describes only consider a given coarse-graining T *taken in isolation* from any other one in the system. This point of view disregards the *relationship between processing stages*, which is crucial in real-world information-processing systems. This leads to the following question: does the relationship between distinct coarse-grainings $T_1, \dots, T_n$ impact their individual information optimality? In this paper, we are mostly interested in a specific kind of relationship: when $T_1, \dots, T_n$ are successively produced in this order, and each new T_i builds on both the previous processing stage T_{i-1} and new information from the fixed source X to extract information about the fixed relevancy Y. This scenario formalises the *incorporation of information along processing stages*—as is the

case in developmental learning, or, more generally, any kind of learning process that goes through identifiable successive steps.

More precisely, consider an informationally bounded agent that extracts information about a relevant variable Y within an environment X. If the agent is informationally optimal, given an affordable complexity cost λ_1, it must maximise the relevant information that it extracts from the environment—resulting in a bottleneck T_1, i.e., a solution to (4.1.2) with parameter λ_1. Then, assume that at, a later stage, the complexity cost that the agent can afford increases to $\lambda_2 > \lambda_1$, while the goal is still to extract information about the same relevant feature Y within the same environment X. To keep being informationally optimal, the agent should thus update the previous bottleneck so it becomes a bottleneck of parameter λ_2. Given this setting, the question we ask is: to which extent can the information content carried by T_1 be leveraged for the production of T_2? It is indeed not intuitively clear that T_2 should keep all the information from T_1. An informal example is the fact that most pedagogical curricula teach knowledge via successive approximations, where, at a more advanced level, the content learned at the beginner level must sometimes be *unlearned* to successively proceed further, even though it was perfectly reasonable—in our language, informationally optimal—to deliver the first beginner sketch to students that would never progress to learn the expert level.

This question has been formalised, in the rate-distortion literature, with the notion of *successive refinement* (SR) (Equitz et al., 1991; Koshelev, 1980; Kostina et al., 2018; Rimoldi, 1994; Tuncel et al., 2003), which, in short, refers to the situation where several-stage processing does not incur any loss of information optimality. More precisely, in the context outlined above, there is successive refinement if the processing cost of first producing a coarse bottleneck T_1 of parameter λ_1 and then refining it to a finer bottleneck T_2 of parameter $\lambda_2 > \lambda_1$ is no larger than the processing cost of directly producing a bottleneck T_2 of parameter λ_2 without any intermediary bottleneck T_1 (see Section 4.2.1 and Appendix D.1.2 for formal definitions). The aim of this work is to push the understanding of successive refinement in the IB framework (Mahvari et al., 2020; Tian et al., 2008; Tuncel, 2009) further, as well as to expand the analysis to a *quantification* of the lack of SR, in cases where the latter does not hold exactly. We start by leveraging general results in existing IB literature (Kline et al., 2022; Kolchinsky et al., 2019) to prove that successive refinement always holds for jointly Gaussian (X, Y), and when Y is a deterministic function of X. However, it is seems crucial, for further progress on more general scenarios, to design specifically tailored mathematical and numerical tools. In this regard, we provide two main contributions.

First, we present a simple geometric characterisation of SR, in terms of convex hulls of the decoder symbol-wise conditional probabilities $q(X|t)$, for t varying in the bottleneck alphabet $\mathcal{T}$. This characterisation is proven in the discrete case under an additional but mild assumption of injectivity of the decoder $q(X|T)$. This new point of view fits well with an ongoing convexity approach to the IB problem (Asoodeh et al., 2020; Benger et al., 2023; Dikshtein et al., 2021; Hsu et al., 2018; Witsenhausen et al., 1975) and might thus help develop a new geometric perspective on the successive refinement of the IB. As an example, we use this geometric characterisation to prove that SR always holds for binary source X and binary relevancy Y. Moreover, this characterisation makes it straightforward to numerically assess, with a linear program checking convex hull inclusions, whether or not two discrete bottlenecks T_1 and T_2 achieve successive refinement. As we demonstrate with minimal numerical examples, this can help in investigating the SR structure of any given IB problem, i.e., how successive refinement depends on the particular combination of trade-off parameters λ_1 and λ_2.

Second, we soften (Catenacci Volpi et al., 2020) the traditional notion of successive refinement and study the *extent to which* several-stage processing incurs a loss of information optimality. More precisely, we propose to measure soft successive refinement with the *unique information* (Bertschinger et al., 2013) (UI) that the coarser bottleneck T_1 holds about the source X, as compared to the finer one T_2. Explicitly, this UI is defined as the minimal

value of $I_q(X;T_1|T_2)$ over all distributions $q := q(X,T_1,T_2)$ whose marginals $q(X,T_1)$ and $q(X,T_2)$ coincide with the corresponding bottleneck distributions (see Section 4.3.1 for details). As a first exploration of soft SR's qualitative features, we investigate the landscapes of unique information over trade-off parameters, for again some simple example distributions $p(X,Y)$. These landscapes seem to unveil a rich structure, which was largely hidden by the traditional notion of SR, that only distinguished between SR being present or absent. Among the general features suggested by these experiments, the most significant are that (*i*) soft SR seems strongly influenced by the trajectories of the decoders $q_\lambda(X|T)$ over λ; (*ii*) the UI often goes through sharp variations at the bifurcations (Ngampruetikorn et al., 2021; Parker et al., 2022; Wu et al., 2020; Zaslavsky et al., 2019) undergone by the bottlenecks (in a fashion compatible with the presence of discontinuities of either the UI itself or its differential w.r.t. to trade-off parameters); and (*iii*) the loss of information optimality seems always small—more precisely, the global bound on the UI was observed to be typically one or two orders of magnitude lower than the system's globally processed information (see Section 4.3.2 for formal statements). These three conclusions are phenomenological and limited to our minimal examples, but they shed light on the kind of structure that can be investigated by further research. They also suggest the relevance that developing this theoretical framework might have for question of incoporating information along processing stages. In particular, the link with IB bifurcations and the overall small loss of information optimality would, if generalisable, have interesting consequences for the structure and efficiency of incremental learning. We also draw a formal equivalences of SR with specific decision problems (Bertschinger et al., 2013, 2014), which shows the link that successive refinement has with the study of soft invariances as studied in Chapter 2.

In the next Section 4.1.3, we review related work. After having established notations and recalled some general notions in Section 4.1.4, we formally introduce the notion of the successive refinement of the IB in Section 4.2.1, where we also prove successive refinability in the case of Gaussian vectors and deterministic channel $p(Y|X)$. We then present the convex hull characterisation in Section 4.2.2, before using it to prove successive refinement for the case of binary source and relevancy variables. The following Section 4.2.3 leverages the convex hull characterisation to gather some first insights from minimal experiments. These experiments suggest an intuition for defining soft successive refinement, which we formalise in Section 4.3.1 through a measure of unique information (Bertschinger et al., 2013), where we provide theoretical motivations for our choice. This new measure is explored in Section 4.3.2 with additional numerical experiments that highlight the general features described above. The alternative interpretation of SR in terms of decision problems is developed in Sections 4.4. We then describe the limitations and potential future work in Section 4.5, and conclude in Section 4.6.

4.1.3 Related Work

The notion of successive refinement has been long studied in the rate-distortion literature (Equitz et al., 1991; Koshelev, 1980; Kostina et al., 2018; Rimoldi, 1994; Tuncel et al., 2003). However, classic rate-distortion theory (Cover et al., 2009) usually considers distortion functions defined on the random variables' *alphabets*, whereas the IB framework can be regarded as a rate-distortion problem only if one allows the distortion to be defined on the space of probability *distributions* (Zaidi et al., 2020). Successive refinement thus needed to be adapted to the IB framework, which was achieved starting from various perspectives.

In (Tian et al., 2008; Tuncel, 2009), successive refinement is formulated within the IB framework. Then, (Mahvari et al., 2020) goes further by considering the informationally optimal limits of several-stage processing in general, without comparing it to single-stage processing. In both these works, the problem is initially defined in asymptotic coding terms,

and only then given a single-letter characterisation. On the contrary, we will directly define successive refinement from a single-letter perspective. It turns out that our single-letter definition and the operational multi-letter definition from (Tian et al., 2008; Tuncel, 2009) are equivalent. The two latter works—as well as (Mahvari et al., 2020)—thus provide our single-letter definition with an operational interpretation that also formalises the intuition of an informationally optimal incorporation of information (see Proposition 4.2.2 and Appendix D.1.2).

The link between successive refinement and the IB theory of deep learning (Achille et al., 2018a; Elad et al., 2019; Kawaguchi et al., 2023; Lorenzen et al., 2022; Saxe et al., 2019; Shwartz-Ziv et al., 2017, 2019; Tishby et al., 2015) has been noted since the inception of the latter research agenda (Tishby et al., 2015). It has been developed further in (Yang et al., 2017) and (Yousfi et al., 2020). The latter works, in particular (Yang et al., 2017), are formally very similar to the SR problem considered here. For the sake of conciseness, we will not present these links further here, but they are described in detail in the published version of the work presented in this chapter (Charvin et al., 2023a).

Note that while the phenomenon of IB bifurcations has been studied from a variety of perspectives (see, e.g., (Ngampruetikorn et al., 2021; Parker et al., 2022; Wu et al., 2020; Zaslavsky et al., 2019)), here, we adopt that of (Zaslavsky et al., 2019), which frames IB bifurcations as parameter values where the minimal number of symbols required to represent a bottleneck increases.

In (No, 2019), successive refinability is proved for discrete source X and relevancy $Y = X$. Our Proposition 4.2.5 generalises this result to either discrete or continuous source X, with relevancy Y being an arbitrary function of X, with a similar argument as that in (No, 2019).

In (Kline et al., 2022), links between the IB framework and renormalisation group theory are exhibited. Even though the questions addressed in the latter work are thus distinct from those addressed here, the Gaussian IB's *semigroup structure* defined and proven in (Kline et al., 2022) implies the successive refinability of Gaussian vectors (see Proposition 4.2.4 and Appendix 4.2.4).

Our convex hull characterisation of SR is complementary with the convexity approach to the IB (Asoodeh et al., 2020; Benger et al., 2023; Dikshtein et al., 2021; Hsu et al., 2018; Witsenhausen et al., 1975): while the latter references do not consider SR, the proof of Proposition 4.2.7 shows that they are useful to leverage our convex hull characterisation.

The loss of information optimality induced by several-stage processing has already been studied in (Lastras et al., 2001) (see next paragraph), but a quantification of it based on *soft Markovianity* was, to the best of our knowledge, only considered in (Catenacci Volpi et al., 2020). Here, we take inspiration in the latter work to quantify soft successive refinement, but we explicitly address the problem that joint distributions over distinct bottlenecks are not uniquely defined. This leads us to use the *unique information* defined in (Bertschinger et al., 2013) within the context of partial information decomposition (Griffith et al., 2014; Harder et al., 2013; Williams et al., 2010) as our measure of soft SR. This unique information has tight links with the Blackwell order (Bertschinger et al., 2014; Blackwell, 1953), which allows us in Section 4.4 to provide a second alternative interpretation of (exact and soft) successive refinement in terms of decision problems.

Ref. (Lastras et al., 2001) proves the near-successive refinability of rate-distortion problems when the distortion measure is the squared error. However, the latter work's approach is different from ours in two respects. First, the distortion measures are different: in particular, as mentioned above, the IB distortion is defined over the space of probability distributions on symbols, unlike the squared error, which is defined on the space of symbols itself. Second, (Lastras et al., 2001) quantifies the lack of SR as the respective differences between sequences of optimal rates (for given distortion sequences) of a several-stage processing system and the

corresponding optimal rates (for the same distortions) of a single-stage processing system. Here, we quantify the lack of SR with a single quantity: the unique information defined by bottlenecks with different granularities. We are, at this stage, not aware of a link between this value of unique information and differences in one-stage and several-stage optimal rates.

4.1.4 Technical Preliminaries

In this section, we fix the notations and conventions that we will use along the paper and recall some general notions that we will need.

Notations and Conventions

The random variables are denoted by capital letters, e.g., X, their alphabets by calligraphic ones, e.g., $\mathcal{X}$, and their symbols by lower-case letters, e.g., x. Sometimes, we will mix upper- and lower-case notations to denote a family where some symbols vary, while others are fixed, e.g., $q(X|t) := \big(q(x|t)\big)_{x\in\mathcal{X}}$, or $q(x|T) := \big(q(x|t)\big)_{t\in\mathcal{T}}$. Throughout the whole paper, X is the fixed source and Y the fixed relevancy of the IB problem. The variable T defined by the solution $q(T|X)$ to the primal IB problem (4.1.2) is called a *primal* bottleneck. We use the same symbol T for *Lagrangian* bottlenecks, i.e., variables defined by solutions $q(T|X)$ to the Lagrangian bottleneck problem (see Equation (4.1.4) below). By "bottleneck" without further specification, we refer to either a primal or Lagrangian bottleneck. The fixed source-relevancy distribution is denoted $p(X,Y)$, and any distribution involving at least one bottleneck is denoted with the letter q, e.g., $q(X,Y,T)$. When it is necessary to make the trade-off parameter explicit, we index the corresponding objects by λ, e.g., $q_\lambda(T|X)$ or $I_\lambda(Y;T)$. Unless explicitly stated otherwise, the source X, relevancy Y, and any considered bottleneck T are defined as either all discrete or all continuous. Probability simplices, and sometimes some of their subsets are written using the generic symbol Δ; for instance, the source simplex is denoted by $\Delta_{\mathcal{X}}$.

Without loss of generality, we always restrict X, Y, and the bottleneck T to their respective supports so that, in particular, all the conditional distributions are unambiguously well-defined.

We will denote by I_Y the function from $\mathbb{R}_+$ to $\mathbb{R}_+$ defined by $I_Y(\lambda) := I(Y;T)$, where T is a solution to the primal IB problem (4.1.2) for the parameter λ. The *information curve*, defined above in Equation (4.1.3), is thus also the graph of the function I_Y.

General Facts and Notions

The following properties of the IB framework will be useful (Asoodeh et al., 2020; Witsenhausen et al., 1975):

- A bottleneck must saturate the information constraint, i.e., solutions T to (4.1.2) must satisfy $I_\lambda(X;T) = \lambda$. In other words, the primal trade-off parameter is the complexity cost of the corresponding bottleneck.
- The function $I_Y : \lambda \mapsto I_\lambda(T;Y)$ is constant at least for $\lambda \geq H(X)$. We will thus always assume, without loss of generality, that $\lambda \in [0, H(X)]$.
- In the discrete case, choosing a bottleneck cardinality $|\mathcal{T}| = |\mathcal{X}| + 1$ is enough to obtain optimal solutions. Thus, we always assume, without loss of generality, that $|\mathcal{T}| \leq |\mathcal{X}| + 1$, where $|\mathcal{T}| < |\mathcal{X}| + 1$ might occur if needed to make T full support .

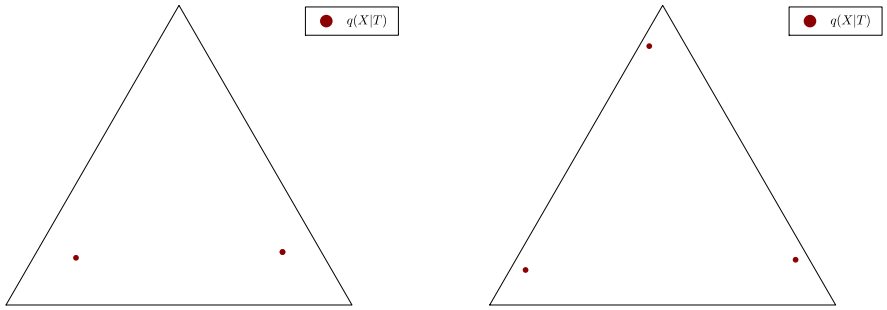


FIGURE 4.1: Examples of distributions $q(X|T)$, visualised as families of points $\{q(X|t),\, t \in \mathcal{T}\}$ on the source simplex $\Delta_{\mathcal{X}}$ with $|\mathcal{X}| = 3$. Each of the triangle's vertices represents the Dirac probability on some $x \in \mathcal{X}$. The bottleneck's effective cardinality is $k = 2$ on the left and $k = 3$ on the right.

To compute bottleneck solutions, instead of directly solving the primal problem (4.1.2), following common practice, we will solve its Lagrangian relaxation (Lemaréchal, 2001):

$$\underset{q(T|X)\,:\,T-X-Y,}{\arg\min} \quad I(X;T) \;-\; \beta\, I(Y;T), \tag{4.1.4}$$

where the complexity-relevancy trade-off is now parameterised by $\beta \geq 0$, which corresponds to the inverse of the information curve's slope (Parker et al., 2022). As the information curve is known to be concave, the Lagrangian parameter β is an increasing function of the primal parameter $\lambda = I(X;T)$. Moreover, we can, without loss of generality, assume that $\beta \geq 1$ (Zaslavsky et al., 2019). (Note that when the information curve is not strictly concave, the Lagrangian formulation does not allow one to obtain all the solutions to the primal problem (Benger et al., 2023; Kolchinsky et al., 2017). However, in our simple numerical experiments, we always obtained strictly concave information curves.)

We also recall that the *effective cardinality* of a bottleneck T defined by a compression channel $q(T|X)$, and of parameter λ, is the minimum bottleneck cardinality obtained from a post-processing of $q(T|X)$ that still produces a bottleneck for the same parameter λ. It will be denoted here by $k(T)$, and we saw in Chapter 2, Appendix B.4.3 that it coincides with the cardinality of the partition of $\mathrm{supp}(q(T))$ defined by the equivalence relation $t \sim t' \Leftrightarrow q(X|t) = q(X|t')$. Similarly, the cardinality of a Lagrangian bottleneck is defined as the number of symbols t with distinct $q(X|t)$. A discrete (primal or Lagrangian) bottleneck T will be said to be a *canonical bottleneck*, or in *canonical form*, if the cardinality of the space $\mathcal{T}$ on which it is defined coincides with the bottleneck's effective cardinality.

For Lagrangian bottlenecks, our definition of effective cardinality is slightly different from but equivalent to that from (Zaslavsky et al., 2019) (see Appendix A.1 in the published version of this paper (Charvin et al., 2023a) for an explicit proof of this statement). It is straightforward that every (primal or Lagrangian) bottleneck can be reduced to its canonical form by merging the symbols with identical $q(X|t)$. We will be particularly interested in the *change* of effective cardinality, which has been identified in (Zaslavsky et al., 2019) as characterising the bottleneck phase-transitions, or *bifurcations*.

In Figure 4.1, we present examples of bottleneck conditional distributions $q(X|T)$, visualised as the family of points $\{q(X|t),\, t \in \mathcal{T}\}$ on the source simplex $\Delta_{\mathcal{X}}$, where, here, $|\mathcal{X}| = 3$, and the bottleneck is computed with $|\mathcal{T}| = 3$ in both examples. However, in Figure 4.1 (left), there are only two distinct $q(X|t)$, so there must be two equal pointwise probabilities $q(X|t_1)$ and $q(X|t_2)$; thus, $k = 2$ and the canonical form of T is obtained by merging t_1 and t_2. On the contrary, in Figure 4.1 (right), there are three distinct $q(X|t)$, so, here, $k = 3$ and the bottleneck is already in canonical form.

Eventually, the notions of *consistency* and *extension* will be crucial.

Definition 4.1.1. Let $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ be a Cartesian product of (continuous or discrete) alphabets. For $C = \{c_1, \dots, c_r\} \subseteq \{1, \dots, m\}$ a subset of coordinates, we write

$$\bigtimes_{c \in C} \mathcal{A}_c := \mathcal{A}_{c_1} \times \cdots \times \mathcal{A}_{c_r}.$$

For each $1 \leq i \leq n$, we consider a subset of coordinates C_i and a probability distribution q_i over $\bigtimes_{c \in C_i} \mathcal{A}_c$. The distributions $q_1, \dots, q_n$ are said to be *consistent* if, for every $1 \leq i, j \leq n$, the respective marginals of q_i and q_j on their common coordinates $\bigtimes_{c \in C_i \cap C_j} \mathcal{A}_c$ are equal.

For instance, if T_1 and T_2 are two bottlenecks, they define consistent distributions $q_1(X, Y, T_1)$ and $q_2(X, Y, T_2)$ because, by definition, their respective marginals on their common coordinates $\mathcal{X} \times \mathcal{Y}$ are $q_1(X, Y) = q_2(X, Y) = p(X, Y)$.

Definition 4.1.2. Let $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ be a Cartesian product of (continuous or discrete) alphabets, and $q_1, \dots, q_n$ be consistent probability distributions over distinct but potentially overlapping coordinates of $\mathcal{A}$. A distribution q over the whole $\mathcal{A}$ is called an *extension* of the family of distributions $\{q_1, \dots, q_n\}$ if it is consistent with each q_i.

Consider bottlenecks $T_1, \dots, T_n$ of same source X and relevancy Y for resp. parameters $\lambda_1, \dots, \lambda_n$. They define a consistent family of distributions $\{q_{\lambda_i}(X, T_i),\ 1 \leq i \leq n\}$. One of the central mathematical objects of this work is the set of their extensions into *joint* distributions $q(X, T_1, \dots, T_n)$:

Notation. For given bottlenecks $T_1, \dots, T_n$ of respective parameters $\lambda_1, \dots, \lambda_n$, we denote by $\Delta_{\lambda_1, \dots, \lambda_n}$ the set of extensions $q(X, T_1, \dots, T_n)$ of the family of distributions $\{q_{\lambda_i}(X, T_i),\ 1 \leq i \leq n\}$.

In general, for a fixed family of bottlenecks, there is a multitude of possible ways to extend them into a joint distribution; indeed, $\Delta_{\lambda_1, \dots, \lambda_n}$ traces a polytope on the simplex $\Delta_{\mathcal{X} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_n}$ of joint distributions (see Appendix A in (Bertschinger et al., 2013)). This feature is the formal version of our previous statement that the IB framework does not entirely specify the *relationship* between bottlenecks $T_1, \dots, T_n$: it only constrains it through the set $\Delta_{\lambda_1, \dots, \lambda_n}$. Questions about possible relationships between information bottlenecks are thus questions about properties of the set $\Delta_{\lambda_1, \dots, \lambda_n}$.

4.2 Exact Successive Refinement of the IB

4.2.1 Formal Framework and First Results

Here, we formally describe, within the IB framework, the rate-distortion-theoretic notion of *successive refinement* (SR) (Equitz et al., 1991; Koshelev, 1980; Kostina et al., 2018; Rimoldi, 1994). We propose a purely single-letter definition (i.e., we only consider single source, relevancy, and bottleneck variables), which makes the presentation simpler but still conveys the intuition of information incorporation. After having presented the notion of SR in the IB framework, we describe its Markov chain characterisation (see Proposition 4.2.2), which mirrors the characterisation of SR for classic rate-distortion problems (Equitz et al., 1991), and makes our formulation equivalent to previous multi-letter operational definitions, which also formalise the intuition of information incorporation (Mahvari et al., 2020; Tian et al., 2008; Tuncel, 2009). We then leverage this characterisation to prove SR in the case of Gaussian vectors and deterministic channel $p(Y|X)$.

Intuitively, there is successive refinement when a finer bottleneck T_2 does not discard any of the information extracted by a coarser bottleneck T_1. This can be imposed by requiring

that $T_2 = (T_1, S_2)$ for some variable S_2, which encodes the "supplement" of information that "refines" T_1 into T_2. In the general case:

Definition 4.2.1. Let $0 < \lambda_1 < \cdots < \lambda_n$, and a discrete or continuous $p(X, Y)$ be given. There is *successive refinement* (SR) for parameters $(\lambda_1, \dots, \lambda_n)$ if there exist variables $(T_1, S_2, S_3, \dots, S_n)$ such that

- T_1 is a bottleneck with parameter λ_1;
- For every $2 \leq i \leq n$, the variable $T_i := (T_{i-1}, S_i)$ is a bottleneck with parameter λ_i.

Note that even though it does not appear explicitly in this definition, the relevancy variable Y is indeed crucial to it, as it defines what a bottleneck is (see Equation (4.1.2)). If the conditions of Definition 4.2.1 hold, we will also say that the IB problem defined by $p(X, Y)$ is $(\lambda_1, \dots, \lambda_n)$-refinable. If bottlenecks $T_1, \dots, T_n$ satisfy the definition's conditions, we will say that they *achieve* successive refinement, or, simply, that there is successive refinement between these bottlenecks. If there is successive refinement for all combinations $0 < \lambda_1 < \cdots < \lambda_n$ of trade-off parameters, we will say that the corresponding IB problem is successively refinable. Eventually, when it will be needed in later sections to contrast this notion with that of soft successive refinement, we will refer to it as *exact* successive refinement.

For instance, let $0 < \lambda_1 < \lambda_2$ and $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. We consider $Y := X \oplus Z$, where $\oplus$ denotes the modulo-2 addition, and X and Z are Bernouilli variables with parameters $\frac{1}{2}$ and a, respectively,

It is helpful to visualise SR on the information plane, i.e., that on which lies the information curve. Indeed, successive refinement can be understood in terms of specific translations on the information plane: those resulting from concatenating an already existing variable T_{i-1} with a new variable S_i—let us call them "accumulative translations" because they result from a processing that does not discard any of the information already collected. Let us focus on the case $n = 2$ and first note that, whether or not (T_1, S_2) is a bottleneck, we have

$$I(X; T_1, S_2) = I(X; T_1) + I(X; S_2|T_1),$$

and, similarly,

$$I(Y; T_1, S_2) = I(Y; T_1) + I(Y; S_2|T_1).$$

In other words, the measure of both the complexity cost and relevance for (T_1, S_2) can be decomposed into the same measures first for T_1 and then for the "supplement" of information S_2, conditionally on the "already collected" information T_1. In Figure 4.2 (left and right), we first fix a coarse bottleneck T_1, understood here as a point $\big(I(X; T_1), I(Y; T_1)\big)$ on the information curve. Once T_1 is known, we supplement it with a new variable S_2, which incurs both an additional complexity cost $I(X; S_2|T_1)$ and an additional relevant information gain $I(Y; S_2|T_1)$. The question of successive refinement is that of whether the additional complexity cost can be leveraged enough for the resulting relevant information gain to take (T_1, S_2) "up to the information curve", i.e., to be such that $\big(I(X; T_1, S_2), I(Y; T_1, S_2)\big)$ is on the information curve. This is the case in Figure 4.2, right, and not the case in Figure 4.2, left. In short, there is successive refinement between two points on the information curve if and only if there exists an "accumulative translation" from the coarser one to the finer one.

Let us now describe a more formal characterisation, where point (ii) will mirror the characterisation of SR for classic rate-distortion problems (Equitz et al., 1991).

Proposition 4.2.2. *Let $0 < \lambda_1 < \cdots < \lambda_n$. The following are equivalent:*

(i) *There is successive refinement for parameters $(\lambda_1, \dots, \lambda_n)$;*

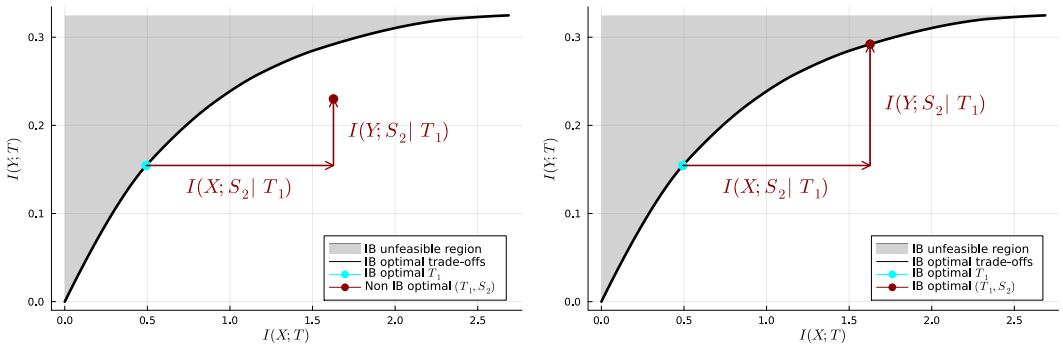


FIGURE 4.2: SR visualised on the information plane. Adding the information from the supplement variable S_2 can be enough (right) or not (left) to achieve SR (values of $I(X;S_2|T_1)$ and $I(Y;S_2|T_1)$ chosen for illustrative purposes).

(*ii*) *There exist bottlenecks $T_1,\dots,T_n$, of common source X and relevancy Y, with respective parameters $\lambda_1,\dots,\lambda_n$, and an extension $q(X,T_1,\dots,T_n)$ of the $q_i := q_i(X,T_i)$, such that, under q, we have the Markov chain*

$$X - T_n - \cdots - T_1. \tag{4.2.1}$$

(*iii*) *There exist bottlenecks $T_1,\dots,T_n$, of common source X and relevancy Y, with respective parameters $\lambda_1,\dots,\lambda_n$, and an extension $q(Y,X,T_1,\dots,T_n)$ of the $q_i := q_i(Y,X,T_i)$, such that, under q, we have the Markov chain*

$$Y - X - T_n - \cdots - T_1. \tag{4.2.2}$$

Proof. See Appendix D.1.1. It is relatively straightforward because we started directly from a single-letter definition. □

Proposition 4.2.2 was already known to be a characterisation of SR of the IB (Mahvari et al., 2020; Tian et al., 2008; Tuncel, 2009). However, as the latter references start from an operational problem in terms of asymptotic rates and distortions for multi-letter systems, here, Proposition 4.2.2 shows that our single-letter Definition 4.2.1 is equivalent to the operational definitions in (Mahvari et al., 2020; Tian et al., 2008; Tuncel, 2009). See Appendix D.1.2 for more details.

Remark 4.2.3. Crucially, the order of the indexing in (4.2.1) and (4.2.2) depends only on the order of the trade-off parameters $\lambda_1 < \cdots < \lambda_n$, and not on the order in which the bottlenecks T_i are produced, which is just the interpretation we started from. In particular, Proposition 4.2.2 makes equally legitimate the interpretation of bottlenecks produced from the finest one to the coarsest one, each new bottleneck thus implementing a further coarsening of the source X. This alternative interpretation makes successive refinement relevant to the Blackwell order (see Section 4.4). For ease of presentation, though, we will stick to the information incorporation interpretation along most of the paper.

Moreover, from Proposition 4.2.2, we can leverage existing IB literature to prove the successive refinability of two specific settings. (For an explicit definition of SR for the *Lagrangian* IB problem, see Appendix D.1.3.)

Proposition 4.2.4. *If X,Y are jointly Gaussian vectors, then the Lagrangian IB problem defined by $p(X,Y)$ is $(\lambda_1,\dots,\lambda_n)$-refinable for all $\lambda_1 < \cdots < \lambda_n$.*

This result is a direct consequence of a property proven in (Kline et al., 2022): in short, in the Gaussian case, iterating the operation of coarse graining a variable by computing a

bottleneck—where, at each iteration, the previous bottleneck becomes the source of the next IB problem—still outputs bottlenecks for the original problem.

Eventually, in the case of deterministic channel $p(Y|X)$, the question of successive refinement can be addressed using a known explicit solution to the IB problem (Kolchinsky et al., 2019):

Proposition 4.2.5. *Let X be a discrete or continuous variable, and Y be a deterministic function of X. Then, the IB problem defined by $p(X,Y)$ is successively refinable for all trade-off parameters $\lambda_1 < \cdots < \lambda_n$.*

Proof. See Appendix D.1.4. A proof was already proposed, from an asymptotic coding perspective, for discrete X and $Y = X$, in (No, 2019). We use a similar argument here. □

However, the solution used here to prove successive refinement is, as noted in (Kolchinsky et al., 2019), not very interesting: it is nothing more than an increasingly noisy version of Y. Proposition 4.2.5 will in any case be useful to set aside the deterministic case in the proof of SR for binary X and Y (see Proposition 4.2.7 below).

Until now, we used existing results from the IB literature that, even though not originally aimed at it, happen to yield interesting consequences for the problem of the successive refinement of the IB. However, it seems crucial, for further progress on the latter topic, to design specifically tailored mathematical and numerical tools. This is the purpose of the following sections of this paper; in particular, in the next section, we present a simple geometric characterisation of the IB's successive refinability.

4.2.2 The Convex Hull Characterisation and the Case $|\mathcal{X}| = |\mathcal{Y}| = 2$

In this section, we present our convex hull characterisation of successive refinement. We then show its relevance both to numerical computations—thanks to a linear program for checking the condition—and to proving new mathematical results—which we exemplify by proving, thanks to this new characterisation, the successive refinability of binary variables. Here, as in our subsequent numerical experiments in Section 4.2.3, we will focus on discrete variables and $n = 2$ processing stages, even though our results are thought of as a first step towards a generalisation to continuous variables and an arbitrary number of processing stages.

The convexity approach that we propose hinges upon changing the perspective on the IB problem (4.1.2) from an optimisation over the "encoder" channels $q(T|X)$ to an optimisation over the *"decoder"* channels $q(X|T)$; indeed, (4.1.2) can be equivalently presented as the "reversed" optimisation problem

$$\underset{\substack{(q(T),q(X|T)):\\ \sum_t q(t)q(X|t)=p(X)\\ T-X-Y,\ I(X;T)\leq\lambda}}{\arg\max} \quad I(Y;T). \tag{4.2.3}$$

Formulations (4.1.2) and (4.2.3) yield the same solutions because, through the Markov chain $T - X - Y$, the joint distribution $q(X,Y,T)$ is equivalently determined by specifying some $q(T|X)$ or specifying some pair $\big(q(T), q(X|T)\big)$ that satisfies the consistency condition

$$\sum_t q(t)q(X|t) = p(X).$$

Moreover, this formulation leads to a crucial intuition concerning the relationship between successive refinement and the set $\mathcal{H}_T := \mathsf{Hull}\{q(X|t), t \in \mathcal{T}\}$, where, for a set $E \subseteq \mathbb{R}^n$, we denote by $\mathsf{Hull}(E)$ the convex hull of E, i.e., the set of points obtained as convex combinations of points in E. First, note that, for a bottleneck T, the set $\mathcal{H}_T$ is reduced to a single point if and

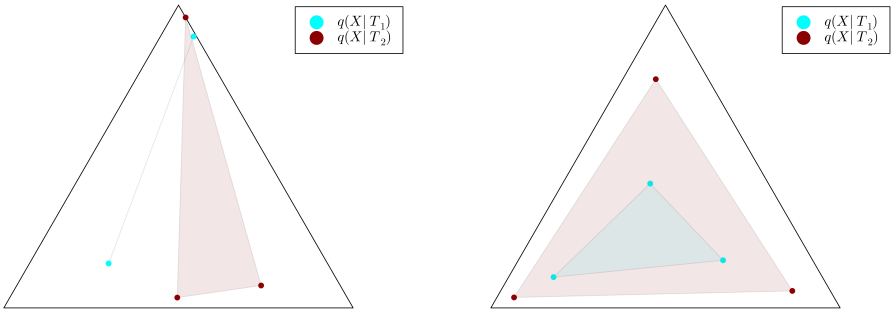


FIGURE 4.3: Illustration of the convex hull condition (satisfied on the right but not on the left). Red triangle: $\mathsf{Hull}\{q(X|t_2), t_2 \in \mathcal{T}_2\}$, with T_2 the finer bottleneck. Cyan triangle/segment: similarly for the coarser bottleneck T_1.

only if T is independent from the source X. Conversely, $\mathcal{H}_T$ coincides with the whole source simplex $\Delta_{\mathcal{X}}$ if and only if T captures all the information from the source, i.e., if $I(X;T) = H(X)$. These edge cases suggests the intuition that $\mathcal{H}_T$ describes the *information content* held by the bottleneck T about the source X. Now, let us recall that SR from a coarse bottleneck T_1 to a finer bottleneck T_2 means intuitively that T_2 can be obtained without discarding any of the information extracted by T_1 about the source X; in other words, that the information content of T_1 about the source X *is included* in that of T_2. Combining this latter intuition with the previous one suggests the following characterisation of SR:

$$\mathsf{Hull}\{q(X|t_1), t_1 \in \mathcal{T}_1\} \subseteq \mathsf{Hull}\{q(X|t_2), t_2 \in \mathcal{T}_2\}, \tag{4.2.4}$$

where T_1 and T_2 are bottlenecks of parameters $\lambda_1 < \lambda_2$, respectively. This condition is visualised in Figure 4.3. The characterisation indeed holds, at least under a mild assumption of injectivity of the finer bottleneck's decoder:

Proposition 4.2.6. *Let $0 < \lambda_1 < \lambda_2$, and assume that $p(X,Y)$ is discrete.*

If there is successive refinement for parameters (λ_1, λ_2), then there exist bottlenecks T_1, T_2 of parameters λ_1, λ_2, respectively, such that the convex hull condition (4.2.4) is satisfied.

Conversely, if there exist bottlenecks T_1, T_2 of parameters λ_1, λ_2, respectively, such that the convex hull condition (4.2.4) holds and such that the decoder $q(X|T_2)$, seen as a probability transition matrix, is injective, then there is successive refinement for parameters (λ_1, λ_2). Moreover in this latter case, if T_1, T_2 are bottlenecks that achieve successive refinement, the extension $\tilde{q}(X, T_1, T_2)$ of $q(X, T_1)$ and $q(X, T_2)$ such that $X - T_2 - T_1$ holds is uniquely defined.

Proof. See Appendix D.1.5. The idea consists in translating the Markov chain characterisation $X - T_2 - T_1$ into the convex hull condition (4.2.4). The direct sense is straightforward. For the converse direction, observe that, even though as soon as (4.2.4) is satisfied it provides a joint distribution $\tilde{q}(X, T_1, T_2)$ that satisfies the Markov chain $X - T_2 - T_1$, it is not clear whether this distribution is consistent with $q(X, T_1)$. The potential problem stems from the fact that $\tilde{q}$ must be such that the channel $\tilde{q}(T_2|T_1)$ maps the marginal $q(T_1)$ to the marginal $q(T_2)$. The injectivity assumption, however, provides a sufficient condition for it to be the case. This assumption happens to also imply the uniqueness of the extension. □

Even though the injectivity assumption might seem restrictive, in practice, in our numerical experiments below (see Sections 4.2.3 and 4.3.2), we always found that the decoder channel $q(X|T_2)$ could be chosen as injective by reducing it to its effective cardinality (see Section 4.1.4)—a process that leaves the convex hull condition (4.2.4) unchanged because it leaves the points $q(X|t_2)$ unchanged.

Our new characterisation provides a simple way of checking whether or not two bottlenecks T_1 and T_2 achieve SR. Recall that the Markov chain characterisation (Proposition 4.2.2, point (ii)) shows that SR is a feature of the space Δ_{q_1,q_2} of all extensions $q(X,T_1,T_2)$ of individual bottleneck distributions $q_1(X,T_1)$ and $q_2(X,T_2)$. While this set might, *a priori*, be difficult to study directly, our characterisation (4.2.4) reduces the problem to a simple geometric property relating only two explicitly given conditional distributions: $q(X|T_1)$ and $q(X|T_2)$. Moreover, note that (4.2.4) is equivalent to

$$\forall t_1 \in \mathcal{T}_1, \quad q(X|t_1) \in \mathsf{Hull}\{q(X|t_2), t_2 \in \mathcal{T}_2\},$$

and that checking whether a point is in the convex hull of a finite set of other points can be cast as a linear programming problem (Matousek et al., 2007). This program is described in Appendix D.1.6 and will be used in Section 4.2.3.

We deem this convex hull characterisation to be important for theory as well. Indeed, it reduces the question of successive refinement to a question about the structure of the trajectories, on the source probability simplex $\Delta_{\mathcal{X}}$, of the points $q_\lambda(X|t)$ for varying λ. Thus, any theoretical progress on the description of these bottleneck trajectories might lead to theoretical progress on the side of successive refinement. As a first step in this direction, we show that this geometric point of view, in combination with the convexity approach to the IB (Asoodeh et al., 2020; Benger et al., 2023; Dikshtein et al., 2021; Hsu et al., 2018; Witsenhausen et al., 1975), solves the question of SR in the case of a binary source and relevancy.

Proposition 4.2.7. *If $|\mathcal{X}| = |\mathcal{Y}| = 2$, then, for any discrete distribution $p(X,Y)$ and any trade-off parameters $\lambda_1 < \lambda_2$, the IB problem defined by $p(X,Y)$ is (λ_1,λ_2)-successively refinable.*

Proof. Let us here outline the proof presented in Appendix D.1.7. The case of deterministic $p(Y|X)$ was already dealt with in Proposition 4.2.5, so we can assume that $p(Y|X)$ is not deterministic. In this case, the IB problem with $|\mathcal{X}| = |\mathcal{Y}| = 2$ and $n = 2$ has been extensively studied in (Witsenhausen et al., 1975). In short, the latter approach leverages the fact that a pair $(q(T), q(X|T))$ is a solution to the IB problem (4.2.3) if the convex combination of the points $F_\beta(q(X|t))$, with weights given by $q(T)$, achieves the lower convex envelope of the function F_β, where F_β is a well-chosen function on the source simplex $\Delta_{\mathcal{X}}$ and β is the information curve's inverse slope. This work, along with considerations from (Asoodeh et al., 2020), which uses the same convexity approach, yields in particular that (i) the points $q_\beta(X|t)$ are the extreme points of a non-empty open segment uniquely defined by β, and (ii) this latter segments grows as a function of the inverse slope β and thus, by concavity, as a function of λ. This implies that the convex hull condition is always satisfied for $\lambda_1 < \lambda_2$. As point (i) also implies that, here, $q_{\lambda_2}(X|T_2)$ must be injective, Theorem 4.2.6 allows us to conclude the successive refinability for $n = 2$ processing stages. □

4.2.3 Numerical Results on Synthetic Examples

In this section, we leverage our new convex hull characterisation to investigate successive refinement on simple numerical examples, i.e., with discrete and low-cardinality distributions $p(X,Y)$. Our experiments suggest that, in general, successive refinement does not always hold exactly. However, they also highlight two other features: first, it seems that successive refinement is often shaped by IB bifurcations (Ngampruetikorn et al., 2021; Parker et al., 2022; Wu et al., 2020; Zaslavsky et al., 2019). Second, even though successive refinement is often not satisfied exactly, visualisations suggest that it is often "close" to being satisfied. The formalisation of this latter intuition will be the topic of the next section.

We consider the Lagrangian form (4.1.4) of the IB problem (see Section 4.1.4). We compute solutions to it with the Blahut–Arimoto (BA) algorithm (Tishby et al., 2000), combined

with reverse deterministic annealing (Rose, 1998; Zaslavsky et al., 2022), starting from $\beta \approx \infty$ (i.e., in practice, $\beta \gg 1$) at the IB solution $T = X$ (we noticed that regular deterministic annealing sometimes yielded sub-optimal solutions because they followed sub-optimal branches at IB bifurcations (Gedeon et al., 2012; Tishby et al., 2000), which was not the case for reverse annealing). We always obtained that $I(X;T)$ was a strictly increasing function of the Lagrangian parameter β, so it makes sense to index the solutions by $\lambda = I(X;T)$ rather than β; for instance, in this section and Section 4.3.2, we will write $q_\lambda(T|X)$ for our algorithm's output for a β such that $I(X;T) = \lambda$.

In all our numerical experiments, after reducing a bottleneck T to its canonical form (see Section 4.1.4), the decoder channel $q_\lambda(X|T)$ was injective. Therefore, thanks to Theorem 4.2.6, the convex hull condition (4.2.4) being satisfied here does imply successive refinement. We thus use the convex hull condition as a proxy for numerically assessing successive refinement.[2] This condition can be investigated in two ways. First, for two distinct trade-off parameters $\lambda_1 < \lambda_2$, we can compute whether the convex hull condition (4.2.4) holds or not with the linear program described in Appendix D.1.6. Second, for $|\mathcal{X}| \leq 3$, we can visualise the whole trajectories, for varying λ, of the points $q_\lambda(X|t)$ on the source simplex $\Delta_{\mathcal{X}}$. As we will see, this yields interesting qualitative insights.

As a sanity check for our algorithm, we compute bottleneck solutions for binary X and Y, which we proved in Proposition 4.2.7 to be successively refinable for all trade-off parameters. We used the linear program to check the convex hull condition numerically for all pairs $\lambda_1 < \lambda_2$ and for distributions $p(X,Y)$ uniformly sampled on the joint probability simplex $\Delta_{\mathcal{X}\times\mathcal{Y}}$. We find that the convex hull condition is indeed always numerically satisfied.

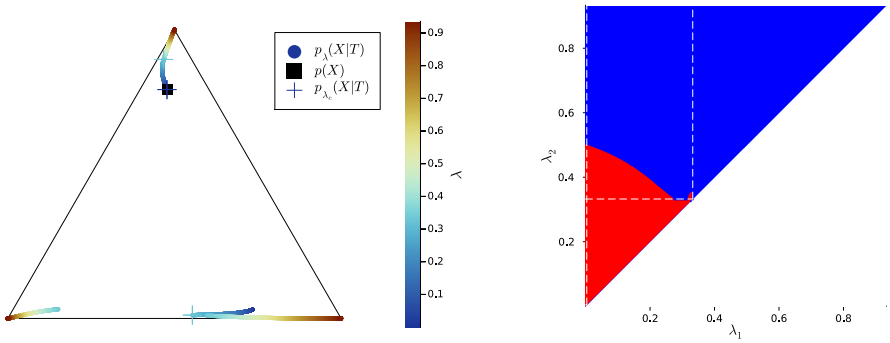
Then, we study the case $|\mathcal{X}| = |\mathcal{Y}| = 3$, once again uniformly sampling example distributions $p(X,Y)$ on $\Delta_{\mathcal{X}\times\mathcal{Y}}$. Figure 4.4 shows, for representative examples, visualisations of the trajectories over λ of the $q_\lambda(X|t)$ (left)—which we will refer to as the *bottleneck trajectories*—along with the corresponding computations of the convex hull condition as a function of λ_1 and $\lambda_2 \geq \lambda_1$ (right)—which we will refer to as the *SR patterns* (The correspodning $p(Y|X)$ are plotted in Appendix D.3, and $p(X)$ is shown in Figures 4.4a–4.4c (left). The explicit $p(X,Y)$ corresponding to each of these paper's figures can be found at: `https://gitlab.com/uh-adapsys/successive-refinement-ib/`(accessed on 14 October 2025).

Let us first give a general description of the bottleneck trajectories. For $\lambda \approx 0$, the $q_\lambda(X|t)$ all coincide with the source distribution $p(X)$. This should be the case, as, for $0 = \lambda = I(X;T)$, the bottleneck T is independent of X. Then, when λ increases, the trajectories seem piecewise continuous, where each discontinuity corresponds to a symbol split, i.e., a change in effective cardinality (see Section 4.1.4). We mark with a cross, for each $t \in \mathcal{T}$, the $q_\lambda(X|t) = q_{\lambda_c}(X|t)$ located just before such a change in effective cardinality.

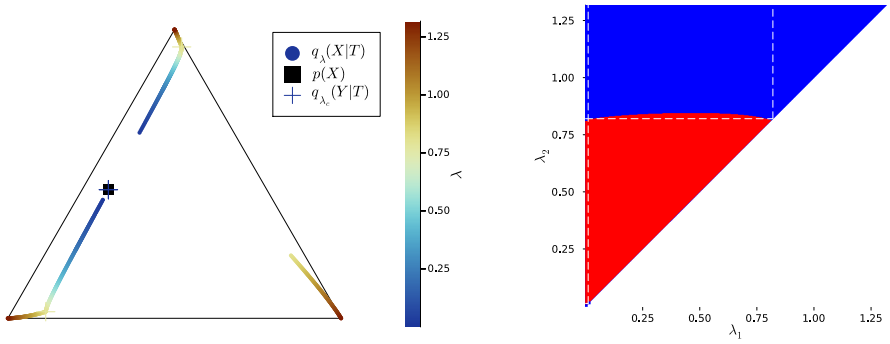
In the examples of Figure 4.4, as $|\mathcal{X}| = 3$, there are two symbol splits, corresponding to that from one to two and two to three symbols, respectively. Eventually, for large λ, the last continuous segment of bottleneck trajectories corresponds to effective cardinality $k(T_\lambda) = |\mathcal{X}|$, and, for the maximal λ, each corner of the source simplex $\Delta_{\mathcal{X}}$ is reached by $q(X|t)$ for some $t \in \mathcal{T}$. This means that for maximum λ, there is a deterministic bijective relationship between T and X. The latter is expected: for maximum λ, bottlenecks are minimal sufficient statistics of X for Y (Shamir et al., 2010); where for $p(X,Y)$ sampled uniformly on the simplex, these minimal sufficient statistics are, with probability 1, just permutations of X.

Definition 4.2.8. In the following, we refer to the piece of trajectory where the bottleneck's effective cardinality $k = k(T_\lambda)$ is equal to the integer i as the *"segment $k = i$"*, i.e., it is the segment where $q_\lambda(X|T)$ corresponds to exactly i distinct points on the source simplex $\Delta_{\mathcal{X}}$;

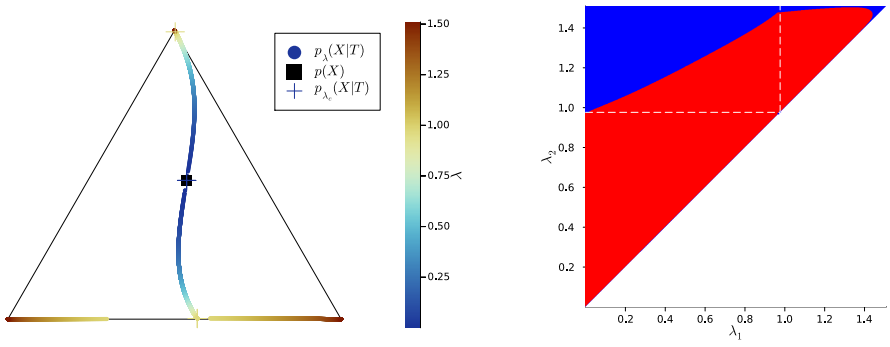
[2] However, at this stage, we do not have a formal guarantee that the convex hull condition failing for two specific bottlenecks implies that SR does not hold.



(A) Example 1.



(B) Example 2.



(C) Example 3.

FIGURE 4.4: On each line, we fix an example distribution $p(X, Y)$. Left: bottleneck trajectories for with $|\mathcal{X}| = |\mathcal{Y}| = 3$: i.e., trajectory of the family of points $\{q_\lambda(X|t), t \in \mathcal{T}\}$ on $\Delta_\mathcal{X}$, as a function of $\lambda = I(X;T)$ (crosses: value of $q_{\lambda_c}(X|T)$ just before a symbol split at a critical parameter λ_c; note that $\lambda_c(1) \approx 0$). The conditional distribution $q_\lambda(X|T)$ is defined by the single point $p(X)$ for $\lambda = 0$ (dark blue cross on the black square); by two points between the first and second symbol splits (dark blue to cyan); by three points after the second symbol split (cyan to red). Note the discontinuity of $q_\lambda(X|T)$ at each symbol split. Right: corresponding SR pattern, i.e., output for the convex hull condition (blue: satisfied; red: not satisfied; dashed white lines: critical values $\lambda_c(i)$ of either λ_1 or λ_2). E.g., $\lambda_c(2) \approx 0.33$ corresponds, on the bottleneck trajectories (left), to the symbol split from two to three symbols (cyan crosses). See Appendix D.3 for the $p(Y|X)$ corresponding to each line.

for instance, in Figure 4.4a, the segment $k = 2$ corresponds to the first piece of trajectory spanning colors from dark blue to cyan.

Notation. We denote by $\lambda_c(i)$ the trade-off parameter's critical value corresponding to the i-th change in effective cardinality, i.e., the symbol split from i to $i+1$ symbols. Here, we will only need to consider the critical values $\lambda_c(1) = 0$ and $\lambda_c(2)$, corresponding to the splits from one to two and two to three symbols, respectively.

Let us now come back to the question of successive refinement: for which parameters $\lambda_1 < \lambda_2$ is the convex hull condition satisfied? The right-hand sides of Figures 4.4a–4.4c provide the answers corresponding to trajectories on the respective left-hand sides—where blue and red mean that the condition is and is not satisfied, respectively. Moreover, we highlight with dashed white vertical and horizontal lines the critical parameter values $\lambda_1 = \lambda_c(i)$ and $\lambda_2 = \lambda_c(i)$, respectively, at which the symbol split occurs (see Appendix D.1.8 for details on the computation of these symbols splits). Note that we always have $\lambda_c(1) \approx 0$, which is expected, as a bottleneck T corresponding to some $\lambda = I(X;T) > 0$ must necessarily define at least two distinct $q_\lambda(X|t)$.

First, in these examples as in most non-reported examples, the convex hull condition (right) breaks as long as $\lambda_2 < \lambda_c(2)$, i.e., as long as the finer bottleneck's effective cardinality is at most $k = 2$. This can also be read from the bottleneck trajectories (left): if the condition was satisfied for all $\lambda_1 < \lambda_2 < \lambda_c(2)$, for instance, then the segment $k = 2$ would be a line segment. This is clearly not the case in Figures 4.4a and 4.4c, and even though visually it virtually seems to be the case in Figure 4.4b, the segment $k = 2$ happens to be very slightly curved, which is enough to break the convex hull condition. In other words, for $\lambda_1 < \lambda_2 < \lambda_c(i)$, several-stage processing seems to induce, in these examples, a nonzero loss of information optimality.

Then, for $\lambda_2 > \lambda_c(2)$, even though there is no single general pattern, the trajectory's structure at the bifurcation seems to impact successive refinement. Indeed, at the bifurcation at $\lambda_c(2)$, the set $\mathsf{Hull}\{q_{\lambda_2}(X|t), t \in \mathcal{T}\}$ opens up along a new, third dimension, and keeps widening when λ_2 increases. This allows it to (gradually in Figures 4.4a and 4.4c, or virtually straight away in Figure 4.4b) encompass the segment $k = 2$ because it "overcomes" the curvature of this piece of trajectory. For instance, in Figure 4.4a, because the segment $k = 2$ is strongly curved, the convex hull condition gets satisfied for all $\lambda_1 < \lambda_c(2)$ only if λ_2 is significantly larger than $\lambda_c(2)$. On the contrary, because in Figure 4.4b, the segment $k = 2$ is virtually not curved, it is almost as soon as $\lambda_2 > \lambda_c(2)$ that the convex hull condition is satisfied for all $\lambda_1 < \lambda_c(2)$.

In Figure 4.4c, the lack of successive refinement for $\lambda_2 > \lambda_c(2)$ does not seem to be due to the same phenomenon as the one just described. Generally speaking, we observed a whole variety of SR patterns (see Appendix F in (Charvin et al., 2023a) for more examples), and our aim here is not to try to interpret all of them. However, despite this diversity, the SR patterns that we studied typically shared a common qualitative feature: the bifurcation structure of the bottleneck trajectories seemingly participates in shaping these SR patterns. Mostly, it seems typically necessary, for SR to hold, that the larger parameter λ_2 has crossed the bifurcation value $\lambda_c(2)$, because the non-zero curvature of the segment $k = 2$ can only be "overcome" by opening the set $\mathsf{Hull}\{q_{\lambda_2}(X|t), t \in \mathcal{T}\}$ along a new dimension, through the symbol split at $\lambda_2 = \lambda_c(2)$. This phenomenon will be explored in more details in Section 4.3.2.

Besides this relationship between SR and the structure of bottleneck bifurcations, this numerical study suggests a generalisation of the notion of successive refinement. Indeed, in Figure 4.4b for instance, even though the right-hand side asserts that successive refinement does not hold for $\lambda_1 < \lambda_2 < \lambda_c(2)$, the virtually linear piece of trajectory on the left-hand side suggests that this is "almost" the case. In the next section, we formalise this intuition.

4.3 Soft Successive Refinement of the IB

The minimal experiments from Section 4.2.3 suggest the intuition that even though successive refinement might not always hold exactly, when broken, it might still be "close" to being satisfied. More generally speaking, let us recall that we are trying here to understand the informationally optimal limits of several-stage information processing. As our numerical experiments suggest that the IB problem is not always successively refinable, it is desirable to *quantify* the lack of successive refinement—i.e., the lack of informational optimality induced by several-stage processing. These considerations lead to the notion of *soft successive refinement* (Catenacci Volpi et al., 2020), which we define and motivate in this section. As we will see, this generalisation of exact SR does not depend on the specific structure of the IB setting; rather, it can also be used as a generalisation of exact SR for *any* rate-distortion scenario.

4.3.1 Formalism

Let us first focus on the case $n = 2$: we thus want to quantify the amount of information captured by a coarse bottleneck T_1 and then discarded by a finer bottleneck T_2. Let us recall that, from Proposition 4.2.2, bottlenecks T_1 and T_2 achieve successive refinement if there exists an extension $q(X, T_1, T_2)$ of $q_1(X, T_1)$ and $q_2(X, T_2)$ such that, under q, we have the Markov chain $X - T_2 - T_1$, which is equivalent to $I_q(X; T_1|T_2) = 0$. It thus seems natural to quantify soft successive refinement with the conditional mutual information $I_q(X; T_1|T_2)$. However, the IB method does not entirely define the relationship between distinct bottlenecks; formally, there is a whole polytope $\Delta_{q_1,q_2} \subseteq \Delta_{\mathcal{X}\times\mathcal{T}_1\times\mathcal{T}_2}$ of possible extensions $q(X, T_1, T_2)$ of $q_1(X, T_1)$ and $q_2(X, T_2)$ (see Section 4.1.4). Among these possible extensions, it seems natural to search for those that minimise the violation of the SR condition $I_q(X; T_1|T_2) = 0$. This leads us to use the *unique information* (Bertschinger et al., 2013)

$$UI(X : T_1 \setminus T_2) := \min_{q\in\Delta_{q_1,q_2}} I_q(X; T_1|T_2). \tag{4.3.1}$$

This quantity was already defined in (Bertschinger et al., 2013) in the context of partial information decomposition (Griffith et al., 2014; Harder et al., 2013; Williams et al., 2010), and it happens to be relevant to us for several reasons.

First of all, it depends only on the distributions $q_1(X, T_1)$ and $q_2(X, T_2)$, which are indeed the only distributions provided by the IB framework. Second, from Proposition 4.2.2, there is successive refinement if and only if there are two bottlenecks T_1 and T_2 such that $UI_{q_1,q_2}(X : T_1 \setminus T_2) = 0$. Third, it is thoroughly argued in (Bertschinger et al., 2013) that (4.3.1) is a good measure of the information that only T_1, and not T_2, has about X, which is an interpretation that coincides neatly with the intuition that we want to operationalise here. Eventually, Proposition 4.3.3 below, which first requires some definitions, provides an information-geometric justification.

Definition 4.3.1. For Δ a probability simplex and $E_1, E_2 \subseteq \Delta$, we define

$$D_{KL}(E_1||E_2) := \inf_{r_1\in E_1, r_2\in E_2} D_{KL}(r_1||r_2),$$

where D_{KL} is the Kullback–Leibler divergence: $D_{KL}(r_1||r_2) := \sum_{a\in\mathcal{A}} r_1(a) \log\left(\frac{r_1(a)}{r_2(a)}\right)$, if the probability distributions r_1 and r_2 are defined on the discrete alphabet $\mathcal{A}$.

Definition 4.3.2. The *successive refinement set* $\Delta_{SR,n} \subseteq \Delta_{\mathcal{X}\times\mathcal{T}_1\times\cdots\times\mathcal{T}_n}$ is the set of distributions r on $\mathcal{X} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_n$ such that, under r, the Markov chain $X - T_n - \cdots - T_1$ holds.

Note that $\Delta_{SR,n}$ does not require its elements to be extensions of any fixed bottleneck distributions $q_i(X, T_i)$ but imposes the Markov chain that characterises SR (see Propostion

4.2.2). SR is achieved for bottlenecks $q_1(X,T_1),\dots,q_n(X,T_n)$ if and only if the successive refinement set $\Delta_{SR,n}$ and the extension set $\Delta_{q_1,\dots,q_n}$ share a common distribution $q \in \Delta_{SR,n} \cap \Delta_{q_1,\dots,q_n}$. In general (for $n=2$), the following proposition can easily be derived:

Proposition 4.3.3. *For fixed distributions $q_1 = q_1(X,T_1)$, $q_2 = q_2(X,T_2)$, we have*

$$UI(X : T_1 \setminus T_2) = D_{KL}(\Delta_{q_1,q_2} || \Delta_{SR,2}). \tag{4.3.2}$$

Proof. See Appendix D.2.1. □

In this sense, $UI(X : T_1 \setminus T_2)$ quantifies "how far" the joint distributions extending the bottlenecks T_1 and T_2 are from making the successive refinement condition $X - T_2 - T_1$ hold true, where the "distance" is understood as a minimised Kullback–Leibler divergence.

Our new measure of soft SR is continuous:

Proposition 4.3.4 ((Rauh et al., 2019), Property P.7)**.** *The unique information $UI(X : T_1 \setminus T_2)$ is a continuous function of the probabilities $q_1(X,T_1)$ and $q_2(X,T_2)$.*

Remark 4.3.5. In particular, if $UI(X : T_1 \setminus T_2)$ has a discontinuity as a function of the parameter λ_1 or λ_2, which define the bottleneck distribution $q_{\lambda_1}(X,T_1)$ or $q_{\lambda_2}(X,T_2)$, respectively, then this can only be a consequence of a discontinuity of the probability $q_{\lambda_1}(X,T_1)$ as a function of λ_1 or $q_{\lambda_2}(X,T_2)$ as a function of λ_2, itself, respectively. This consideration will be useful for analysing our numerical experiments in Section 4.3.2.

Moreover, the formulation (4.3.2) of unique information suggests a natural generalisation to an arbitrary number of processing stages:

Definition 4.3.6. Let $T_1,\dots,T_n$ be bottlenecks with respective parameters $\lambda_1 < \dots < \lambda_n$, and $q_i(X,T_i)$ their respective individual distributions. One can quantify *soft successive refinement*, or, equivalently, the *lack of successive refinement*, through the divergence $D_{KL}(\Delta_{q_1,\dots,q_n} || \Delta_{SR,n})$.

While (Banerjee et al., 2018) proposes a provably convergent algorithm to compute $UI(X : T_1 \setminus T_2)$, to the best of our knowledge, there currently exists no provably convergent algorithm to compute $D_{KL}(\Delta_{q_1,\dots,q_n} || \Delta_{SR,n})$ for $n > 2$. Our numerical investigations (see Section 4.3.2) will stick to the case $n = 2$.

For the sake of completeness, let us point out that for each λ, there is a whole set of solutions $q_\lambda(T|X)$—or, equivalently, $q_\lambda(X,T)$—to the IB problem (4.1.2). Thus, the unique information, which is defined as a function of specific bottleneck distributions $q_1(X,T_1)$ and $q_2(X,T_2)$, could *a priori* not be uniquely defined by the corresponding trade-off parameters λ_1 and λ_2.

4.3.2 Numerical Results on Simple Examples

A provably convergent algorithm that computes, in the discrete case, the unique information (4.3.1), was provided in (Banerjee et al., 2018). In this section, we use the authors' implementation of this algorithm (`https://github.com/infodeco/computeUI`, accessed on 14 October 2025) to qualitatively investigate, on minimal examples, the landscapes of unique information (UI) and their relationship to the bottleneck trajectories on the simplex.

In Figures 4.5a–4.5c (left), we plot again the same bottlenecks trajectories as in Figures 4.4a–4.4c (left), but compare them this time with the unique information $UI(X : T_1 \setminus T_2)$, plotted as a function of λ_1 and λ_2 (right). We also plot, in Figures 4.6a–4.6c, some representative examples of the exact SR patterns (left) and UI landscapes (right) for slightly larger source and relevancy cardinalities, where $p(X,Y)$ is, as above, uniformly sampled — the explicit distributions $p(X,Y)$ corresponding to Figures 4.6a–4.6c can be found at `https://gitlab.com/uh-adapsys/successive-refinement-ib/`.

Once again, we highlight with dashed white vertical and horizontal lines the critical parameter values $\lambda_1 = \lambda_c(i)$ and $\lambda_2 = \lambda_c(i)$, respectively, where, as expected, $\lambda_c(1) \approx 0$. We will first describe, for a fixed $p(X, Y)$, the relative variations in unique information as a function of λ_1 and λ_2. Then, we will compare the absolute values of unique information to the information globally processed by the system.

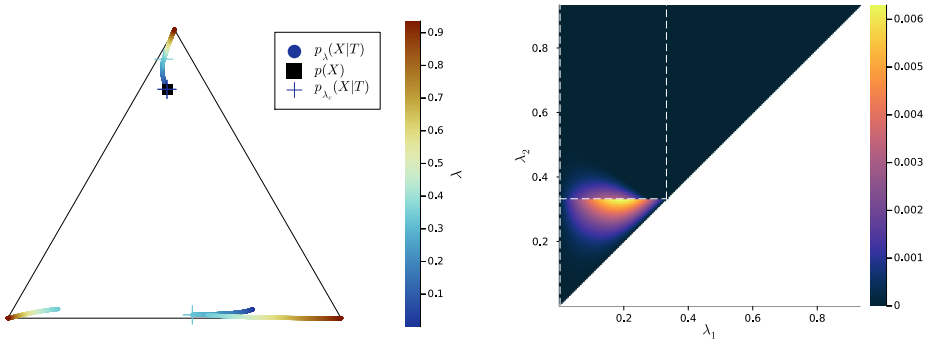
For all Figures from Figure 4.5a to Figure 4.5c, the UI landscape partly mirrors the respective exact SR pattern of Figures 4.4a–4.4c (right). However, within the region where these latter figures answered a binary "no" to the question of exact SR, Figures 4.5a–4.5c reveal a sharply uneven variation in the violation of SR, where, for important ranges of trade-off parameters, the unique information is negligible comparative to others. For instance, even though Figure 4.4b (right) seems to indicate that SR does not hold for $\lambda_1 < \lambda_2 < \lambda_c(2)$, the corresponding UI in Figure 4.5b (right) is virtually zero on a large part of this set of parameters, while still peaking for λ_2 close to $\lambda_c(2)$. This richer structure of the unique information landscape is further evidenced by Figures 4.6a–4.6c.

Moreover, the unique information landscapes seem shaped by the bottleneck trajectories. Most importantly, the influence of IB bifurcations on SR can be seen even more clearly with soft than with exact SR. In particular, in Figures 4.6a–4.6c, it seems that along the lines where one of the trade-off parameters crosses a critical value, the UI often goes through discontinuities, or at least sharp variations in either λ_1, λ_2, or both directions. In particular, even though patterns widely vary across different example distributions $p(X, Y)$, unique information can significantly *drop* when λ_2 crosses a critical value from below—a feature observed in both shown and non-shown examples. As we know that the unique information is continuous, the apparent discontinuity should be one of the bottleneck probability $q_{\lambda_2}(X, T_2)$ itself (see Proposition 4.3.4 and Remark 4.3.5). This is consistent with the observation from Section 4.2.3 that, at symbol splits, the trajectory of $q_\lambda(X|T)$ often seems to go through a discontinuity. Further, the fact that the sharp variation in UI is a *decrease* in this quantity (in increasing order of λ_2) is intuitively consistent with the fact that the bottleneck trajectory's discontinuity often induces a sudden "widening" (in increasing order of λ) of

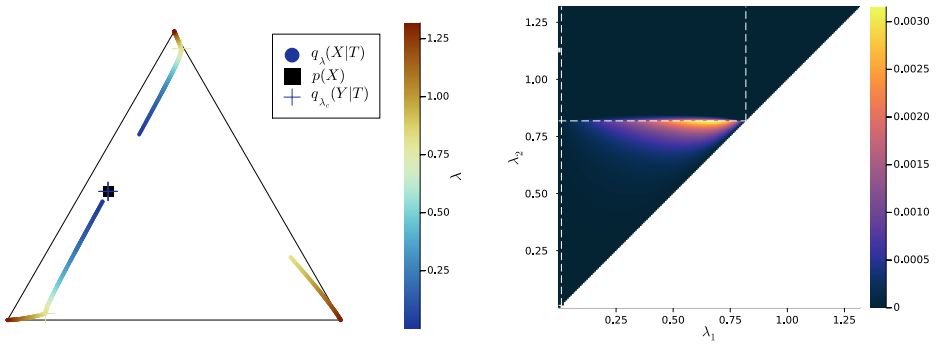
$$\mathcal{H}_T := \mathsf{Hull}\{q(X|t),\, t \in \mathcal{T}\}.$$

Indeed, for fixed λ_1, when λ_2 crosses a critical value from below, the corresponding symbol split means that $\mathcal{H}_{T_2}$ "widens" by opening up a new dimension, so it "more easily" encompasses $\mathcal{H}_{T_1}$, yielding as a consequence a drop in unique information. Recalling our intuition (see Section 4.2.2) that $\mathcal{H}_T$ describes the information content that a bottleneck T contains about the source X, the feature just described can be interpreted in the following way: the IB bifurcations seem to induce a sudden "expansion" (in increasing order of λ) of the information content carried by the bottleneck about the source, which makes the latter's content more easily contain the information content of coarser bottlenecks.

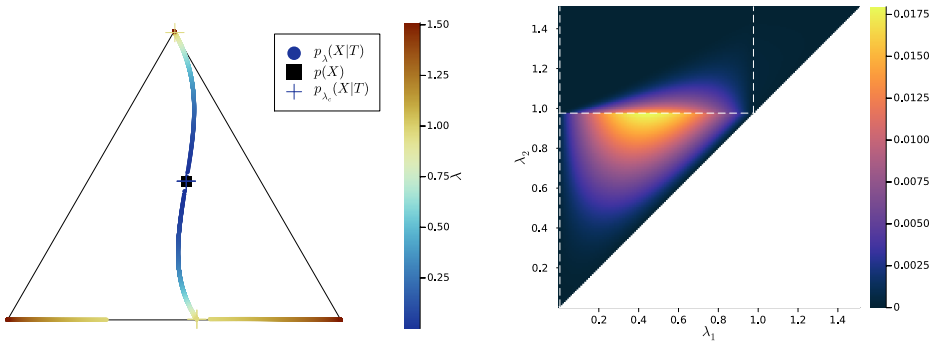
Note, however, that these simple numerical results do not allow one to discriminate between the interpretation of the UI's sharp variations at bifurcations as a discontinuity with regard to trade-off parameters, or a discontinuity of the UI's *differential*. For instance, if the derivative with regard to λ_2 discontinuously takes a value close to $-\infty$ for λ_2 slightly larger than some λ_c, then the UI graph can seem discontinuous at finite numerical resolution, even if, formally, only the UI's differential is so. On the other hand, as an example, bifurcations can be characterised precisely as points of discontinuities of the derivatives, with regard to the trade-off parameter, of $I(T; X)$ and $I(T; Y)$ (Chechik et al., 2005; Zaslavsky et al., 2019), even though the functions themselves are continuous (Chechik et al., 2005; Gilad-Bachrach et al., 2003). A more involved analysis distinguishing discontinuities of UI from those of its differential is left to future work. In any case, the interpretation as a discontinuity of the differential rests on a weaker assumption, which is still sufficient for explaining the numerical



(A) Example 1.

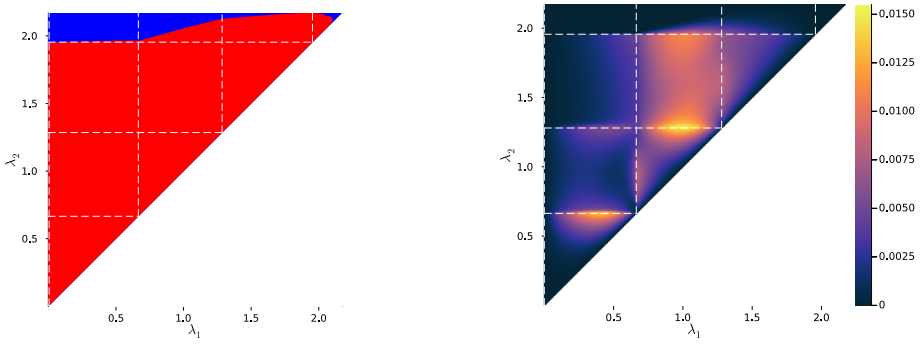


(B) Example 2.

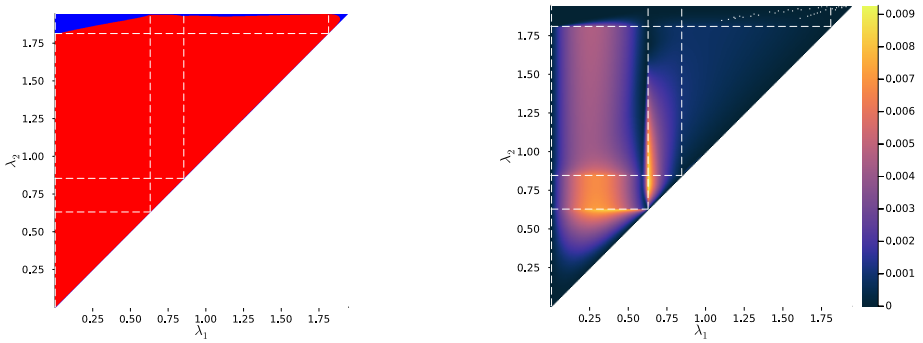


(C) Example 3.

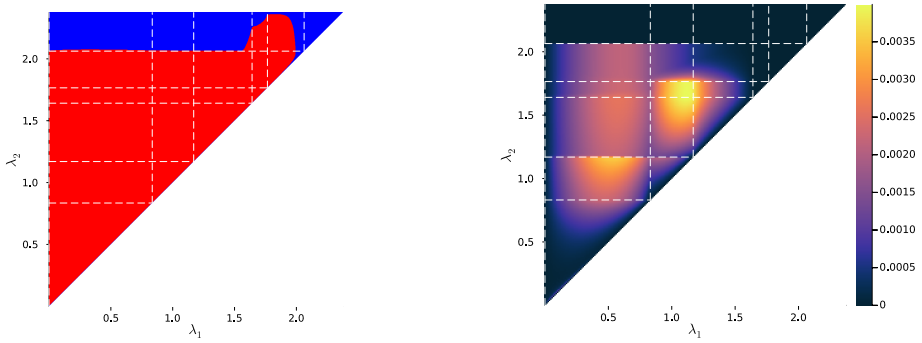
FIGURE 4.5: On each line, we fix the same $p(X, Y)$ as in resp. Figures 4.4a, 4.4b and 4.4c. Left: example trajectory of $q_\lambda(X|T)$ (same as in Figure 4.4). Right: corresponding unique information, in bits (color), expressed as a function of the pair of trade-off parameters (white dashed lines indicate critical values $\lambda_c(i)$ of either λ_1 or λ_2.). For instance, the critical value $\lambda_c(2) \approx 0.33$ (right) corresponds, on the bottleneck trajectories (left), to the symbol split from two to three symbols (cyan crosses). See Appendix D.3 for the $p(Y|X)$ corresponding to this figure.



(A) New example $p(X, Y)$, with $|\mathcal{X}| = 5$ and $|\mathcal{Y}| = 3$.



(B) New example $p(X, Y)$, with $|\mathcal{X}| = 5$ and $|\mathcal{Y}| = 3$. White dots in the top right corner correspond to values of (λ_1, λ_2) for which the algorithm did not converge (see main text for a comment on this lack of convergence).



(C) New example $p(X, Y)$, with $|\mathcal{X}| = 7$ and $|\mathcal{Y}| = 5$.

FIGURE 4.6: On each line: example, for a new $p(X, Y)$, of exact SR patterns and the corresponding UI landscapes over trade-off parameters $\lambda_1 < \lambda_2$. Left: exact SR pattern, i.e., output for the convex hull condition (blue: satisfied, red: not satisfied). Right: corresponding UI landscape, in bits (color). White dashed lines indicate critical values $\lambda_c(i)$ of either λ_1 or λ_2. Note that (*i*) the binary notion of exact SR (left) does not reflect most of the structure unveiled by UI (right), (*ii*) the UI landscape seems highly impacted by IB bifurcations, and (*iii*) the UI is in any case always small, even though not entirely negligible. See main text for more details.

results.

More generally, these results suggest that for a several-stage processing that is IB-optimal at each stage, to minimise the information discarded along stages, the trade-off parameters should rather lie close to well-chosen IB bifurcations. If this happens to be a general feature of the IB framework, it would have implications for incremental learning. Indeed, coming back to the modelling of embodied agents (see Section 4.1), for instance, it would mean that organisms that are poised close to information optimality by evolution should have a very specific structure of developmental learning, where the stages of learning should be discrete and determined by the right trade-off parameters.

Eventually, a last crucial feature was also satisfied on these minimal examples: whatever the structure of bottleneck trajectories, the maximal UI was significantly lower than the mutual information $I(X;T_1,T_2)$ between the external source X and the system's pair of bottlenecks (T_1,T_2). More precisely, for an extension $q(X,T_1,T_2)$ of $q_{\lambda_1} := q_{\lambda_1}(X,T_1)$ and $q_{\lambda_2} := q_{\lambda_2}(X,T_2)$ that achieves the minimum in (4.3.1), let us define

$$\sigma(q_{\lambda_1},q_{\lambda_2}) := \frac{UI_{q_{\lambda_1},q_{\lambda_2}}(X : T_1 \setminus T_2)}{I_q(X;T_1,T_2)}.$$

Note that decomposing $I_q(X;T_1,T_2)$, where $q \in \Delta_{q_1,q_2}$, with the chain rule for mutual information shows that this quantity only depends on q_{λ_1} and q_{λ_2}: thus here, $\sigma(q_{\lambda_1},q_{\lambda_2})$ is indeed well-defined by q_{λ_1} and q_{λ_2}. The maximum ratio over all trade-off parameters $\lambda_1 < \lambda_2$ was typically of the order of 1% in our experiments; for instance, it was 1.89%, 0.39%, 1.82%, 2.03%, 1.34%, and 0.31% for the IB problems corresponding to Figures 4.5a–4.6c, respectively. Among all the (shown and non-shown) studied examples, it never exceeded 5.4%, and we did not notice an increase in this maximum ratio when the source or relevancy cardinalities were increased (the largest cardinalities that we experimented with were $|\mathcal{X}| = 20$, $|\mathcal{Y}| = 10$). In short, even though several-stage processing might incur a non-negligible loss of information optimality in the IB sense, these results suggest that this loss could often be significantly limited. Of course, here as in Section 4.2.3, on the one hand, the numerical results are purely phenomenological, and, on the other, it is at this stage far from being clear that the qualitative insights brought by these minimal experiments generalise well to more complex situations. However, they exhibit the potentially crucial qualitative features of exact and soft successive refinement in the IB framework, which can be targeted by further theoretical research.

4.4 Interpretations in terms of Decision Problems

Here, we show that exact and soft successive refinement can be, in the discrete case at least, understood in terms of optimally solving decision problems on arbitrary tasks, through orders on the compression channels $q(T|X)$ (or more precisely, pre-orders: i.e., we will consider binary relations that are reflexive and transitive). We will rely on (Bertschinger et al., 2014), where these orders were considered.

Let us first make clear what we mean here by a decision problem. Consider a state variable X over a finite set $\mathcal{X}$, another finite set $\mathcal{A}$ of possible actions, and a reward function $u = u(x,a)$ that depends on both the value x of the state X, and the chosen action $a \in \mathcal{A}$. The agent's observation is not the state X itself, but only the output T of X through some stochastic channel $\kappa := p(T|X)$ (where we assume here that the observation space $\mathcal{T}$ is finite). To each observation-dependent policy $\pi = \pi(A|T)$ corresponds an expected reward

$$\mathbb{E}_\pi(u(X,A)) := \sum_t p(t)\mathbb{E}_{(X,A)\sim p(X|t)\pi(A|t)}(u(X,A)),$$

where $p(X|t)$ is determined from $\kappa := p(T|X)$, $p(X)$ through the Bayes rule, and $p(X|t)\pi(A|t)$ denotes the product measure of $p(X|t)$ and $\pi(A|t)$. Solving the decision problem $(p(X), \mathcal{A}, u)$ for the observation channel κ means choosing a policy that yields an optimal expected reward

$$\mathcal{R}(p(X), \kappa, u) := \max_{\pi} \mathbb{E}_{\pi}(u(X, A)).$$

We can now define the following order (Bertschinger et al., 2014):

Definition 4.4.1. For two channels κ and κ', we write $\kappa \sqsupseteq_{\mathcal{X}} \kappa'$, if, for any decision problem $(p(X), \mathcal{A}, u)$, we have

$$\mathcal{R}(p(X), \kappa, u) \geq \mathcal{R}(p(X), \kappa', u).$$

In short, $\kappa \sqsupseteq_{\mathcal{X}} \kappa'$ means that, for any conceivable task based on any data distribution $p(X)$ over the fixed data space $\mathcal{X}$, the observation channel κ can yield a performance at least as good as that of the observation channel κ'—if combined with a well-chosen policy. The second order is the *Blackwell order* (Blackwell, 1953):

Definition 4.4.2. For two channels κ and κ', we write $\kappa \sqsupseteq'_{\mathcal{X}} \kappa'$ if there exists a channel γ such that $\kappa' = \gamma \circ \kappa$, where "$\circ$" denotes the composition of channels.

It turns out that successive refinement can be characterised by either of these two orders, thanks to the Sherman–Stein–Blackwell theorem (Bertschinger et al., 2014; Blackwell, 1953). In other words, SR, which is *a priori* not a decision-theoretic statement, turns into one through its equivalence with the Blackwell order:

Proposition 4.4.3. *Let* $0 < \lambda_1 < \lambda_2$. *The following are equivalent:*

(*i*) *There is successive refinement for parameters* (λ_1, λ_2).

(*ii*) *There are bottlenecks* T_1, T_2 *of respective parameters* λ_1, λ_2 *such that*

$$q(T_2|X) \sqsupseteq_{\mathcal{X}} q(T_1|X).$$

(*iii*) *There are bottlenecks* T_1, T_2 *of respective parameters* λ_1, λ_2 *such that*

$$q(T_2|X) \sqsupseteq'_{\mathcal{X}} q(T_1|X).$$

Proof. Using the Markov chain characterisation (point (*ii*) in Proposition 4.2.2), the result is nothing more than a reformulation of Theorem 4 in (Bertschinger et al., 2014) in the language of the present paper. Note that, to use this theorem, we need to assume that the source X is fully supported, but this is indeed an assumption that we are using along the whole paper because it does not incur any loss of generality (see Section 4.1.4). □

Point (*ii*) means that there is SR when the coarser coarse-graining T_1 can be retrieved by post-processing the finer coarse-graining T_2. Now, the equivalence of SR with point (*iii*) relies on the mathematically deep part of the Sherman–Stein–Blackwell theorem (Bertschinger et al., 2014), and provides a new operational meaning to SR. Namely, there is SR when, for *any* distribution $q(X)$ on the source, and *any* reward function, the optimal performance is at least as good when the decisions are based on the output of $q(T_2|X)$, seen as an observation channel, than when they are based on the output of $q(T_1|X)$. Crucially, this property does not directly make reference to the source-relevancy ditribution $p(X, Y)$.

For example, assume that evolution poises the sensors of a given biological organism at optimality in the IB sense (Palmer et al., 2015; van Dijk et al., 2012), i.e., if X is the environment, S some sensor's output (e.g., a retina's ganglion cells activation), and Y a behaviourally

relevant feature (e.g., the edibility of food), then S is a bottleneck for $p(X,Y)$. Successive refinement here means that if the sensor S_2 is finer than S_1 as a bottleneck for the fixed feature Y relevant to a particular task, then S_2 will afford to the organism—if combined with the right decision making—better performances than S_1 on any other task, for any other input distribution $p(X)$. In other words, S_2 is then "universally better" than S_1.

Eventually, the unique information that we chose as our measure of soft SR has initially been thought precisely as measuring the deviation from the order "$\sqsupseteq_{\mathcal{X}}$" (see arguments in (Bertschinger et al., 2014)). Unique information can thus, for instance, be understood as quantifying the deviation from a finer IB-optimal sensor to be "universally better" than a coarser one.

4.5 Limitations and Future Work

Our convex hull characterisation intertwines the question of exact SR with the more fundamental question of the structure of decoder curves

$$\big\{\big(\lambda \mapsto q_\lambda(X|t)\big),\ t \in \mathcal{T}\big\} \tag{4.5.1}$$

on the source simplex $\Delta_{\mathcal{X}}$, a question for which the convexity approach to the IB problem (Asoodeh et al., 2020; Benger et al., 2023; Dikshtein et al., 2021; Hsu et al., 2018; Witsenhausen et al., 1975) seems promising. In short, this approach reformulates the IB problem to that of finding the lower convex envelope of a well-chosen function F_β, defined on the source simplex $\Delta_{\mathcal{X}}$, and parameterised by the information curve's inverse slope β (see Appendix D.1.7). More precisely, bottlenecks are essentially characterised by the fact that the lower convex envelope must be achieved by convex combinations of the points $F_\beta(q(X|t))$; this approach thus provides analytical tools for proving key properties of the set of trajectories (4.5.1), which would then have consequences for SR through the convex hull condition. Despite the limited scope of the result itself, the proof of Proposition 4.2.7 gives an example of such a fruitful interaction, thus suggesting a way forward for further theoretical progress. As a first step, one could try to use the convexity approach to the IB to prove the unicity (up to permutations) and injectivity of $q(X|T)$, for canonical bottlenecks and the strictly concave information curve. We expect such a result to simplify our convex hull characterisation of SR — in the case of the strictly concave information curve. Generally speaking, leveraging, through our convex hull characterisation, the convexity approach to the IB problem might allow one to (i) identify new wholly refinable IB problems, but also (ii) produce general methods to identify, for a given distribution $p(X,Y)$, the combination of parameters for which exact SR holds.

It must be stressed that even though we motivate the successive refinement of the IB by diverse scientific questions in Sections 4.2 and 4.4, in this work, we do not model any concrete system. Rather, our minimal numerical experiments target the qualitative exploration of the formalised problem. Our results might in turn be relevant for future modelling work (see the last paragraph of this section), but the most pressing aspect is to first develop further the theoretical and computational framework. In particular, it seems important to describe formally the apparent discontinuity of UI (or its differential) as a function of the trade-off parameters λ_1 and λ_2 at IB bifurcations (through that of the $q_\lambda(X,T)$ as functions of λ); to describe more formally why the UI tends to peak and then drop close to IB bifurcations; to provide global bounds on UI in general or as functions of the source and relevancy distribution $p(X,Y)$; or to make formal the informal relationship between the "extent to which" the convex hull condition is broken, and variations in UI. Another interesting contribution would be to provide an asymptotic coding interpretation to unique information; indeed, the deviation from successive refinement is more classically quantified as a difference between asymptotic rates or

distortions (see, e.g., (Lastras et al., 2001)), and it is not clear whether or not this interpretation can be made for UI. Numerically speaking, one could design algorithms allowing for the computation of UI for continuous $p(X,Y)$ and/or more than two processing stages. Indeed, the algorithm from (Banerjee et al., 2018) only encompasses the case of discrete variables and two processing stages. One could, for instance, take inspiration from (Banerjee et al., 2018) to formulate the quantity $D_{KL}(\Delta_{q_1,\dots,q_n}||\Delta_{SR,n})$ as a double minimisation problem over separate parameters, allowing for an alternating optimisation algorithm.

Let us point out that our framework considers that the source of information X and the target variable Y are the same along all processing stages. More general frameworks could allow for variations in either the source of information (as in the case in temporal series) or the target variable (as is the case in transfer learning). Frameworks for both these kinds of extensions have already been proposed (Mahvari et al., 2021; Yang et al., 2017), and it would be interesting to study if, in these cases as well, the specific nature of the IB problem imprints the informationally optimal limits of several-stage processing.

Eventually, we deem the interpretation in terms of the incorporation of information to be particularly relevant to modelling adaptive behaviour. For instance, for a given developmental or skill-learning problem on a given task, our framework could help in distinguishing situations where the choice of the successive processing stages' complexity along incrementally learning the task does not matter (i.e., when there is successive refinement) from situations where these complexities must be minutely weighed, so as to avoid as much as possible the "waste" of cognitive work along the way (i.e., when the unique information is not negligible and unevenly distributed). In the latter case, our framework, once mature, might precisely describe those sequences of processing stages' complexity that minimise the "waste" of cognitive work from one learning stage to another, thereby potentially identifying key stages of skill or developmental learning. Future work should keep in mind the horizon of identifying such qualitative features and producing measures capturing the relevant phenomena for experimental research in these areas.

4.6 Conclusions

In this chapter, we develop further the framework of successive refinement of the IB, particularly through a geometric approach to the problem; and we investigate soft successive refinement of the IB.

After motivating the formal problem, we turn to the mathematical analysis of it. We first note that, for jointly Gaussian vectors (X,Y) or for deterministic $p(Y|X)$, successive refinability can be easily drawn from existing IB literature (Kline et al., 2022; Kolchinsky et al., 2019). Then, we propose a new geometric characterisation of SR, which builds on the intuition that what is "known" by a bottleneck is the convex hull of its decoder conditional probabilities. This new point of view, associated with an active approach that reformulates the IB problem as that of finding the lower convex envelope of a well-chosen function (Asoodeh et al., 2020; Benger et al., 2023; Dikshtein et al., 2021; Hsu et al., 2018; Witsenhausen et al., 1975), provides a new tool for theoretical research on this topic. We exemplify this potential fertility by proving, thanks to the combination of our convex hull characterisation with the convexity approach to the IB, the successive refinability of binary source X and binary relevancy Y (Proposition 4.2.7). This convex hull characterisation also allows one to numerically investigate SR with a linear program, which can be helpful for computational studies on this topic. Our own minimal numerical experiments suggest that (i) successive refinement does not always hold for the IB, (ii) the successive refinement patterns are shaped by IB bifurcations, and (iii) even when successive refinement seems to break, sometimes it is "close" to being satisfied, in the sense of the convex hull condition being only "slightly" violated.

To formalise this latter intuition, we propose to soften the traditional notion of SR into a *quantification* of the loss of information optimality incurred by several-stage processing. For that purpose, we call on the measure of unique information (UI) used in (Bertschinger et al., 2013). Intuitively, this quantity measures the information that only the coarser bottleneck T_1, and not the finer one T_2, holds about the source X, and it can be generalised to an arbitrary number of processing stages. Our minimal experiments, in the case of two processing stages, unveil a rich structure of soft SR that was partially hidden by exact SR, which only makes the distinction between vanishing UI (if there is SR) and positive UI (if there is no SR). Even though the UI landscapes depend strongly on the distribution $p(X, Y)$ that defines the IB problem, some qualitative features seem to emerge: (*i*) the "more" the convex hull condition is broken, the higher the unique information; (*ii*) the IB bifurcations crucially shape the UI landscape, with sharp decreases in unique information in particular when the finer trade-off parameter λ_2 crosses a bifurcation critical value; and (*iii*) in any case, this violation of successive refinement seems to always be mild compared to the system's globally processed information.

The features exhibited by these numerical experiments offer a "first outlook" of potentially general properties of exact and soft successive refinement for the IB problem, thus providing a guide for future theoretical research. These potential properties might provide interesting perspectives on the scientific questions that motivate the formalism, particularly in terms of the incorporation of information along processing stages. For instance, the apparently important role of bifurcations in exact and soft successive refinement suggests that informationally optimal several-stage learning or processing should ideally be organised along well-chosen "checkpoints" on the information plane. Moreover, if the loss of information optimality induced by this sequential processing is indeed typically low (even though not entirely negligible) for the IB framework, this could be taken as an indication that incremental learning might be made highly efficient. These potential features thus provide a strong incentive to bring the formal framework presented here closer to maturity—for instance, along the lines of research proposed in Section 4.5.

Chapter 5

Conclusion

This thesis has developed generalisations of different notions of invariance and equivariance, with a strong focus on the information-theoretic counterpart of group-theoretically defined symmetries.

In Chapter 2, I presented a characterisation of channel invariances with the IB method, a similar characterisation of channel equivariances with a novel variation of the IB, and a broader Divergence IB framework that encompasses both and might be instrumental in capturing alternative notions of symmetry. These information-theoretic characterisations yield new principled notions of *soft symmetries*, defined, in short, as transformations that are "made indistinguishable from the identity" by an optimal compression partially preserving a well-chosen information quantity — where the "softness" of the symmetry is parametrised by the trade-off between compression and information preservation.

In Chapter 3, the focus was on the class-pose decomposition framework, which aims at decomposing a group action ρ on a given state-space $\mathcal{X}$ into an invariant part $\mathtt{Id}_C$ on a "class" space C and an equivariant part ξ on a "pose" space $\mathcal{P}$. We saw that this construction was only possible under restrictive assumptions, which lead to a broader *minimal class-pose parametrisation* framework addressing these limitations, where the group action ρ on $\mathcal{X}$ becomes a "maximally isomorphic" factor of the action $\mathtt{Id}_C \otimes \xi$ on the class-pose space $C \times \mathcal{P}$. Crucially, this construction was then adapted to a measure-theoretic setting that can capture closed-loop and stochastic actions, formalised as a Markov Decision Process (MDP) with fixed policy. By extending previous notions from ergodic theory, classes were then reframed as *ergodic components* of the fixed-policy MDP, and poses as the *minimal joining* of the MDPs induced on these ergodic components. These novel notions of class and pose were then characterised in terms of information parsimony. In particular, the characterisation of ergodic components with an instance of the Divergence IB marks an important step in the formalisation of the duality between symmetry and information parsimony, as it explicitly binds transformations of a space to the coarse-grainings that capture the corresponding invariants (see Section 3.6.1 there).

In Chapter 4, I presented a study of *successive refinement* in the IB framework, which describes the informationally optimal limits on several-stage processing. While successive refinement can be interpreted in terms of incremental learning, it is equivalent to the question of whether, given a coarser and a finer bottleneck, the coarser bottleneck can be obtained by coarse-graining the finer one — a question that sheds light on the inclusion relations among soft channel invariances. We obtain a geometric characterisation in terms of convex hulls defined by information channels, and prove the successive refinability of the IB in some specific cases. We then study, on simple synthetic examples, the "lack" of successive refinability through a previously introduced measure of *unique information*, which yields to qualitative insights on the relations between IB coarse-grainings of varying granularity.

As explained in the introductory Chapter 1 and along the rest of the thesis, these mostly formal results are aimed at contributing to research in adaptive behaviour, structure discovery

and sensorimotor perception. To conclude this thesis, I outline below questions that are raised by these results and could provide interesting directions to future research in these fields.

5.1 Abstract symmetries, information parsimony and SMCs

The clearest path for future work opened by this thesis is that outlined in Sections 2.5.2 and 2.5.3. There, our focus is on the information-theory based discovery of *abstract symmetries*, i.e., symmetries of systems that might potentially include an agent's sensorimotor interface, but where the symmetry transformations do not model the agent's own actions. We argue that the transformations defining the symmetry and the compression channels capturing the corresponding invariants should be discovered *jointly*, by solving a mutli-objective optimisation problem over *both* these objects. The question then becomes to identify the appropriate information-theoretic objectives, in particular one that *binds* the compression channel and the transformations, by requiring the former to coarse-grain the features left invariant by the latter. It turns out that the characterisation of ergodic components obtained in Chapter 3 provides a promising tool to achieve this "transformation-based" extension of the Divergence IB framework.

Future work should realise this extension, in the case of channel equivariances but also at the level of generality provided by Divergence IB framework. This flexibility could make these tools applicable in diverse embodiment-relevant settings. For instance, they could allow one to discover the equivariances of the transition function of MDPs — see equation (1.2.1) — which can be seen as a simple formalisation of a sensorimotor interface in a fully observed setting (see Section 1.2.3). Importantly, here, the symmetry group would now be known in advance, but now would be part of what is being discovered. Such an MDP result could then be applied to the ϵ-transducer of input-ouput processes (Barnett et al., 2015; Shalizi, 2001), which have been used to model embodied agents' sensorimotor interfaces (Marzen, 2025; Rosas et al., 2024).[1] Alternatively, one could instantiate these tools in causal Bayesian networks modeling agents' perception-action loop (Langer et al., 2024; Tishby et al., 2011).

Through such formal advances, the tools developed in this thesis could thus help investigate generalised (exact and soft) symmetries of embodied agents' sensorimotor interface — which can be seen as a formalisation of *apparatus-related* SMCs, i.e., as the "lawful regularities" induced by the agent's embodiment (see Section 1.1.3). To take one concrete example: let us come back to the case of sensory substitution devices, where, e.g., vision can be partially "restored" through tactile stimulation induced by a head-mounted camera (Eagleman et al., 2023). SMC theory claims that this is because vision — and visual experience — is ultimately defined not by a specific set of neural pathways and corresponding dynamics, but by the specific "structure of changes" induced by eye and body movements on retinal activations (O'Regan et al., 2001). This claim could be investigated by comparing the generalised symmetries of the sensorimotor interface defined by body movement and retinal stimulation on the one hand, and, on the other hand, by body movement and tactile stimulation induced by the sensory substitution device on the tongue. It is indeed reasonable to expect that this kind of interactive setting can be accurately modeled by the general formalisms of perception-action loop models (Langer et al., 2024; Tishby et al., 2011) and/or input-output processes (Barnett et al., 2015; Rosas et al., 2025; Shalizi, 2001) — so that the corresponding generalised symmetries could be captured by extensions of the tools developed in this thesis. More generally, these abstract symmetries of the sensorimotor interface could provide novel "sensorimotor correlates" of perceptual experience — thus contributing to the operationalisation of ideas at the heart and foundation of SMC theory (O'Regan et al., 2001).

[1] Here, the "input process" is that of agent's actions, transformed by a "sensorimotor interface" channel into an "output process" of resulting sensory inputs.

At this stage, however, such a contribution is still out of reach. First, a substantial amount of mathematical work is required to make the formal framework more mature. But it is of course also necessary to design appropriate data-processing tools to explore these mathematical objects in real-world data. Both these aspects would also be required for implementing the ideas presented here in artificial agents. While I have outlined above possible directions for future work on the formal side, one could address the computational side, e.g., by taking inspiration in the abundant literature on machine learning methods to approximate solutions to the IB problem and its generalisations — see (Hu et al., 2024) for a recent review.

5.2 Invariant/equivariant dynamics emerging from behaviour

While the information-theoretic characterisation of ergodic components from Chapter 3 is relevant to the discovery of abstract symmetries, the minimal class-pose parametrisation framework as a whole opens a new direction to investigate the structure induced by an agent's own actions, in particular its closed-loop behaviour.

For that purpose, it would first be necessary to make the formal framework developed in Chapter 3 more mature, in particular: (*i*) to address the technical subtleties arising with minimal joinings in the non-finite case (especially for an uncountable number of ergodic components); (*ii*) to complete the information-theoretic formalisation of poses in the finite case and both classes and poses in the non-finite case; and (*iii*) to use these information-theoretic characterisations to derive algorithms discovering minimal class-pose parametrisation (see Section 3.7). This framework could then be used to investigate proof-of-concept scenarios with agents in a fully observed environment. A first step could here be to investigate the class-pose structure that emerges from policies that uniformly sample the actions of a group or group-like object: e.g., the pseudogroup of changes of perspective that identifies contiguous surfaces from the structure of light rays in (Tsao et al., 2022).

However, to become relevant to real-world embodied agents, our framework needs to be extended to allow for the sensory outcome of actions to depend on the full sensorimotor history. Relevant formalisms include here Partially Observable MDPs, causal Bayesian networks models of the perception-action loop (Langer et al., 2024; Tishby et al., 2011), or input-output processes and their ϵ-transducers (Barnett et al., 2015; Marzen, 2025; Rosas et al., 2025; Shalizi, 2001).

Moreover, let us recall that our novel notions of class and pose crucially depend on the policy. Of course, this feature becomes particularly interesting once we focus on *behaviourally relevant* policies. Such behavioural relevancy could be defined in many different ways, taking inspiration from, e.g., the field of *intrinsic motivation* (Aubret et al., 2023; Forestier et al., 2022; Salge et al., 2014; Volpi et al., 2023) or normativity perspectives on sensorimotor autonomy (Barandiaran, 2017; Barandiaran et al., 2014; Barrett et al., 2025). Alternatively, instead of having classes and poses that depend on the policy but not the converse, one could also consider cases where the class-pose structure and the policy co-determine each other — where, e.g., the policy is required to make the corresponding classes capture a specific kind of invariant. This might lead to formalisms capturing not only how perceptual structure emerges from information trade-offs, but also how information-trade-offs can *induce the production of structure* in the sensorimotor loop. Note that this would bear some resemblance with the *active inference* framework (Pezzulo et al., 2024), to the extent that a specific requirement on the desired perceptual structure (i.e., here, on the class-pose structure) would induce the agent's behaviour that realises these perceptual requirements — similarly as in active inference, sensory predictions can induce actions that makes these predictions "self-fulfilling". However, in our case, the perceptual requirements could be much different from the optimisation of variational free energy (Pezzulo et al., 2024).

As pointed out in Section 3.8, the self-generated aspect of minimal joinings resonates with the inside-out approach to brain dynamics (Buzsáki et al., 2019), while the fact that classes are ergodic components suggests links with the notion of percepts as attractors from Closed-Loop Perception theory (Ahissar et al., 2016). Future work could explore these links further, including by developing new variations of the formalism that integrates it better with these frameworks.

It is also crucial to study concrete examples of how the class-pose structure can be used by embodied agents for purposeful behaviour. I propose that the stochastic dynamics defined by the minimal joining of ergodic components is a possible formalisation of the *skillful exercise* (O'Regan et al., 2001) of a sensorimotor contingency, i.e., of the practice of the *know-how* (O'Regan et al., 2001) of how a family of similar percepts behaves under a certain behaviour. For example, "skillful writing" would mean being able to "run" the corresponding minimal joining, with any pen sufficiently similar to those one has previously encountered. The fact that all pens are not used exactly the same way, and do not "feel" exactly the same, would here correspond to the fact that the minimal joining corresponding to the "skillful writing" activity is not an isomorphic joining: it is an "abstract" dynamical system that can be enacted by "projecting" it on any pen (through an appropriate *marginalisation map*, see Definition 3.5.6), but is projected in different ways with different pens. Importantly, here, to each pen corresponds a percept (i.e., a class defined as an ergodic component), but the percept is defined by the way the pen is used for "skillful writing": it would be a different percept if, e.g., it was used for drawing, or with the other hand (i.e., if the policy was different). Future work could test whether minimal class-pose parametrisations (in their current MDP form, or after appropriate generalisation) can indeed formalise this kind of intuitions in simple agents, first in a proof-of-concept setting.

The latter formal interpretation of skillful exercise of SMCs is relevant to the question of the role of ongoing brain dynamics in sensorimotor perception (see Section 1.1.3). In the interpretation outlined above, the internal dynamics (i.e., the minimal joining) emerge from concretely enacted behaviours (i.e., from the dynamics of the ergodic components on the original state-space), but once the minimal joining is formed, its dynamics acquire an automomy w.r.t. the sensorimotor level: they can unfold "internally" without being actually enacted (i.e., without being projected back on the original state-space with the marginalisation maps). This might provide an important step forward in the formalisation of ideas from the inside-out approach on brain dynamics, which insists on the internalisation of sensorimotor dynamics by brain circuits (Buzsáki et al., 2019). However, this setting does not address how the unfolding of brain dynamics on the time-scale of perception itself is often inseparable from that of the dynamics of the sensorimotor interface — as is the case, e.g., with eye movements in vision, see Section 1.1.3, and as is exemplified with a minimal dynamical model in (Aguilera et al., 2013). On the long term, it is thus necessary to go further and design formalisms that capture how "minimal joining-like" objects — thought here as emerging on the time-scale of learning and development — can be *coupled*, on the time-scale of perception, with the sensorimotor interface.

Another possible avenue for future work would be to analyse further the structure of the pose coordinate, and how it varies when the policy varies. E.g., can the pose coordinate be factorised into distinct coordinates that correspond to distinct coordinates of the action space — yielding an MDP generalisation of disentanglement (Higgins et al., 2018)? Can we analyse compositionality relations between different policies by establishing some kind of algebraic relations between their corresponding pose coordinates — thus, e.g., yielding concepts analogous to integers' decomposition in prime numbers, but where integers are replaced by an agent's possible behaviours? Would these algebraic relationships have information-theoretic characterisations? Investigating these questions might allow one to investigate in detail the structure of a given SMC.

At the formal level, it is worth noting that while Chapter 3 has focused on minimal joinings of a given family (of stationary MDPs), a similar notion of *maximal common factor* could be defined just as naturally — as a maximal element for a relation that would be defined similarly as we defined the "j-factor" relation, but with joinings replaced by common factors. This hints at a possible theory of the *lattice structure* of stationary MDPs — or more generally, of agentic systems formalised in a given mathematical category. Recent work has already started to investigate the lattice structure of the factors of a given stochastic process (Rosas et al., 2024):[2] I am here proposing to extend it by (*i*) investigating the relations between *arbitrary* processes, through the notions of (maximal) common factor and (minimal) joining, (*ii*) considering stochastic processes with actions — an example of which are stationary MDPs, but we could also consider, e.g., the more general structure of input-output processes (Barnett et al., 2015; Rosas et al., 2025), and (*iii*) investigating the *soft* versions of all these objects, by varying the trade-off parameters of characterisations of these structures with well-chosen information-theoretic multi-objective optimisation problems. Such a formalism could then be used to analyse the lattice structure of general agentic systems' dynamics. If this is possible, it might yield precious new tools to investigate the idea that perception — as well as other activities typically referred to as cognitive — emerge from a subtle intertwining of the agent's history of sensorimotor interactions, and its irreducibly ongoing behaviour.

[2] Here, what I call factor corresponds to what is called *computationally closed coarse-graining* in (Rosas et al., 2024).

Appendix A

Appendix for Chapter 1

Here, we collect basic definitions about groups and group actions.

Definition A.0.1. A *group* is a set $\mathcal{G}$ equipped with a binary operation

$$\begin{aligned} \mathcal{G} \times \mathcal{G} &\to \mathcal{G} \\ (g, g') &\mapsto gg' \end{aligned}$$

which satisfies the following properties:

- Associativity: $(gg')g'' = g(g'g'')$ for all $g, g', g'' \in \mathcal{G}$.
- Identity element: There exists $e \in \mathcal{G}$ such that $eg = ge = g$ for all $g \in \mathcal{G}$. This element is automatically unique.
- Invertibility: For all $g \in \mathcal{G}$, there exists g^{-1} such that $gg^{-1} = g^{-1}g = e$, where e is the identity element.

Definition A.0.2. The action of a group $\mathcal{G}$ on a set $\mathcal{A}$, i.e., is a function[1]

$$\begin{aligned} \rho \ : \ \mathcal{G} \times \mathcal{A} &\to \mathcal{A} \\ (g, a) &\mapsto \rho_g(a) \ := g \cdot a, \end{aligned}$$

such that $(gg') \cdot a = g \cdot (g' \cdot a)$ and $e \cdot a = a$ for all $a \in \mathcal{A}$ and $g, g' \in \mathcal{G}$, where e is the identity element of the group $\mathcal{G}$. For a subset $F \subseteq \mathcal{A}$, we write $g \cdot F := \{g \cdot a, \ a \in F\}$. A subset $F \subseteq \mathcal{A}$ is *invariant* if $g \cdot F = F$ for all $g \in \mathcal{G}$.

Let us now fix the action of a group $\mathcal{G}$ on a set $\mathcal{A}$.

Definition A.0.3. The *orbit* of a point a is the set $[a]$ of all other points in $\mathcal{A}$ that can be reached from a by transforming it with a group element $g \in \mathcal{G}$, i.e.,

$$[a] := \{g \cdot a, \ g \in \mathcal{G}\} \ \subseteq \ \mathcal{A}.$$

It can be easily verified that the set of all orbits define a partition of the set $\mathcal{A}$. The *projection on orbits* is then the function

$$\begin{aligned} \mathcal{A} &\to \mathcal{A}/\mathcal{G} \\ a &\mapsto [a], \end{aligned}$$

which to each point associates its orbit, where we denote by $\mathcal{A}/\mathcal{G}$ the set of all orbits, also called the *quotient* of the set $\mathcal{A}$ w.r.t. the action of $\mathcal{G}$.

The following property of the partition in orbits underlies the duality between symmetry and information parsimony that we will explore in this thesis:

[1] Both notations $\rho_g(a)$ and $g \cdot a$ will be useful.

Proposition A.0.4. *The partition in orbits is the finest partition of $\mathcal{A}$ into invariant subsets: i.e., for any other partition $\{\mathcal{A}_i\}_{i\in\mathcal{I}}$ into invariant subsets and all $i \in \mathcal{I}$, there exists $a \in \mathcal{A}$ such that $[a] \subseteq \mathcal{A}_i$.*

Proof. Let $i \in \mathcal{I}$ and $a \in \mathcal{A}_i$. As $\mathcal{A}_i$ is invariant, we have $g \cdot a \subseteq \mathcal{A}_i$ for all $g \in \mathcal{G}$, i.e., $[a] \in \mathcal{A}_i$. □

Let us now give a short overview of the notions of *invariance* and *equivariance*. These terms can be used to refer to different mathematical objects (and they will be in these thesis).

"Invariance" generally denotes, in short, a situation where some transformation(s) of a system leave(s) something unchanged in that system. For instance, assume that $\mathcal{A}$, $\mathcal{B}$ are sets, f is a map from $\mathcal{A}$ to $\mathcal{B}$, and ρ is the action a group $\mathcal{G}$ on $\mathcal{A}$ such that $f \circ \rho_g = f$ for all $g \in \mathcal{G}$, where the symbol $\circ$ denotes composition. Then we can alternatively say (i) that ρ is made of *invariances* of the map f (here the focus is on understanding f through ρ), or (ii) that f captures an invariant feature, or invariant for short, of the group action ρ (here the focus is on understanding ρ through f). In this thesis, we are interested in two kinds of invariances: channel invariances, which will be defined in Chapter 2 and correspond to (stochastic versions of) the point of view (i), and projections on obits, which correspond to the point of view (ii).

On the other hand, "equivariance" generally denotes a commutation property of the form $f \circ \rho^{\mathcal{A}}_g = \rho^{\mathcal{B}}_g \circ f$ for all $g \in \mathcal{G}$, where f is a map from a set $\mathcal{A}$ to a set $\mathcal{B}$, while $\rho^{\mathcal{A}}$ and $\rho^{\mathcal{B}}$ denote actions of a group $\mathcal{G}$ on resp. $\mathcal{A}$ and $\mathcal{B}$. Similarly as for invariances, the focus can either be on understanding f through the pair of actions $(\rho^{\mathcal{A}}, \rho^{\mathcal{B}})$, or on understanding the relationship between $\rho^{\mathcal{A}}$ and $\rho^{\mathcal{B}}$ through the map f. In this thesis, we will be interested in two kinds of equivariances: channel equivariances, which will be defined in Chapter 2 and correspond to (stochastic versions of) the former point of view, and *factors* of group actions, which correspond to the latter point of view and will be important in Chapter 3. As the explicit definition of factors and the related notion of isomorphism facilitates the presentation of this thesis' contributions in Section 1.3.2, we include it here.

Definition A.0.5. Let ρ^1, ρ^2 be actions of resp. groups $\mathcal{G}^1$ and $\mathcal{G}^2$ on resp. sets $\mathcal{X}^1$ and $\mathcal{X}^2$. The group action ρ^2 is a *factor* of ρ^1 with *factor map* $\phi : \mathcal{X}^1 \to \mathcal{X}^2$ if ϕ is surjective and, for all $g \in \mathcal{G}$, we have $\phi \circ \rho^1_g = \rho^2_g \circ \phi$: i.e., the following diagram is commutative:

$$\begin{array}{ccc} \mathcal{X}^1 & \xrightarrow{\rho^1_g} & \mathcal{X}^1 \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\phi} \\ \mathcal{X}^2 & \xrightarrow{\rho^2_g} & \mathcal{X}^2 \end{array} \tag{A.0.1}$$

The group actions are *isomorphic* if ϕ is bijective.

Intuitively, the group action ρ^2 is a factor of the group action ρ^1 if it is a "subsystem" of the "dynamical system" defined by the action of ρ^1 on $\mathcal{X}^1$; and two group actions are isomorphic if they are "the same", as one can be transformed into the other using the bijection ϕ.

Appendix B

Appendix for Chapter 2

B.1 A general rate-distortion theorem

Here, we prove a general theorem from which we will derive both Theorems 2.2.3 and 2.3.1. Before that, though, we need to introduce the notion of preimages through channels.

B.1.1 Preimages through stochastic channels

Definition B.1.1. Let $\mathcal{A}$, $\mathcal{B}$ be finite sets and $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$. The *preimage* of a subset $E \subseteq \mathcal{B}$ through the channel γ is the set

$$\gamma^{-1}(E) := \{a \in \mathcal{A} : \ \gamma(E|a) > 0\}. \tag{B.1.1}$$

Note that if $\gamma := \gamma_f$ is defined by a deterministic function $f : \mathcal{A} \to \mathcal{B}$, then for all $E \subseteq \mathcal{B}$, the preimage $\gamma_f^{-1}(E)$ through the channel γ_f coincides with the preimage $f^{-1}(E)$ through the function f, in the usual sense. However, while for deterministic functions, the preimage of E is made of elements of $\mathcal{A}$ that *must* be sent in E, the requirement is weaker for our generilsation to stochastic channels, as we consider all elements of $\mathcal{A}$ that *can* be sent in E (i.e., with non-zero probability). Moreover, some properties of preimages through functions generalise well to the stochastic case:

Lemma B.1.2. *Let $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$ be finite sets, and $\gamma \in \mathscr{K}(\mathcal{A}_1, \mathcal{A}_2)$, $\gamma' \in \mathscr{K}(\mathcal{A}_2, \mathcal{A}_3)$. Then:*

(i) If $E_n \subseteq \mathcal{A}_2$ for all $n \in \mathbb{N}$, then

$$\gamma^{-1}\left(\bigcup_{n\in\mathbb{N}} E_n\right) = \bigcup_{n\in\mathbb{N}} \gamma^{-1}(E_n).$$

(ii) If $E \subseteq \mathcal{A}_3$, then

$$(\gamma' \circ \gamma)^{-1}(E) = \gamma^{-1}\left((\gamma')^{-1}(E)\right).$$

Proof. (i). On the one hand, for all $a \in \mathcal{A}_1$,

$$\exists n \in \mathbb{N} : \ \gamma(E_n|a) > 0 \quad \Rightarrow \quad \gamma\left(\bigcup_{n\in\mathbb{N}} E_n|a\right) \geq \gamma(E_n|a) > 0,$$

so that $\bigcup_{n\in\mathbb{N}} \gamma^{-1}(E_n) \subseteq \gamma^{-1}\left(\bigcup_{n\in\mathbb{N}} E_n\right)$. Conversely, if $a \in \mathcal{A}_1$ is such that $\gamma\left(\bigcup_{n\in\mathbb{N}} E_n|a\right) > 0$, then

$$\sum_{n\in\mathbb{N}} \gamma(E_n|a) \geq \gamma(\bigcup_{n\in\mathbb{N}} E_n|a) > 0,$$

which implies the existence of $n \in \mathbb{N}$ such that $\gamma(E_n|a) > 0$, i.e., such that $a \in \gamma^{-1}(E_n)$.

(ii). For all $a_1 \in \mathcal{A}_1$,

$$\begin{aligned} a_1 \in (\gamma' \circ \gamma)^{-1}(E) &\Leftrightarrow (\gamma \circ \gamma')(E|a_1) > 0 \\ &\Leftrightarrow \sum_{a_2 \in \mathcal{A}_2} \gamma'(E|a_2)\gamma(a_2|a_1) > 0 \\ &\Leftrightarrow \exists a_2 \in \mathcal{A}_2 \ : \ \gamma'(E|a_2) > 0 \text{ and } \gamma(a_2|a_1) > 0 \\ &\Leftrightarrow \gamma\left(\left\{a_2 \in \mathcal{A}_2 \ : \ \gamma'(E|a_2) > 0\right\}|a_1\right) > 0 \\ &\Leftrightarrow \gamma\left((\gamma')^{-1}(E)|a_1\right) > 0 \\ &\Leftrightarrow a_1 \in \gamma^{-1}\left((\gamma')^{-1}(E)\right) > 0. \end{aligned}$$

□

B.1.2 Main result and its proof

Let $\mathcal{A}$ finite and $\mu \in \Delta_{\mathcal{A}}$ full support. For all $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$, we define the distribution $q_\kappa \in \Delta_{\mathcal{A}\times\mathcal{T}}$ by $q_\kappa(a,t) := \mu(a)\kappa(t|a)$ for all $a \in \mathcal{A}$, $t \in \mathcal{T}$, and denote by $I_\kappa(A;T)$ the corresponding mutual information. We then consider the rate-distortion problem

$$R_D(\lambda) := \operatorname*{arg\,min}_{\substack{\kappa \in \mathscr{K}(\mathcal{A},\mathcal{T}) \\ D(\kappa) \geq \lambda}} I_\kappa(A;T), \tag{B.1.2}$$

where $\lambda \in [0, \Lambda]$ for some $\Lambda > 0$, while $D : \mathscr{K}(\mathcal{A}, \mathcal{T}) \to [0, \Lambda]$ is continuous, and seen here as a "distortion" function defined on a set of "compression channels" $\mathscr{K}(\mathcal{A}, \mathcal{T})$.

We also consider a surjective function $\mathsf{pr} : \mathcal{A} \to C$ where C is finite. It defines a unique extension of q_κ to $\Delta_{\mathcal{A}\times\mathcal{T}\times C}$ satisfying the Markov chain $T - A - C$, which we still denote by q_κ. Explicitly, for for $a \in \mathcal{A}$, $c \in C$, $t \in \mathcal{T}$,

$$q_\kappa(a,t,c) := \mu(a)\kappa(t|a)\delta_{\mathsf{pr}(a)=c}.$$

As pr is surjective and μ is full-support, the push-forward $\mathsf{pr}\cdot\mu \in \Delta_C$ is also full-support. Thus the corresponding conditional distribution $\epsilon \in \mathscr{K}(C, \mathcal{A})$ of A given C is uniquely defined, for all $a \in \mathcal{A}, c \in C$, by

$$\epsilon(a|c) := \frac{q_\kappa(a,c)}{q_\kappa(c)} = \frac{\mu(a)}{(\mathsf{pr}\cdot\mu)(c)}\,\delta_{\mathsf{pr}(a)=c}.$$

Note that the second equality above clearly implies

$$\mathsf{pr} \circ \epsilon = \mathsf{Id}_C, \tag{B.1.3}$$

which will be useful below. Moreover, the conditional probability $\kappa_{\mathsf{pr}} \in \mathscr{K}(C, \mathcal{T})$ of T given C is uniquely defined by

$$\kappa_{\mathsf{pr}} := \kappa \circ \epsilon \in \mathscr{K}(C, \mathcal{T}), \tag{B.1.4}$$

i.e., explicitly, for all $c \in C$, $t \in \mathcal{T}$,

$$\kappa_{\mathsf{pr}}(t|c) := \frac{\sum_{a \in \mathsf{pr}^{-1}(c)} \kappa(t|a)\mu(a)}{(\mathsf{pr}\cdot\mu)(c)}. \tag{B.1.5}$$

Informally, κ_{pr} can be seen as the "probabilistic quotient" of κ w.r.t. the partition on $\mathcal{A}$ defined by the preimages $(\mathsf{pr}^{-1}(c))_{c\in C}$, with weights given by $\mu \in \Delta_{\mathcal{A}}$. We then define

$$\bar{\kappa} := \kappa_{\mathsf{pr}} \circ \mathsf{pr} = \kappa \circ \epsilon \circ \mathsf{pr} \in \mathscr{K}(\mathcal{A}, \mathcal{T}), \tag{B.1.6}$$

i.e., explicitly,

$$\bar{\kappa}(t|a) := \sum_{c\in C} \kappa_{\mathsf{pr}}(t|c)\delta_{\mathsf{pr}(a)=c}. \tag{B.1.7}$$

Informally, $\bar{\kappa}$ can be seen as the "projection, w.r.t. μ, of κ on the channels that factorise through pr".[1] We are now ready to state the main result of this section:

Theorem B.1.3. *Let $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$. The following are equivalent:*

$(P1)$ *For all $t \in \mathrm{supp}(\kappa \cdot \mu)$, there exists $c \in C$ such that $\kappa^{-1}(t) \subseteq \mathsf{pr}^{-1}(c)$.*

$(P2)$ *There exists a function $h : \mathcal{T} \to C$ such that $\mathsf{pr} = h \circ \kappa$.*

Moreover, assume that

(a) $D(\kappa) = D(\bar{\kappa})$,

(b) $D(\kappa) = \Lambda$ *if and only if* $(P1)$ *holds (or equivalently, if and only if* $(P2)$ *holds).*

Then

(i) For $\lambda = \Lambda$, a channel $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$ is a solution to the problem (B.1.2) if and only if there exists a congruent channel $\iota \in \mathscr{K}_{\mathsf{cong}}(C, \mathcal{T})$ such that $\kappa = \iota \circ \mathsf{pr}$.

(ii) For all $0 \leq \lambda \leq \Lambda$, all solutions $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$ to the problem (B.1.2) satisfy $\kappa = \gamma \circ \mathsf{pr}$, for some $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{T})$.

The remaining of this section consists of the proof of Theorem B.1.3.

Lemma B.1.4. *For all $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$,*

(i) $\bar{\kappa} \cdot \mu = \kappa \cdot \mu$.

(ii) $I_{\bar{\kappa}}(A;T) \leq I_{\kappa}(A;T)$, *where equality holds if and only if* $\bar{\kappa}(T|A) = \kappa$.

Proof. (i). Under the distribution $q_{\kappa}(A,T)$, the deterministic channel defined by pr is the conditional probability of C given A, while ϵ is the conditional probability of A given C, where the marginal on A is μ. Thus, by a straightforward computation, we have $(\epsilon \circ \mathsf{pr}) \cdot \mu = \mu$. Using the definition (B.1.6) of $\bar{\kappa}$, this implies that

$$\bar{\kappa} \cdot \mu = (\kappa \circ \epsilon \circ \mathsf{pr}) \cdot \mu = \kappa \cdot \big((\epsilon \circ \mathsf{pr}) \cdot \mu\big) = \kappa \cdot \mu.$$

(ii). We have

$$\begin{aligned} I_{\kappa}(A;T) &= \sum_{a\in\mathcal{A}, t\in\mathrm{supp}(\kappa\cdot\mu)} \mu(a)\kappa(t|a) \log\left(\frac{\kappa(t|a)}{(\kappa\cdot\mu)(t)}\right) \\ &= \sum_{t\in\mathrm{supp}(\kappa\cdot\mu)} \sum_{c\in C} \sum_{a\in\mathsf{pr}^{-1}(c)} \mu(a)\kappa(t|a) \log\left(\frac{\kappa(t|a)}{(\kappa\cdot\mu)(t)}\right). \end{aligned}$$

[1] Maybe this intuition can be given an information-geometric meaning, but this will not be necessary here.

From the log-sum inequality (Csiszár et al., 2011), for $t \in \mathrm{supp}(\kappa \cdot \mu)$, $c \in \mathcal{C}$,

$$\begin{aligned}
\sum_{a \in \mathsf{pr}^{-1}(c)} \mu(a)\kappa(t|a) \log\left(\frac{\kappa(t|a)}{(\kappa \cdot \mu)(t)}\right) &= \sum_{a \in \mathsf{pr}^{-1}(c)} \mu(a)\kappa(t|a) \log\left(\frac{\mu(a)\kappa(t|a)}{\mu(a)(\kappa \cdot \mu)(t)}\right) \\
&\geq \left(\sum_{a \in \mathsf{pr}^{-1}(c)} \mu(a)\kappa(t|a)\right) \log\left(\frac{\sum_{a \in \mathsf{pr}^{-1}(c)} \mu(a)\kappa(t|a)}{\sum_{a \in \mathsf{pr}^{-1}(c)} \mu(a)(\kappa \cdot \mu)(t)}\right) \qquad \text{(B.1.8)} \\
&= (\mathsf{pr} \cdot \mu)(c)\kappa_{\mathsf{pr}}(t|c) \log\left(\frac{\kappa_{\mathsf{pr}}(t|c)}{(\kappa \cdot \mu)(t)}\right) \qquad \text{(B.1.9)} \\
&= \sum_{a \in \mathsf{pr}^{-1}(c)} \mu(a)\bar{\kappa}(t|a) \log\left(\frac{\bar{\kappa}(t|a)}{(\kappa \cdot \mu)(t)}\right) \qquad \text{(B.1.10)} \\
&= \sum_{a \in \mathsf{pr}^{-1}(c)} \mu(a)\bar{\kappa}(t|a) \log\left(\frac{\bar{\kappa}(t|a)}{(\bar{\kappa} \cdot \mu)(t)}\right), \qquad \text{(B.1.11)}
\end{aligned}$$

where line (B.1.9) uses the definition (B.1.5) of κ_{pr}; line (B.1.10) uses the definition (B.1.7) of $\bar{\kappa}$; and line (B.1.11) uses point (i) proven above. Moreover, from the equality case of the log-sum inequality (Csiszár et al., 2011), equality holds in (B.1.8) if and only if $\frac{\kappa(t|a)\mu(a)}{(\kappa \cdot \mu)(t)\mu(a)}$ is constant for $a \in \mathsf{pr}^{-1}(c)$, i.e., if and only if $\kappa(t|a)$ is constant for $a \in \mathsf{pr}^{-1}(c)$. Thus, summing (B.1.8) over $c \in C$ and $t \in \mathrm{supp}(\kappa \cdot \mu)$, we get

$$\begin{aligned}
I_\kappa(A;T) &\geq \sum_{t \in \mathrm{supp}(\kappa \cdot \mu)} \sum_{c \in \mathcal{C}} \sum_{a \in \mathsf{pr}^{-1}(c)} \mu(a)\bar{\kappa}(t|a) \log\left(\frac{\bar{\kappa}(t|a)}{(\bar{\kappa} \cdot \mu)(t)}\right) \\
&= I_{\bar{\kappa}}(A;T),
\end{aligned}$$

with equality if and only if for all $t \in \mathrm{supp}(\kappa \cdot \mu)$ and all c, the quantity $\kappa(t|a)$ is constant for $a \in \mathsf{pr}^{-1}(c)$ — which means more precisely that for all $a \in \mathsf{pr}^{-1}(c)$, we have $\kappa(t|a) = \kappa_{\mathsf{pr}}(t|c)$ (use equation (B.1.5)). Using equation (B.1.7) thus yields that there is equality if and only if

$$\forall a \in \mathcal{A},\ \forall t \in \mathrm{supp}(\kappa \cdot \mu), \quad \kappa(t|a) = \bar{\kappa}(t|a). \qquad \text{(B.1.12)}$$

But from point (i) above, we have $\mathrm{supp}(\kappa \cdot \mu) = \mathrm{supp}(\bar{\kappa} \cdot \mu)$. This implies that condition (B.1.12) is equivalent to $\kappa = \bar{\kappa}$. □

Proposition B.1.5. *For all $0 \leq \lambda \leq \Lambda$, every solution $\kappa \in R_D(\lambda)$ of (B.1.2) satisfies $\kappa = \bar{\kappa}$, i.e., $\kappa = \kappa_{\mathsf{pr}} \circ \mathsf{pr}$.*

Proof. For all $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$, assumption (a) in Theorem B.1.3 shows that $\bar{\kappa}$ achieves the same value $D(\bar{\kappa}) = D(\kappa)$ of the constraint in the rate-distortion problem (B.1.2); while point (ii) in Lemma B.1.4 shows that the corresponding value $I_{\bar{\kappa}}(A;T)$ of the target function is lower than or equal to $I_\kappa(A;T)$, with equality if and only if $\bar{\kappa} = \kappa$. In particular, if κ solves the IB problem, then we must have $\kappa = \bar{\kappa}$, i.e., $\kappa = \kappa_{\mathsf{pr}} \circ \mathsf{pr}$. □

This proves point (ii) in Theorem B.1.3.

Lemma B.1.6. *For all $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$, the following are equivalent:*

$(P1)$ *There exists a function $h : \mathcal{T} \to \mathcal{C}$ such that $\mathsf{pr} = h \circ \kappa$.*

$(P2)$ *For all $t \in \mathrm{supp}(\kappa \cdot \mu)$, there exists some $c \in \mathcal{C}$ such that $\kappa^{-1}(t) \subseteq \mathsf{pr}^{-1}(c)$.*

$(P3)$ *The channel $\kappa_{\mathsf{pr}} \in \mathscr{K}(\mathcal{C}, \mathcal{T})$ defined in (B.1.4) is congruent (see Definition 2.2.2).*

Proof. $(P1) \Rightarrow (P2)$. For all $c \in \mathcal{C}$,

$$\begin{aligned}\mathsf{pr}^{-1}(c) &= (h \circ \kappa)^{-1}(c) \\ &= \kappa^{-1}\left(h^{-1}(c)\right) \\ &= \bigcup_{t \in h^{-1}(c)} \kappa^{-1}(t),\end{aligned}$$

where we used Lemmma B.1.2. Thus for all $t \in \mathcal{T}$, we have $\kappa^{-1}(t) \subseteq \mathsf{pr}^{-1}(h(t))$.

$(P2) \Rightarrow (P3)$. As $\{\mathsf{pr}^{-1}(c)\}_{c \in \mathcal{C}}$ is a partition, for each $t \in \mathrm{supp}(\kappa \cdot \mu)$ we have $\kappa^{-1}(t) \neq \emptyset$ and there exists a *unique* $c \in \mathcal{C}$ such that $\kappa^{-1}(t) \subseteq \mathsf{pr}^{-1}(c)$. Define $h(t)$ as this unique c if $t \in \mathrm{supp}(\kappa \cdot \mu)$, and arbitrarily otherwise. Then, using that μ is full-support, we obtain $\mathsf{pr} = h \circ \kappa$.

$(P2) \Leftrightarrow (P3)$. For all function $h : \mathcal{T} \to \mathcal{C}$:

$$\begin{aligned} h \circ \kappa_{\mathsf{pr}} = \mathsf{Id}_{\mathcal{C}} \quad &\Leftrightarrow \quad h \circ \kappa \circ \epsilon = \mathsf{Id}_{\mathcal{C}} \\ &\Rightarrow \quad h \circ \kappa = \mathsf{pr} \\ &\Rightarrow \quad h \circ \kappa \circ \epsilon = \mathsf{Id}_{\mathcal{C}} \\ &\Leftrightarrow \quad h \circ \kappa_{\mathsf{pr}} = \mathsf{Id}_{\mathcal{C}}, \end{aligned}$$

where we used the definition (B.1.4) of κ_{pr} in the first and last line; while the second line composes each side of the equality by pr on the right and uses equation (B.1.3), and the third line composes each side by ϵ on the right and uses equation (B.1.3) again. Thus we have $h \circ \kappa_{\mathsf{pr}} = \mathsf{Id}_{\mathcal{C}} \Leftrightarrow h \circ \kappa = \mathsf{pr}$ for all $h : \mathcal{T} \to \mathcal{C}$, which yields the result. □

Lemma B.1.6 proves the first part of Theorem B.1.3. It remain to show point (i), i.e., that

$$R_D(\Lambda) = \mathscr{K}_{\mathsf{pr},\mathsf{cong}} := \left\{\iota \circ \mathsf{pr}, \quad \iota \in \mathscr{K}_{\mathsf{cong}}(\mathcal{C}, \mathcal{T})\right\}.$$

Let us first prove the inclusion $R_D(\Lambda) \subseteq \mathscr{K}_{\mathsf{pr},\mathsf{cong}}$. Fix a solution $\kappa \in R_D(\Lambda)$. Proposition B.1.5 proves that $\kappa = \kappa_{\mathsf{pr}} \circ \mathsf{pr}$. But because we must have $D(\kappa) = \Lambda$, combining assumption (b) in Theorem B.1.3 with Lemma B.1.6 yields that κ_{pr} is here congruent. Let us now prove the converse inclusion $\mathscr{K}_{\mathsf{pr},\mathsf{cong}} \subseteq R_D(\Lambda)$.

Lemma B.1.7. *For all $\kappa \in \mathscr{K}_{\mathsf{pr},\mathsf{cong}}$, the channel κ_{pr} is congruent, and we have $D(\kappa) = \Lambda$ and $I_\kappa(A;T) = H(\mathsf{pr}(A))$.*

Proof. Let $\kappa = \iota \circ \mathsf{pr}$, with $\iota \in \mathscr{K}_{\mathsf{cong}}(\mathcal{C}, \mathcal{T})$. We must have $\iota = \kappa_{\mathsf{pr}}$, as for all $c \in \mathcal{C}$, $t \in \mathcal{T}$, from (B.1.5),

$$\kappa_{\mathsf{pr}}(t|c) = \sum_{a \in \mathsf{pr}^{-1}(c)} \frac{\mu(a)\kappa(t|a)}{\mu(c)} = \sum_{a \in \mathsf{pr}^{-1}(c)} \frac{\mu(a)\iota(t|\mathsf{pr}(a))}{\mu(c)} = \iota(t|c).$$

Thus κ_{pr} is congruent, so that from assumption (b) in Theorem B.1.3 and Lemma B.1.6, we have $D(\kappa) = \Lambda$. Let now $f : \mathcal{T} \to \mathcal{C}$ such that $f \circ \iota = \mathsf{Id}_{\mathcal{C}}$. Then $f \circ \kappa = f \circ \iota \circ \mathsf{pr} = \mathsf{pr}$. Thus, using the data-processing inequality (Cover et al., 2009),

$$I_\kappa(A;T) \geq I_\kappa(A; f(T)) = I(A; \mathsf{pr}(A)) = H(\mathsf{pr}(A)),$$

where the last equality holds because pr is deterministic. On the other hand, $\kappa = \iota \circ \mathsf{pr}$, implies the Markov chain $A - \mathsf{pr}(A) - T$, and therefore $I_\kappa(A;T) \leq I(A; \mathsf{pr}(A)) = H(\mathsf{pr}(A))$. □

Now, as D is continuous, $D^{-1}([\Lambda, +\infty[) = D^{-1}(\Lambda)$ is compact, as it is closed in the compact set $\mathscr{K}(\mathcal{A}, \mathcal{T})$. Therefore problem (B.1.2) is defined as the minimisation of a continuous function on a compact domain, and has at least one solution, say κ_*, which we know belongs to $\mathscr{K}_{\mathsf{pr,cong}}$ from the the inclusion $R_D(\Lambda) \subseteq \mathscr{K}_{\mathsf{pr,cong}}$ that we already proved. But Lemma B.1.7 then implies that for all $\kappa \in \mathscr{K}_{\mathsf{pr,cong}}$, we have $D(\kappa) = D(\kappa_*)$ and $I_\kappa(A;T) = I_{\kappa_*}(A;T)$. Thus any $\kappa \in \mathscr{K}_{\mathsf{pr,cong}}$ must also be a solution, i.e., $\kappa \in R_D(\Lambda)$. This ends the proof.

B.2 Proof of Theorem 2.2.3

Let us set $\mathcal{A} := \mathcal{X}$, $\mu := \mu(X,Y)$, $D(\kappa) := I_\kappa(T,Y)$ for all $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$, and let $\mathsf{pr} : \mathcal{X} \to \mathcal{C}$ be the projection on the equivalence classes $(\mathcal{X}^c)_{c\in\mathcal{C}}$ of the relation $\sim_{\mathcal{X}}$ defined in (2.2.3). Then $R_D(\lambda) = \mathrm{IB}(\lambda)$ for all $0 \le \lambda \le \Lambda := I(X;Y)$ (where $R_D(\lambda)$ and $\mathrm{IB}(\lambda)$ are resp. defined in (B.1.2) and (2.2.1)), and the conclusions of Theorem B.1.3 directly yield points (i) and (ii) in Theorem 2.2.3. Let us thus first prove that the assumptions (a) and (b) from Theorem B.1.3 are indeed satisfied here.

We will keep using the other notations defined at the beginning of Section B.1.2. In particular, using equations (B.1.5) and (B.1.7), here, for all $x \in \mathcal{X}$, $t \in \mathcal{T}$,

$$\bar\kappa(t|x) = \sum_{c\in\mathcal{C}} \delta_{x\in\mathcal{X}^c} \sum_{x'\in\mathcal{X}^c} \mu(x') \frac{\kappa(t|x')}{\mu(\mathcal{X}^c)}, \tag{B.2.1}$$

Moreover, for all $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$, we denote by $I_\kappa(Y;T)$ the mutual information defined by the marginal $q_\kappa(Y;T)$ of $q_\kappa(X,Y,T)$.

Lemma B.2.1. *For all $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$, we have $q_{\bar\kappa}(Y;T) = q_\kappa(Y,T)$. In particular $I_{\bar\kappa}(Y;T) = I_\kappa(Y;T)$.*

Proof. For all $y \in \mathcal{Y}, t \in \mathcal{T}$,

$$\begin{aligned}
q_{\bar\kappa}(y,t) &= \sum_{x\in\mathcal{X}} \mu(y,x)\bar\kappa(t|x) = \sum_{c\in\mathcal{C}} \sum_{x\in\mathcal{X}^c} \mu(y,x)\bar\kappa(t|x) \\
&= \sum_{c\in\mathcal{C}} \sum_{x,x'\in\mathcal{X}^c} \mu(y,x)\mu(x') \frac{\kappa(t|x')}{\mu(\mathcal{X}^c)} && \text{(B.2.2)} \\
&= \sum_{c\in\mathcal{C}} \sum_{x,x'\in\mathcal{X}^c} \mu(y,x')\mu(x) \frac{\kappa(t|x')}{\mu(\mathcal{X}^c)} && \text{(B.2.3)} \\
&= \sum_{c\in\mathcal{C}} \sum_{x'\in\mathcal{X}^c} \mu(y,x')\kappa(t|x') \\
&= \sum_{x\in\mathcal{X}} \mu(y,x')\kappa(t|x') = q(y,t),
\end{aligned}$$

where the first and last lines use that $\{\mathcal{X}^c\}_{c\in\mathcal{C}} = \{\mathsf{pr}^{-1}(c)\}_{c\in\mathcal{C}}$ is a partition of $\mathcal{X}$; line (B.2.2) uses equation (B.2.1); and line (B.2.3) uses the definition of the equivalence relation $\sim_{\mathcal{X}}$ (see equation (2.2.3)): i.e., for $x, x' \in \mathcal{X}^c$, we have for all $y \in \mathcal{Y}$, $\mu(y|x) = \mu(y|x')$, which is equivalent to $\mu(y,x)\mu(x') = \mu(y,x')\mu(x)$. □

Lemma B.2.2. *For all $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$, we have $I_\kappa(T;Y) \le I(X;Y)$, with equality if and only if for all $t \in \mathrm{supp}(q_\kappa(T))$, there exists $c \in \mathcal{C}$ such that $\kappa^{-1}(t) \subseteq \mathsf{pr}^{-1}(c)$.*

Proof. We have

$$I_\kappa(T;Y) = \sum_{y\in\mathrm{supp}(\mu_Y),t\in\mathrm{supp}(q_\kappa(T))} \mu(y)q_\kappa(t|y)\log\left(\frac{q_\kappa(t|y)}{q_\kappa(t)}\right)$$
$$= \sum_{y\in\mathrm{supp}(\mu_Y),t\in\mathrm{supp}(q_\kappa(T))} \mu(y)\left(\sum_x \mu(x|y)\kappa(t|x)\right)\log\left(\frac{\sum_x \mu(x|y)\kappa(t|x)}{\sum_x \mu(x)\kappa(t|x)}\right).$$

But for all $y \in \mathrm{supp}(\mu_Y)$ and $t \in \mathrm{supp}(q_\kappa(T))$, from the log-sum inequality (Csiszár et al., 2011), with the convention $0\log(\frac{0}{0}) := 0$,

$$\left(\sum_x \mu(x|y)\kappa(t|x)\right)\log\left(\frac{\sum_x \mu(x|y)\kappa(t|x)}{\sum_x \mu(x)\kappa(t|x)}\right) \leq \sum_x \mu(x|y)\kappa(t|x)\log\left(\frac{\mu(x|y)\kappa(t|x)}{\mu(x)\kappa(t|x)}\right). \tag{B.2.4}$$

So that, summing over y and t, we get $I_\kappa(Y;T) \leq I(X;Y)$, with equality if and only if for all $y \in \mathrm{supp}(\mu_Y), t \in \mathrm{supp}(q_\kappa(T))$, it holds in (B.2.4). From the equality case of the log-sum inequality (Csiszár et al., 2011), the latter is equivalent to the existence of nonzero constants $(\alpha_{y,t})_{y\in\mathrm{supp}(\mu_Y),t\in\mathrm{supp}(q_\kappa(T))}$ such that

$$\forall x \in \mathcal{X}, \quad \mu(x)\kappa(t|x) = \alpha_{y,t}\mu(x|y)\kappa(t|x),$$

i.e., such that, for all $y \in \mathrm{supp}(\mu_Y), t \in \mathrm{supp}(q_\kappa(T))$, the quantity $\frac{\mu(x|y)}{\mu(x)}$ is constant on the subset of elements $x \in \mathcal{X}$ for which $\kappa(t|x) > 0$. But the latter subset is precisely the preimage $\kappa^{-1}(t)$, and

$$\frac{\mu(x|y)}{\mu(x)} = \frac{1}{\mu(y)}\mu(y|x),$$

where $\frac{1}{\mu(y)}$ does not depend on x. Thus we proved that $I(Y;T) = I(X;Y)$ holds if and only if for all $t \in \mathrm{supp}(q_\kappa(T))$, the distribution $\mu(Y|x)$ does not depend on $x \in \kappa^{-1}(t)$: i.e., if and only if for all $t \in \mathrm{supp}(q_\kappa(T))$, there exists $c \in C$ such that $\kappa^{-1}(t) \subseteq \mathsf{pr}^{-1}(c)$ (use the definition (2.2.3) of the equivalence relation $\sim_{\mathcal{X}}$ defining the projection pr). □

Lemmas B.2.1 and B.2.2 prove that the assumptions resp. (*a*) and (*b*) in Theorem B.1.3. Thus the conclusions of the latter theorem hold in our case, which implies point (*i*) and (*ii*) in Theorem 2.2.3. Assumptions (*a*) and (*b*) being here satisfied also implies that all the intermediary lemmas from Section B.1.2 hold as well. This will be useful for the proof of points (*iii*) to (*v*) in Theorem 2.2.3, to which we now turn.

Proposition B.1.5 ensures that for all $\lambda \in [0,\Lambda]$ and $\kappa \in \mathrm{IB}(\lambda)$, we have the factorisation $\kappa = \kappa_{\mathsf{pr}}\circ\mathsf{pr}$, with κ_{pr} defined in (B.1.4). Thus, for all $\phi \in \mathrm{Bij}(\mathcal{X})$,

$$\begin{aligned}
\phi \in \mathcal{G}_{\mathrm{ci}} &\Leftrightarrow \forall x \in \mathcal{X}, \quad \mu(Y|x) = p(Y|\phi\cdot x)\\
&\Leftrightarrow \forall x \in \mathcal{X}, \quad x \sim_{\mathcal{X}} \phi\cdot x\\
&\Leftrightarrow \mathsf{pr}\circ\phi = \mathsf{pr}\\
&\Rightarrow \kappa_{\mathsf{pr}}\circ\mathsf{pr}\circ\phi = \kappa_{\mathsf{pr}}\circ\mathsf{pr}\\
&\Leftrightarrow \kappa\circ\phi = \kappa,
\end{aligned} \tag{B.2.5}$$

which yields point (*iv*) of Theorem 2.2.3. Moreover, if we assume that $\lambda = \Lambda$, then from Lemma B.1.7, here κ_{pr} is a congruent channel, i.e., there exists a function f such that $f\circ\kappa_{\mathsf{pr}}$

is the identity on C. Therefore the only implication in (B.2.5) becomes an equivalence as well, which yields point (iii) of Theorem 2.2.3.

Let us now prove point (v) in Theorem 2.2.3. The statement is equivalent to proving that the equivalence relation defined by the partition in orbits under $\mathcal{G}_{\text{ci}}$, which we denote here by $\sim_{\text{ci}}$, coincides with the equivalence relation $\sim_{\mathcal{X}}$ defined in (2.2.3). Moreover, by definition of an orbit, $x \sim_{\text{ci}} x'$ means that there exitst $\phi \in \text{Bij}(\mathcal{X})$ such that 1) $\phi \in \mathcal{G}_{\text{ci}}$, i.e., $\mu(Y|\phi \cdot x'') = \mu(Y|x'')$ for all $x'' \in \mathcal{X}$, and 2) $x' = \phi \cdot x$.

Thus $x \sim_{\text{ci}} x'$ clearly implies $\mu(Y|x) = \mu(Y|x')$, i.e., $x \sim_{\mathcal{X}} x'$. Conversely, let us fix $x, x' \in \mathcal{X}$ such that $x \sim_{\mathcal{X}} x'$. We define ϕ as the transposition that permutes x and x', and fixes all the other elements of $\mathcal{X}$. It is straightforward to verify that ϕ satisfies points 1) and 2) above, i.e., that we have $x \sim_{\text{ci}} x'$.

This ends the proof of Theorem 2.2.3.

B.3 Appendix for Section 2.3

B.3.1 On the projection on the exponential family

Let us denote by $\mathsf{cl}\,\mathcal{E}$ the topological closure of the exponential family $\mathcal{E}$ in $\Delta_{\mathcal{A}}$. Here, we denote by $\tilde{\mu} \in \mathsf{cl}\,\mathcal{E}$ the unique distribution (Ay et al., 2017) which achieves the minimum in $\inf_{\nu \in \mathsf{cl}\,\mathcal{E}} D(\mu||\nu)$. Note that we always have $\text{supp}(\mu) \subseteq \text{supp}(\tilde{\mu})$, because otherwise

$$D(\mu||\tilde{\mu}) = +\infty > \inf_{\nu \in \mathcal{E}} D(\mu||\nu).$$

I.e., as here μ is assumed full support, the distribution $\tilde{\mu}$ is full support as well. Thus $\tilde{\mu}$ is both in $\mathsf{cl}\,\mathcal{E}$ and the interior of the simplex $\Delta_{\mathcal{A}}$, which implies that $\tilde{\mu} \in \mathcal{E}$: in particular, it achieves the minimum in $\inf_{\nu \in \mathcal{E}} D(\mu||\nu)$.

Let us now prove that $\tilde{\mu}$ also achieves the latent space divergence, i.e., for all $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$, we automatically have $D(\kappa \cdot \mu||\kappa \cdot \tilde{\mu}) = \inf_{\nu \in \mathcal{E}} D(\kappa \cdot \mu||\kappa \cdot \nu)$. Indeed, for all $\nu \in \mathsf{cl}\,\mathcal{E}$ and with the convention $0 \log(\frac{0}{0}) := 0$,

$$\begin{aligned}
D(\kappa \cdot \mu||\kappa \cdot \tilde{\mu}) - D(\kappa \cdot \mu||\kappa \cdot \nu) &= \sum_{t \in \mathcal{T}} (\kappa \cdot \mu)(t) \log\left(\frac{(\kappa \cdot \nu)(t)}{(\kappa \cdot \tilde{\mu})(t)}\right) \\
&= \sum_{t \in \mathcal{T}} \sum_{a \in \mathcal{A}} \mu(a)\kappa(t|a) \log\left(\frac{\sum_{a \in \mathcal{A}} \nu(a)\kappa(t|a)}{\sum_{a \in \mathcal{A}} \tilde{\mu}(a)\kappa(t|a)}\right) \\
&\leq \sum_{t \in \mathcal{T}} \sum_{a \in \mathcal{A}} \mu(a)\kappa(t|a) \log\left(\frac{\nu(a)\kappa(t|a)}{\tilde{\mu}(a)\kappa(t|a)}\right) \qquad \text{(B.3.1)} \\
&= \sum_{a \in \mathcal{A}} \mu(a) \log\left(\frac{\nu(a)}{\tilde{\mu}(a)}\right) \\
&= D(\mu||\tilde{\mu}) - D(\mu||\nu) \\
&\leq 0, \qquad \text{(B.3.2)}
\end{aligned}$$

where line (B.3.1) uses the log-sum inequality (Csiszár et al., 2011).

B.3.2 Proof of Theorem 2.3.1

Here, we set $\mu := \mu(A)$, $D(\kappa) := D(\kappa \cdot \mu||\kappa \cdot \tilde{\mu})$ for all $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$, and $\mathsf{pr} : \mathcal{A} \to \mathcal{C}$ is the projection on the equivalence classes $(\mathcal{A}^c)_{c \in \mathcal{C}}$ of the relation $\sim$ defined in (2.3.4). Then $R_D(\lambda) = \text{DIB}(\lambda)$ for all $0 \leq \lambda \leq \Lambda := D(\mu||\tilde{\mu})$ (where $R_D(\lambda)$ and $\text{DIB}(\lambda)$ are resp. defined in (B.1.2) and (2.3.1)), and the conclusions of Theorem B.1.3 directly yield Theorem 2.3.1.

Let us thus first prove that the assumptions (a) and (b) from Theorem B.1.3 are indeed satisfied here.

We will keep using the other notations defined at the beginning of Section B.1.2. In particular, using equations (B.1.5) and (B.1.7), here, for all $a \in \mathcal{A}$, $t \in \mathcal{T}$,

$$\bar{\kappa}(t|a) = \sum_{c \in C} \delta_{a \in \mathcal{A}^c} \sum_{a \in \mathsf{pr}^{-1}(c)} \frac{\kappa(t|a)\mu(a)}{\mu(\mathcal{A}^c)}. \tag{B.3.3}$$

We also write $\tilde{q}_\kappa := \tilde{\mu}\kappa \in \Delta_{\mathcal{A}\times\mathcal{T}}$ for all $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$, where we recall that $\tilde{\mu}$ is the projection of μ on the exponential family $\mathcal{E}$ (see Sections 2.3.1 and B.3.1).

Lemma B.3.1. *For all $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$, we have $\bar{\kappa} \cdot \mu = \kappa \cdot \mu$ and $\bar{\kappa} \cdot \tilde{\mu} = \kappa \cdot \tilde{\mu}$. In particular, $D(\bar{\kappa} \cdot \mu || \bar{\kappa} \cdot \tilde{\mu}) = D(\kappa \cdot \mu || \kappa \cdot \tilde{\mu})$.*

Proof. The equality $\bar{\kappa} \cdot \mu = \kappa \cdot \mu$ is point (i) in Lemma B.1.4. Moreover, for all $t \in \mathcal{T}$,

$$\begin{aligned}
(\bar{\kappa} \cdot \tilde{\mu})(t) &= \sum_{a \in \mathcal{X}} \tilde{\mu}(a)\bar{\kappa}(t|x) \\
&= \sum_{c \in C} \sum_{a,a' \in \mathcal{A}^c} \tilde{\mu}(a)\mu(a')\frac{\kappa(t|a')}{\mu(\mathcal{A}^c)} && \text{(B.3.4)} \\
&= \sum_{c \in C} \sum_{a,a' \in \mathcal{A}^c} \mu(a)\tilde{\mu}(a')\frac{\kappa(t|a')}{\mu(\mathcal{A}^c)} && \text{(B.3.5)} \\
&= \sum_{c \in C} \sum_{a' \in \mathcal{A}^c} \tilde{\mu}(a')\kappa(t|a') \\
&= \sum_{a \in \mathcal{X}} \tilde{\mu}(a')\kappa(t|a') = (\kappa \cdot \tilde{\mu})(t),
\end{aligned}$$

where the first and last lines use that $(\mathcal{A}^c)_{c\in C} = (\mathsf{pr}^{-1}(c))_{c\in C}$ is a partition of $\mathcal{X}$; line (B.3.4) uses equation (B.3.3); and line (B.3.5) uses the definition of the equivalence relation $\sim$ (see equation (2.3.4)): i.e., for $a, a' \in \mathcal{A}^c$, we have $\tilde{\mu}(a)\mu(a') = \mu(a)\tilde{\mu}(a')$. □

Lemma B.3.2. *For all $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$, we have $D(q_\kappa(T)||q_{\bar{\kappa}}(T)) \leq D(\mu||\tilde{\mu})$, with equality if and only if for all $t \in \mathrm{supp}(\kappa \cdot \mu)$, there exists $c \in C$ such that $\kappa^{-1}(t) \subseteq \mathsf{pr}^{-1}(c) := \mathcal{A}^c$.*

Proof. We have

$$D(\kappa \cdot \mu||\kappa \cdot \tilde{\mu}) = \sum_{t \in \mathrm{supp}(\kappa\cdot\mu)} \left(\sum_{a \in \mathrm{supp}(\mu)} \kappa(t|a)\mu(a) \right) \log \left(\frac{\sum_{a \in \mathrm{supp}(\mu)} \kappa(t|a)\mu(a)}{\sum_{a \in \mathrm{supp}(\mu)} \kappa(t|a)\tilde{\mu}(a)} \right),$$

But for all $t \in \mathrm{supp}(\kappa \cdot \mu)$, from the log-sum inequality (Csiszár et al., 2011), with the convention $0 \log(\frac{0}{0}) := 0$,

$$\left(\sum_{a \in \mathrm{supp}(\mu)} \kappa(t|a)\mu(a) \right) \log \left(\frac{\sum_{a \in \mathrm{supp}(\mu)} \kappa(t|a)\mu(a)}{\sum_{a \in \mathrm{supp}(\mu)} \kappa(t|a)\tilde{\mu}(a)} \right) \leq \sum_{a \in \mathrm{supp}(\mu)} \kappa(t|a)\mu(a) \log \left(\frac{\kappa(t|a)\mu(a)}{\kappa(t|a)\tilde{\mu}(a)} \right). \tag{B.3.6}$$

So that, summing over t, we get $D(\kappa \cdot \mu||\kappa \cdot \tilde{\mu}) \leq D(\mu||\tilde{\mu})$, with equality if and only if for all $t \in \mathcal{T}$, it holds in (B.3.6). From the equality case of the log-sum inequality (Csiszár et al., 2011), the latter is equivalent to the existence of nonzero constants $(\alpha_t)_{t \in \mathrm{supp}(\kappa\cdot\mu)}$ such that

$$\forall a \in \mathcal{A} \quad \kappa(t|a)\mu(a) = \alpha_t \kappa(t|a)\tilde{\mu}(a),$$

i.e., such that, for all $t \in \mathrm{supp}(\kappa \cdot \mu)$, we have $\mu(a) = \alpha_t \tilde{\mu}(a)$ for all $a \in \mathcal{A}$ such that $\kappa(t|a) > 0$, i.e.,

$$\forall t \in \mathrm{supp}(\kappa \cdot \mu),\ \forall a \in \kappa^{-1}(t),\ \frac{\mu(a)}{\tilde{\mu}(a)} = \alpha_t$$

In the above, note that the fraction $\frac{\mu(a)}{\tilde{\mu}(a)}$ does make sense, because here $\mathrm{supp}(\tilde{\mu}) = \mathcal{A}$ (see Section B.3.1). Thus we proved that equality holds in (B.3.6) if and only if for all $t \in \mathrm{supp}(\kappa \cdot \mu)$, the quotient ${}^{\mu(a)}/_{\tilde{\mu}(a)}$ is constant on $\kappa^{-1}(t)$, i.e., if and only if for all $t \in \mathrm{supp}(\kappa \cdot \mu)$, there exists an c such that $\kappa^{-1}(t) \subseteq \mathsf{pr}^{-1}(c) = \mathcal{A}^c$ (see the definition (2.3.4) of the equivalence relation defining the projection pr.) □

Lemmas B.2.1 and B.2.2 prove that the assumptions resp. (a) and (b) in Theorem B.1.3. Thus the conclusions of the latter theorem hold in our case, which ends the proof of Theorem 2.2.3.

Moreover, note that assumptions (a) and (b) being here satisfied also implies that all the intermediary lemmas from Section B.1.2 hold as well. This will be useful for the proof of Theorem 2.3.4 below.

B.3.3 Proof of Lemma 2.3.3

Writing $P_{Y|X}$ the column transition matrix corresponding to the channel $\mu(Y|X)$,

$$\begin{aligned}(\sigma, \tau) \in \mathcal{G}_{\mathrm{ce}} \quad &\Leftrightarrow \quad P_{Y|X} P_\sigma = P_\tau P_{Y|X} \\ &\Leftrightarrow \quad P_{Y|X} = P_\tau P_{Y|X} P_{\sigma^{-1}} \\ &\Leftrightarrow \quad P_{Y|X} = P_{\tau \cdot Y|\sigma \cdot X},\end{aligned}$$

where the last equivalence comes from the fact that the left multiplication of $P_{Y|X}$ by the permutation matrix P_τ induces the permutation τ of the rows of $P_{Y|X}$; whereas the right multiplication of $P_{Y|X}$ by the permutation matrix $P_{\sigma^{-1}}$ induces the permutation $(\sigma^{-1})^{-1} = \sigma$ of the columns of $P_{Y|X}$. This proves the equivalence $(i) \Leftrightarrow (ii)$. But this implies that, denoting by $[(x, y)]$ the orbit of a point (x, y) under the equivariance group's action,

$$\begin{aligned}&\mathsf{pr}_{\mathrm{ce}}(x_1, y_1) = \mathsf{pr}_{\mathrm{ce}}(x_2, y_2) \\ \Leftrightarrow \quad &[(x_1, y_1)] = [(x_2, y_2)] \\ \Leftrightarrow \quad &\exists\, (\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y}) \ : \ \begin{cases} (\phi, \psi) \in \mathcal{G}_{\mathrm{ce}}, \\ (\phi \otimes \psi)(x_1, y_1) = (x_2, y_2). \end{cases} \\ \Leftrightarrow \quad &\exists\, (\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y}) \ : \ \begin{cases} \forall\, (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad \mu(y|x) = \mu(\tau \cdot y|\sigma \cdot x), \\ (\phi \otimes \psi)(x_1, y_1) = (x_2, y_2). \end{cases}\end{aligned}$$

From this equivalence, the equivalence $(ii) \Leftrightarrow (iii)$ easily follows. This ends the proof of Lemma 2.3.3.

B.3.4 Proof of Theorem 2.3.4

Points (i) and (ii) are direct applications of Theorem 2.3.1.

$(iii) - (iv)$. The proof is almost identical to that of points $(iii) - (iv)$ of Theorem 2.2.3 (see Appendix B.2). Let here pr denote the projection defined by the relation

$$(x, y) \sim (x', y') \quad \Leftrightarrow \quad \mu(y|x) = \mu(y'|x')$$

Proposition B.1.5 ensures that for all λ and $\kappa \in \mathrm{DIB}_{\mathrm{ce}}(\lambda)$, we have the factorisation $\kappa = \bar{\kappa} \circ \mathsf{pr}$, where $\bar{\kappa} = \bar{q}(T|S_J)$ is defined in (B.1.4). Thus, for all $(\phi, \psi) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$,

$$\begin{aligned}
(\phi,\psi) \in \mathcal{G}_{\mathrm{ce}} \quad &\Leftrightarrow \quad \forall (x,y) \in \mathcal{X}\times\mathcal{Y}, \quad \mu(y|x) = \mu(\psi\cdot y|\phi\cdot x) \\
&\Leftrightarrow \quad \forall (x,y) \in \mathcal{X}\times\mathcal{Y}, \quad (x,y) \sim (\phi\cdot x, \psi\cdot y) \\
&\Leftrightarrow \quad \mathsf{pr}\circ(\phi,\psi) = \mathsf{pr} \\
&\Rightarrow \quad \bar{\kappa}\circ\mathsf{pr}\circ(\phi,\psi) = \bar{\kappa}\circ\mathsf{pr} \\
&\Leftrightarrow \quad \kappa\circ(\phi,\psi) = \kappa,
\end{aligned} \tag{B.3.7}$$

where the first equivalence uses Lemma 2.3.3. This yields point (iv) in Theorem 2.3.4. Moreover, if we assume that $\lambda = \Lambda$, then from Lemma B.1.7, here $\bar{\kappa}$ is a congruent channel, i.e., there exists a function f such that $f \circ \bar{\kappa}$ is the identity on $\mathcal{C}$. Thus the only implication in (B.3.7) becomes an equivalence as well, which yields point (iii) of Theorem 2.3.4.

(v). Here, the reasoning used for the proof of point (v) in Theorem 2.2.3 does not work. Indeed the transposition $\Phi \in \mathrm{Bij}(\mathcal{X}\times\mathcal{Y})$ that permutes two pairs (x,y) and (x',y') and fixes all the other ones does not have a split form $\phi \otimes \psi$ for some $(\phi,\psi) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$. Let us exhibit an explicit counter-example. Let $\mathcal{X} = \{1,2,3\}$ and $\mathcal{Y} = \{1,2\}$, with $\mu(X)$ uniform and $\mu(Y|X)$ defined through the row transition matrix

$$\begin{pmatrix} c & p_{12} \\ p_{21} & c \\ p_{31} & p_{32} \end{pmatrix}$$

where we choose $c, p_{12}, p_{21}, p_{31}$, and p_{32} pairwise distinct. It can be easily shown that this channel has no non-trivial equivariances, i.e., $\mathcal{G}_{\mathrm{ce}} = \{e_{\mathcal{X}\times\mathcal{Y}}\}$, so that the projection on orbits $\mathsf{pr}_{\mathrm{ce}}$ is the identity of $\mathcal{X}\times\mathcal{Y}$. Yet the projection pr defined by the the relation $\sim$ will here identify the two pairs (x,y) such that $\mu(y|x) = c$. Therefore $\mathsf{pr}_{\mathrm{ce}} \neq \mathsf{pr}$.

This ends the proof of Theorem 2.3.4.

B.3.5 Relation to the Intertwining IB

The present work (which was published in (Charvin et al., 2025)) is a follow-up on that in (Charvin et al., 2023b); in this section we explicitly relate the two. In the latter reference, we considered what we called the *Intertwining IB* problem:

$$\mathrm{IIB}(\lambda) \;:=\; \underset{\substack{\kappa \in \mathscr{K}(\mathcal{X}\times\mathcal{Y},\mathcal{T}) \,:\\ D(\kappa\cdot\mu \,||\, \kappa\cdot(\mu_{\mathcal{X}}\otimes\mu_{\mathcal{Y}})) \geq \lambda}}{\arg\min} I_{\kappa}(X,Y;T), \tag{B.3.8}$$

where $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$ are the marginals of μ on resp. $\mathcal{X}$ and $\mathcal{Y}$. It can easily be verified this is a DIB problem with $\mathcal{E} = \Delta_{\mathcal{X}} \otimes \Delta_{\mathcal{Y}}$ and $\mathscr{K}_{\mathsf{shape}} = \mathscr{K}(\mathcal{X}\times\mathcal{Y}, \mathcal{T})$.

Problem (B.3.8) is used in (Charvin et al., 2023b) to characterise equivariances under specific conditions: if (i) the distribution μ is discrete and full support, and (ii) $\mu_{\mathcal{Y}}$ is uniform, then the solution κ to (B.3.9) with $\lambda = I(X;Y)$ are such that a pair $(\phi,\psi) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$ is an equivariance if and only if $\kappa\circ(\phi\otimes\psi) = \kappa$.

Thus, Section 2.3.2 is an improvement on the latter result: here, we replace the Intertwining IB problem by the similar problem $\mathrm{DIB}_{\mathrm{ce}}$, from which we obtain the same characterisation as above, except that the assumption (ii) that $\mu_{\mathcal{Y}}$ is uniform can now be dropped. Note that the latter assumption holds precisely when the hierarchical model $\mathcal{E}_{\mathrm{IIB}} := \Delta_{\mathcal{X}} \otimes \Delta_{\mathcal{Y}}$ from (Charvin et al., 2023b) coincides with the hierachical model $\mathcal{E}_{\mathrm{ce}} := \Delta_{\mathcal{X}} \otimes \{\upsilon_{\mathcal{Y}}\}$ from Section 2.3.2 (where $\upsilon_{\mathcal{Y}}$ is the uniform distribution on $\mathcal{Y}$). In this sense, we have clarified that

$\mathcal{E}_{\text{ce}}$, and not $\mathcal{E}_{\text{IIB}}$, is the "correct" hierarchical model to characterise — and soften — equivariances.

B.3.6 The classic IB is a Divergence IB

Ref. (Charvin et al., 2023b) proves that the classic IB can be formulated as an Intertwining IB with specific constraints on the shape of compression channels. More precisely, define $\mathcal{T}$ as $\mathcal{T} := \mathcal{T}_{\text{IB}} \times \mathcal{Y}$ with $\mathcal{T}_{\text{IB}} := \mathbb{N}$, [2] and consider the set

$$\mathscr{K}_{\text{IB}} := \{\kappa_{\mathcal{X}} \otimes \mathtt{Id}_{\mathcal{Y}} \ : \ \kappa_{\mathcal{X}} \in \mathscr{K}(\mathcal{X}, \mathcal{T}_{\text{IB}})\} \subset \mathscr{K}(\mathcal{X} \times \mathcal{Y}, \mathcal{T}_{\text{IB}} \times \mathcal{Y})$$

of channels that can compress the $\mathcal{X}$ coordinate but copy the $\mathcal{Y}$ coordinate. This leads to the problem

$$\text{IIB}_{\mathscr{K}_{\text{IB}}}(\lambda) := \underset{\substack{\kappa \in \mathscr{K}_{\text{IB}(X,Y)} : \\ D(\kappa\cdot\mu||\kappa\cdot(\mu_{\mathcal{X}}\otimes\mu_{\mathcal{Y}}))=\lambda}}{\arg\min} I_{\kappa}(X,Y;T), \tag{B.3.9}$$

where $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$ are the marginals of μ on resp. $\mathcal{X}$ and $\mathcal{Y}$ Then:

Proposition B.3.3 ((Charvin et al., 2023b), Prop. 5)**.** *For every $0 \leq \lambda \leq I(X;Y)$, a channel $\kappa_{\mathcal{X}} \otimes \mathtt{Id}_{\mathcal{Y}} \in \mathscr{K}_{\text{IB}}(X,Y)$ solves the problem (B.3.9) if and only if $\kappa_{\mathcal{X}}$ solves the IB problem (2.2.1).*

In this sense, the classic IB is equivalent to the problem (B.3.9). Importantly, in can be easily verified that the latter is a DIB with $\mathscr{K}_{\mathsf{shape}} = \mathscr{K}_{\text{IB}(X,Y)}$ and $\mathcal{E} = \Delta_{\mathcal{X}} \otimes \Delta_{\mathcal{Y}}$.

However, for the sake of consistency with the results presented in this work, let us also prove that the classic IB is equivalent to a DIB with still $\mathscr{K}_{\mathsf{shape}} = \mathscr{K}_{\text{IB}(X,Y)}$, but now

$$\mathcal{E} = \mathcal{E}_{\text{ce}} := \{\nu_{\mathcal{X}}\upsilon_{\mathcal{Y}},\ \nu_{\mathcal{X}} \in \Delta_{\mathcal{X}}\},$$

which is the exponential family used in Section 2.3.2 to fully characterise channel equivariances.

As mentioned in Section 2.3.2, we have $D(\mu||\mathcal{E}_{\text{ce}}) = D(\mu||\mu_{\mathcal{X}} \otimes \upsilon_{\mathcal{Y}})$, where $\mu_{\mathcal{X}}$ is the marginal of μ on $\mathcal{X}$ and $\upsilon_{\mathcal{Y}}$ the uniform distribution on $\mathcal{Y}$. Moreover for $\kappa = \kappa_{\mathcal{X}} \otimes \mathtt{Id}_{\mathcal{Y}} \in \mathscr{K}_{\text{IB}}$, we have $\kappa \cdot \mu = q_{\kappa}(T,Y)$, while $\kappa \cdot (\mu_{\mathcal{X}} \otimes \upsilon_{\mathcal{Y}}) = q_{\kappa}(T) \otimes \upsilon_{\mathcal{Y}}$ and $\kappa \cdot (\mu_{\mathcal{X}} \otimes \mu_{\mathcal{Y}}) = q_{\kappa}(T) \otimes \mu_{\mathcal{Y}}$, where we recall that $q_{\kappa}(X,Y,T) := \mu\kappa$. Thus

$$\begin{aligned} D(\kappa \cdot \mu||\kappa \cdot \mathcal{E}_{\text{ce}}) &= D(q_{\kappa}(T,Y)||q_{\kappa}(T) \otimes \upsilon_{\mathcal{Y}}) \\ &= D(q_{\kappa}(T,Y)||q_{\kappa}(T) \otimes \mu_{\mathcal{Y}}) + D(\mu_{\mathcal{Y}}||\upsilon_{\mathcal{Y}}) \\ &= D(\kappa \cdot \mu||\kappa \cdot (\mu_{\mathcal{X}} \otimes \mu_{\mathcal{Y}})) + D(\mu_{\mathcal{Y}}||\upsilon_{\mathcal{Y}}). \end{aligned}$$

Therefore the constraint function in the DIB defined in (B.3.9), and the constraint function for the same problem but with $\mathcal{E}_{\text{IIB}}$ replaced by $\mathcal{E}_{\text{ce}}$, differ by a constant K that depends only on $\mu_{\mathcal{Y}}$, which is here fixed. In particular, the corresponding DIB problems are equivalent, in that for all $0 \leq \lambda \leq D(\mu||\mu_{\mathcal{X}} \otimes \mu_{\mathcal{Y}})$,

$$\text{IIB}_{\mathscr{K}_{\text{IB}}}(\lambda) = \text{DIB}_{\mathcal{E}_{\text{ce}},\mathscr{K}_{\text{IB}}}(\lambda + K).$$

As Proposition B.3.3 proves that $\text{IIB}_{\mathscr{K}_{\text{IB}}}$ is equivalent to the classic IB, this proves that $\text{DIB}_{\mathcal{E}_{\text{ce}},\mathscr{K}_{\text{IB}}}$ is also equivalent to the classic IB (up to shifting the trade-off parameter λ by a constant K).

In particular, our framework captures channel invariances — which are a special case of channel equivariances with trivial action on the output space — by using the exponential

[2] This choice is formally equivalent to $\mathcal{T} := \mathbb{N}$, as there is a bijection between $\mathbb{N}$ and $\mathbb{N} \times \mathcal{Y}$.

family $\mathcal{E}_{\text{ce}}$ that captures equivariances, and imposing the additional constraint $\mathscr{K}_{\text{IB}}$ of only compressing the input space but leaving the output space unchanged.

B.3.7 Proof of Theorem 2.3.6

The proof is almost identical to that of points (iv) and (v) of Theorem 2.2.3 (see Appendix B.2). Here DIB(Λ) denotes the solutions to the specific DIB problem defined in Section 2.3.3, and pr the corresponding projection defined by

$$a \sim a \quad \Leftrightarrow \quad \mu(a) = \mu(a') \tag{B.3.10}$$

Point (ii) in Theorem 2.3.1 ensures that for all λ and $\kappa \in \text{DIB}(\lambda)$, we have the factorisation $\kappa = \bar{\kappa} \circ \text{pr}$, with $\bar{\kappa}$ defined in (B.1.6). Thus, for all $\Phi \in \text{Bij}(\mathcal{A})$,

$$\begin{aligned} \Phi \in \mathcal{G}_{\text{di}} \quad &\Leftrightarrow \quad \forall a \in \mathcal{A}, \quad \mu(a) = \mu(\Phi \cdot a) \\ &\Leftrightarrow \quad \forall a \in \mathcal{A}, \quad a \sim \Phi \cdot a \\ &\Leftrightarrow \quad \text{pr} \circ \Phi = \text{pr} \\ &\Rightarrow \quad \bar{\kappa} \circ \text{pr} \circ \Phi = \bar{\kappa} \circ \text{pr} \\ &\Leftrightarrow \quad \kappa \circ \Phi = \kappa. \end{aligned} \tag{B.3.11}$$

This yields point (iii) of Theorem 2.3.1. Moreover, if we assume that $\lambda = \Lambda$, then from Lemma B.1.7, here $\bar{\kappa}$ is a congruent channel, i.e., there exists a function f such that $f \circ \bar{\kappa}$ is the identity on $\bar{\mathcal{S}}$. Therefore the only implication in (B.3.11) becomes an equivalence as well, which yields point (i).

(iii). The statement is equivalent to proving that the equivalence relation defined by the partition in orbits under $\mathcal{G}_{\text{di}}$, which we denote here by $\sim_{\text{di}}$, coincides with the equivalence relation $\sim$ defined in (B.3.10). Moreover, by definition of an orbit, $a \sim_{\text{di}} a'$ means that there exitst $\Phi \in \mathcal{G}_{\text{di}}$ such that (i) $\Phi \in \mathcal{G}_{\text{di}}$, i.e., $p(\Phi \cdot a'') = p(a'')$ for all $a'' \in \mathcal{A}$, and (ii) $a' = \Phi \cdot a$.

Thus $a \sim_{\text{di}} a'$ clearly implies $\mu(a) = \mu(a')$, i.e., $a \sim a'$. Conversely, let us fix $a, a' \in \mathcal{A}$ such that $a \sim a'$. We define $\Phi \in \text{Bij}(\mathcal{A})$ as the transposition that permutes a and a', and fixes all the other elements of $\mathcal{A}$. It is straightforward to verify that Φ satisfies points (i) and (ii) above, i.e., that we have $a \sim_{\text{di}} a'$.

This ends the proof of Theorem 2.3.6.

B.4 Appendix for section 2.3.4

In this appendix, the distribution $\mu(A)$ is allowed to not be full support, and we denote by S this support. In this case, there is still a unique distribution $\tilde{\mu}$ in the closure $\text{cl}\,\mathcal{E}$ of $\mathcal{E}$ such that $D(\mu||\tilde{\mu}) = \inf_{r \in \mathcal{E}} D(\mu||\tilde{\mu})$ (Ay et al., 2017). We denote by $\tilde{S}$ the support of $\tilde{\mu}$. Note that $D(\mu||\tilde{\mu}) = \inf_{r \in \mathcal{E}} D(\mu||r) < +\infty$ implies $S \subseteq \tilde{S}$. Whenever this yields no confusion, we remove the subscript κ from the distribution $q_\kappa(A,T) := q\kappa$ or any of its marginals and conditional distributions: e.g., $q(T) := q_\kappa(T)$. Similarly $\tilde{q}(A,T) := \tilde{\mu}\kappa$. We also assume now that $\mathcal{T}$ is finite, and we define the DIB Lagrangian, on $\mathscr{K}(\mathcal{A}, \mathcal{T})$, as

$$\mathcal{L}_\beta(\kappa) := I_\kappa(A;T) - \beta D(q(T)||\tilde{q}(T)). \tag{B.4.1}$$

After technical preliminaries in Section B.4.1, we prove in Section B.4.2 the following necessary condition for local minimisers κ of (B.4.1): for all $a \in \text{supp}(\mu(A))$ and $t \in \text{supp}(q_\kappa(T))$,

$$\kappa(t|a) = \frac{1}{Z(a,\beta)} q(t) \exp\left[-\beta\left(\frac{q(t)\tilde{\mu}(a)}{\tilde{q}(t)\mu(a)} - \log\left(\frac{q(t)\tilde{\mu}(a)}{\tilde{q}(t)\mu(a)}\right) - 1\right)\right], \tag{B.4.2}$$

where $q(t) := \sum_a \mu(a)\kappa(t|a)$ and $\tilde{q}(t) := \sum_a \tilde{\mu}(a)\kappa(t|a)$, with $Z(a,\beta)$ a normaliser. From this fixed-point equation, we obtain a Blahut-Arimoto (BA) algorithm with the same guarantees as BA for the classic IB (Tishby et al., 2000) (see Section B.4.2). Eventually, we provide more details on effective cardinality in Section B.4.3.

B.4.1 Minimisers on $\mathcal{S} \subseteq \mathcal{A}$ yield minimisers on $\mathcal{A}$

In this section, we reduce the minimisation of $\mathcal{L}_\beta$ on $\mathscr{K}(\mathcal{A},\mathcal{T})$ to a minimisation over channels defined only on the support $\mathcal{S}$ of $\mu = \mu(A)$. More precisely, we show that a minimiser of $\mathcal{L}_\beta$ can always be obtained the following way: choose a minimiser $\kappa \in \mathscr{K}(\mathcal{S},\mathcal{T})$ of the Lagrangian $\mathcal{L}_\beta$ restricted to $\mathscr{K}(\mathcal{S},\mathcal{T})$, and extend it to a channel in $\mathscr{K}(\mathcal{A},\mathcal{T})$ by sending $\mathcal{A}\setminus\mathcal{S}$ on a dummy symbol $t_0 \notin \mathrm{supp}(\kappa\cdot\mu)$. This allows us, in our numerical experiments, to reduce the computation of minimisers of (B.4.1) to that of the same Lagrangian restricted to channels defined only on the support $\mathcal{S} \subseteq \mathcal{A}$. This restriction to the support allows us to then use the BA algorithm (see Section B.4.2), which indeed can only be applied on the support $\mathcal{S}$.

Let $\kappa \in \mathscr{K}(\mathcal{A},\mathcal{T})$. We write $\kappa_\mathcal{S} \in \mathscr{K}(\mathcal{S},\mathcal{T})$ and $\mu_\mathcal{S}(A)$ the restrictions of κ, resp. μ, to $\mathscr{K}(\mathcal{S},\mathcal{T})$, resp. $\Delta_\mathcal{A}$: i.e., $\kappa_\mathcal{S}(t|a) := \kappa(t|a)$ and $\mu_\mathcal{S}(a) := \mu(a)$ for all $a \in \mathcal{S}, t \in \mathcal{T}$ We extend all the notations relating to κ in Section B.3 to $\kappa_\mathcal{S}$; in particular, for $a \in \mathcal{S}, t \in \mathcal{T}$,

$$\begin{aligned} q_\mathcal{S}(t) &:= \sum_{a\in\mathcal{S}} q_\mathcal{S}(t,a) := \sum_{a\in\mathcal{S}} \mu_\mathcal{S}(a)\kappa_\mathcal{S}(t|a) = \sum_{a\in\mathcal{S}} \mu(a)\kappa(t|a) = q(t), \\ \tilde{q}_\mathcal{S}(t) &:= \sum_{a\in\mathcal{S}} \tilde{q}_\mathcal{S}(t,a) := \sum_{a\in\mathcal{S}} \tilde{\mu}_\mathcal{S}(a)\kappa_\mathcal{S}(t|a) := \sum_{a\in\mathcal{S}} \tilde{\mu}(a)\kappa(t|a), \end{aligned} \tag{B.4.3}$$

or

$$\begin{aligned} \mathcal{L}_{\beta,\mathcal{S}}(\kappa_\mathcal{S}) &:= I_{\kappa_\mathcal{S}}(A;T) - \beta D(q_\mathcal{S}(T)||\tilde{q}_\mathcal{S}(T)) \\ &:= \sum_{a\in\mathcal{S},t\in\mathrm{supp}(q_\mathcal{S}(T))} \mu_\mathcal{S}(a)\kappa_\mathcal{S}(t|a)\log\left(\frac{\kappa_\mathcal{S}(t|a)}{q_\mathcal{S}(t)}\right) - \beta \sum_{t\in\mathrm{supp}(q_\mathcal{S}(T))} q_\mathcal{S}(t)\log\left(\frac{q_\mathcal{S}(t)}{\tilde{q}_\mathcal{S}(t)}\right). \end{aligned} \tag{B.4.4}$$

We also denote by $\tilde{\mathcal{S}} \subseteq \mathcal{A}$ the support of $\tilde{\mu} = \tilde{\mu}(A)$.

Proposition B.4.1. *Let $\kappa \in \mathscr{K}(\mathcal{A},\mathcal{T})$. Then κ is a global minimum of $\mathcal{L}_\beta$ if and only If*

(i) $\kappa_\mathcal{S}$ is a global minimum of $\mathcal{L}_{\beta,\mathcal{S}}$,

(ii) For all $t \in \mathrm{supp}(q_\kappa(T))$ and $a \in \tilde{\mathcal{S}} \setminus \mathcal{S}$, we have $\kappa(t|a) = 0$.

In particular, if $\kappa_\mathcal{S}$ is a global minimum $\mathcal{L}_{\beta,\mathcal{S}}$, we obtain a global minimum of $\mathcal{L}_\beta$ with the extension $\kappa'(T|A) \in \mathscr{K}(\mathcal{A},\mathcal{T})$ of $\kappa_\mathcal{S}$ defined through

$$\kappa'(T|a) := \begin{cases} \kappa_\mathcal{S}(T|a) & \text{if } a \in \mathcal{S}, \\ \delta_{t_0} & \text{if } a \in \mathcal{A}\setminus\mathcal{S}, \end{cases} \tag{B.4.5}$$

where we chose $t_0 \in \mathcal{T} \setminus \mathrm{supp}(q_\kappa(T))$.

Before proving this result, let us recall that $q(t) = \sum_{a\in\mathcal{S}} \mu(a)\kappa(t|a)$, so that $\mathrm{supp}(q(T)) = \mathrm{supp}(q_\mathcal{S}(T))$ can be seen as the "probabilistic image of $\mathcal{S}$ through the channel $\kappa_\mathcal{S}$", and does not depend on the values of $\kappa(t|a)$ for $a \in \tilde{\mathcal{S}} \setminus \mathcal{S}$. Thus the condition (ii) in Proposition B.4.1 means that κ sends the elements of $\mathcal{S}$ and $\tilde{\mathcal{S}} \setminus \mathcal{S}$ on distinct subsets of bottleneck symbols in $\mathcal{T}$. Moreover, intuitively, the channel $\kappa'(T|A)$ extends $\kappa_\mathcal{S}$ by sending all the elements a outside $\mathcal{S}$ on a "dummy" symbol t_0 which lies outside the image $\mathrm{supp}(q_\mathcal{S}(T))$ of $\mathcal{S}$ through $\kappa_\mathcal{S}$.

Proof. We have

$$
\begin{aligned}
D(q(T)||\tilde{q}(T)) &= \sum_{t\in\mathrm{supp}(q(T))} q(t)\log\left(\frac{q(t)}{\tilde{q}(t)}\right)\\
&= \sum_{t\in\mathrm{supp}(q(T)),a\in\mathcal{A}} \kappa(t|a)\mu(a)\log\left(\frac{\sum_{a\in\mathcal{A}}\kappa(t|a)\mu(a)}{\sum_{a\in\mathcal{A}}\kappa(t|a)\tilde{\mu}(a)}\right)\\
&= \sum_{t\in\mathrm{supp}(q(T)),a\in\mathcal{S}} \kappa(t|a)\mu(a)\log\left(\frac{\sum_{a\in\mathcal{S}}\kappa(t|a)\mu(a)}{\sum_{a\in\mathcal{S}}\kappa(t|a)\tilde{\mu}(a)+\sum_{a\in\tilde{\mathcal{S}}\setminus\mathcal{S}}\kappa(t|a)\tilde{\mu}(a)}\right)\\
&\leq \sum_{t\in\mathrm{supp}(q(T)),a\in\mathcal{S}} \kappa(t|a)\mu(a)\log\left(\frac{\sum_{a\in\mathcal{S}}\kappa(t|a)\mu(a)}{\sum_{a\in\mathcal{S}}\kappa(t|a)\tilde{\mu}(a)}\right) \qquad \text{(B.4.6)}\\
&= \sum_{t\in\mathrm{supp}(q(T)),a\in\mathcal{S}} \kappa'(t|a)\mu(a)\log\left(\frac{\sum_{a\in\mathcal{S}}\kappa'(t|a)\mu(a)}{\sum_{a\in\mathcal{S}}\kappa'(t|a)\tilde{\mu}(a)}\right)\\
&= \sum_{t\in\mathrm{supp}(q(T)),a\in\mathcal{A}} \kappa'(t|a)\mu(a)\log\left(\frac{\sum_{a\in\mathcal{A}}\kappa'(t|a)\mu(a)}{\sum_{a\in\mathcal{A}}\kappa'(t|a)\tilde{\mu}(a)}\right) \qquad \text{(B.4.7)}\\
&= \sum_{t\in\mathcal{T},a\in\mathcal{A}} \kappa'(t|a)\mu(a)\log\left(\frac{\sum_{a\in\mathcal{A}}\kappa'(t|a)\mu(a)}{\sum_{a\in\mathcal{A}}\kappa'(t|a)\tilde{\mu}(a)}\right) \qquad \text{(B.4.8)}\\
&= D(q'(T)||\tilde{q}'(T)),
\end{aligned}
$$

where we defined the marginals $q'(T) := \sum_{t\in\mathcal{T}}\kappa'(t|a)\mu(a)$ and $\tilde{q}'(T) := \sum_{t\in\mathcal{T}}\kappa'(t|a)\tilde{\mu}(a)$. Note that (B.4.7) uses $\kappa'(t|a) = 0$ for $a \in \mathcal{A}\setminus\mathcal{S}$, $t \in \mathrm{supp}(q(T))$, and (B.4.8) uses the definition (B.4.5) of $\kappa'(T|A)$, which implies

$$\kappa'(t_0|a)\mu(a) = (0\times\mu(a))\,\delta_{a\in\mathcal{S}} + (\kappa'(t_0|a)\times 0)\,\delta_{a\in\mathcal{A}\setminus\mathcal{S}} = 0,$$

and thus

$$\sum_{a\in\mathcal{A}}\kappa'(t_0|a)\mu(a)\log\left(\frac{\sum_{a\in\mathcal{A}}\kappa'(t_0|a)\mu(a)}{\sum_{a\in\mathcal{A}}\kappa'(t_0|a)\tilde{\mu}(a)}\right) = 0.$$

Moreover, the r.h.s. of (B.4.6) coincides with $D(q_{\mathcal{S}}(T)||q_{\mathcal{S}}(T))$. On the other hand it is straightforward to verify that $I_\kappa(A;T) = I_{\kappa'}(A;T) = I_{\kappa_{\mathcal{S}}}(A;T)$. Thus

$$\mathcal{L}_\beta(\kappa) \geq \mathcal{L}_\beta(\kappa') = \mathcal{L}_{\beta,\mathcal{S}}(\kappa_{\mathcal{S}}), \qquad \text{(B.4.9)}$$

and equality is achieved in (B.4.9) if and only if it is achieved in (B.4.6). The latter is equivalent to $\sum_{a\in\tilde{\mathcal{S}}\setminus\mathcal{S}}\kappa(t|a)\tilde{\mu}(a) = 0$ for all $t \in \mathrm{supp}(q(T))$, i.e., to $\kappa(t|a) = 0$ for all $t \in \mathrm{supp}(q(T))$ and $a \in \tilde{\mathcal{S}}\setminus\mathcal{S}$, i.e., to point (ii) in Proposition B.4.1.

Assume now that κ minimises $\mathcal{L}_\beta$. Then equation (B.4.9) and its equality case clearly imply point (ii) in Proposition B.4.1. Moreover, if $\kappa_{\mathcal{S},1}$ is another channel in $\mathscr{K}(\mathcal{S},\mathcal{T})$, we can extend it to a channel κ'_1 similarly as in (B.4.5), which yields

$$\mathcal{L}_{\beta,\mathcal{S}}(\kappa_{\mathcal{S},1}) = \mathcal{L}_\beta(\kappa'_1) \geq \mathcal{L}_\beta(\kappa) = \mathcal{L}_\beta(\kappa') = \mathcal{L}_{\beta,\mathcal{S}}(\kappa_{\mathcal{S}}),$$

whence point (i) in Proposition B.4.1.

Conversely, assume that points (i) and (ii) hold. Fix an arbitrary channel $\kappa_1 \in \mathscr{K}(\mathcal{A},\mathcal{T})$

and write $\kappa_{S,1} \in \mathscr{K}(S,\mathcal{T})$, resp. $\kappa_1' \in \mathscr{K}(\mathcal{A},\mathcal{T})$, the restriction of κ_1 to S, resp. the corresponding channel defnied similarly as in (B.4.5). Then

$$\begin{aligned}\mathcal{L}_\beta(\kappa_1) &\geq \mathcal{L}_\beta(\kappa_1') = \mathcal{L}_{\beta,S}(\kappa_{S,1}) \\ &\geq \mathcal{L}_{\beta,S}(\kappa_S) = \mathcal{L}_\beta(\kappa') = \mathcal{L}_\beta(\kappa),\end{aligned}$$

where the last equality uses point (ii) and the equality case of (B.4.9). Therefore κ is indeed a global minimum of $\mathcal{L}_\beta$. □

B.4.2 Self-consistent equation and Blahut-Arimoto algorithm

Here we describe a Blahut-Arimoto (BA) iterative algorithm to compute the minimisers of the DIB Lagrangian (B.4.1). Following Proposition B.4.1, we aim at a minimiser κ_S of the Lagrangian $\mathcal{L}_{\beta,S}$ restricted to $S := \mathrm{supp}(\mu(A))$ (see equation (B.4.4)), which automatically yields solutions for channels defined on the whole alphabet $\mathcal{A}$ (see equation (B.4.5)). To alleviate notations, in this section we will directly write κ and $\mathcal{L}_\beta$ instead of κ_S and $\mathcal{L}_{\beta,S}$. As we will see, our algorithm does not provably converge to a global minimum of the DIB Lagrangian, but it has the same guarantees as the BA algorithm for the classic IB (Tishby et al., 2000): namely, the values of the Lagrangian decrease at each step and converge to a fixed value, and the limit of a corresponding convergent sequence $(\kappa_i)_{i\in\mathbb{N}}$ must satisfy equation (B.4.2).

Critical points are characterised by a self-consistent equation

Taking into account the constraints $\sum_{t\in\mathcal{T}} \kappa(t|a) = 1$ for all $a \in S$, but not the inequality constraints $\kappa(t|a) \geq 0$ for all $a \in S, t \in \mathcal{T}$, we obtain the extended Lagrangian

$$\mathcal{L}_{\beta,\alpha}(\kappa) := I_\kappa(A;T) - \beta D(q_\kappa(T)||\tilde{q}(T)) + \sum_{a\in S, t\in\mathcal{T}} \alpha_a \kappa(t|a), \tag{B.4.10}$$

for some family of real parameters $(\alpha_a)_{a\in\mathcal{A}}$. We derive $\mathcal{L}_{\beta,\alpha}$ on the open quadrant

$$\mathcal{Q}_+ := \{(\kappa(t|a))_{a\in S,t\in\mathcal{T}} \ : \ \forall a \in S, \forall t \in \mathcal{T},\ \kappa(t|a) > 0\} = (\mathbb{R}_+)^{|S||\mathcal{T}|}.$$

First, $q(t) := \sum_{x'} \mu(a')\kappa(t|a')$ and $\tilde{q}(t) := \sum_{x'} \tilde{\mu}(a')\kappa(t|a')$, so that

$$\begin{aligned}\partial_{\kappa(t|a)} q(t) &= \mu(a), \\ \partial_{\kappa(t|a)} \tilde{q}(t) &= \tilde{\mu}(a).\end{aligned}$$

Moreover, note that $q(T)$ and $\tilde{q}(T)$ are strictly positive for $(\kappa(t|a))_{a,t} \in \mathcal{Q}_+$. Thus we can write

$$\begin{aligned}
\partial_{\kappa(t|a)} I_\kappa(A;T) &= \partial_{\kappa(t|a)} \sum_{a',t'} \mu(a')\kappa(t'|a') \log\left(\frac{\kappa(t|a')}{q(t)}\right) \\
&= \mu(a)\log\left(\frac{\kappa(t|a)}{q(t)}\right) + \sum_{a'} \mu(a')\kappa(t|a')\frac{q(t)}{\kappa(t|a')}\frac{q(t)\delta_{a'=a} - \mu(a)\kappa(t|a')}{q(t)^2} \\
&= \mu(a)\log\left(\frac{\kappa(t|a)}{q(t)}\right) + \sum_{a'} \mu(a')\left(\delta_{a'=a} - \frac{\mu(a)\kappa(t|a')}{q(t)}\right) \\
&= \mu(a)\log\left(\frac{\kappa(t|a)}{q(t)}\right) + \mu(a) - \mu(a)\frac{q(t)}{q(t)} \\
&= \mu(a)\log\left(\frac{\kappa(t|a)}{q(t)}\right),
\end{aligned}$$

and

$$\begin{aligned}
\partial_{\kappa(t|a)} D(q(T)||\tilde{q}(T)) &= \partial_{\kappa(t|a)} \sum_{a',t'} \mu(a')\kappa(t'|a') \log\left(\frac{q(t)}{\tilde{q}(t)}\right) \\
&= \mu(a)\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + \sum_{a'} \mu(a')\kappa(t|a')\frac{\tilde{q}(t)}{q(t)}\frac{\mu(a)\tilde{q}(t) - \tilde{\mu}(a)q(t)}{\tilde{q}(t)^2} \\
&= \mu(a)\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + \left(\sum_{a'} \mu(a')\kappa(t|a')\right)\left(\frac{\mu(a)}{q(t)} - \frac{\tilde{\mu}(a)}{\tilde{q}(t)}\right) \\
&= \mu(a)\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + q(t)\left(\frac{\mu(a)}{q(t)} - \frac{\tilde{\mu}(a)}{\tilde{q}(t)}\right) \\
&= \mu(a)\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + \mu(a) - \frac{q(t)}{\tilde{q}(t)}\tilde{\mu}(a).
\end{aligned}$$

Therefore

$$\partial_{\kappa(t|a)} \mathcal{L}_{\beta,\alpha}(\kappa) = \mu(a)\log\left(\frac{\kappa(t|a)}{q(t)}\right) - \beta\left(\mu(a)\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + \mu(a) - \frac{q(t)}{\tilde{q}(t)}\tilde{\mu}(a)\right) + \alpha_a.$$

Now, recall that here the input set of $\kappa = \kappa_S$ is $S = \mathrm{supp}(\mu(A))$. Hence, we can absorb $\mu(a)$ into the constant α_a, and get that a necessary condition for local minimisers of the DIB Lagrangian $\mathcal{L}_\beta$ on $\mathcal{Q}_+$ is the existence of constants $(\alpha_a)_a \in \mathbb{R}^{|S|}$ such that

$$\log\left(\frac{\kappa(t|a)}{q(t)}\right) - \beta\left(\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + 1 - \frac{q(t)\tilde{\mu}(a)}{\tilde{q}(t)\mu(a)}\right) + \alpha_a = 0$$

i.e., such that

$$\kappa(t|a) = q(t)\exp\left[\beta\left(\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + 1 - \frac{q(t)\tilde{\mu}(a)}{\tilde{q}(t)\mu(a)}\right) + \alpha_a\right].$$

Thus we proved that local minimisers of the DIB Lagrangian $\mathcal{L}_\beta$ over the set of channels $\kappa \in \mathscr{K}(S,\mathcal{T})$ with strictly positive entries satisfy the necessary condition

$$\kappa(t|a) = \frac{1}{Z(a,\beta)} q(t)\exp\left[-\beta\left(\frac{q(t)\tilde{\mu}(a)}{\tilde{q}(t)\mu(a)} - \log\left(\frac{q(t)\tilde{\mu}(a)}{\tilde{q}(t)\mu(a)}\right) - 1\right)\right], \tag{B.4.11}$$

where $Z(a,\beta)$ is a positive normaliser. Note that in (B.4.11), we added the factor $\frac{\tilde{\mu}(a)}{\mu(a)}$ in the logarithm. This equivalent reformulation is more suited to the implementation of the Blahut-Arimoto algorithm described below. Indeed, in this form, the expression in the exponential is always non-positive (as shown by the study of the function $x \mapsto x - \log(x) - 1$), which avoids overflow for large β.

A priori, there might also be local minimisers of $\mathcal{L}_\beta$ on the border of $\mathscr{K}(S,\mathcal{T})$. For the sake of completeness, let us outline an argument showing that this is actually not the case. The computations above show that, deriving $\mathcal{L}_\beta(\kappa)$ as a function on $\mathcal{Q}_+$, we get

$$\partial_{\kappa(t|a)}\mathcal{L}_\beta(\kappa) = \mu(a)\left[\log\left(\frac{\kappa(t|a)}{q(t)}\right) - \beta\left(\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + 1 - \frac{q(t)\tilde{\mu}(a)}{\tilde{q}(t)\mu(a)}\right)\right].$$

In particular, for $\kappa \in \mathscr{K}(S,\mathcal{T})$ strictly positive but with at least one coordinate approaching 0, the directional derivative w.r.t this coordinate diverges to $-\infty$. Indeed, $D(\mu||\tilde{\mu}) < +\infty$ implies $S = \mathrm{supp}(\mu(A)) \subseteq \mathrm{supp}(\tilde{\mu}(A))$, while each $\kappa(T|a)$ is a probability, with $\mu(A)$ and $\tilde{\mu}(A)$ fixed; so that there are strictly positive constants k and K such that $k \leq q(t) := \sum_a \mu(a)\kappa(t|a) \leq K$ and $k \leq \tilde{q}(t) := \sum_a \tilde{\mu}(a)\kappa(t|a) \leq K$ for all $t \in \mathrm{supp}(q(T))$ and all $\kappa \in \mathcal{Q}_+$. Thus the term $\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + 1 - \frac{q(t)\tilde{\mu}(a)}{\tilde{q}(t)\mu(a)}$ remains bounded as well. But as $q(t) \geq k$, on the other hand $\log\left(\frac{\kappa(t|a)}{q(t)}\right)$ diverges to $-\infty$ when $\kappa(t|a)$ goes to 0.

Using classic arguments, we can then use the divergence to $-\infty$ of the gradient close to the border, along with the continuity of $\mathcal{L}_\beta$ on the whole closed set $\mathscr{K}(S,\mathcal{T})$, to prove that κ cannot be a local minimum of $\mathcal{L}_\beta$ over $\mathscr{K}(S,\mathcal{T})$ if it has a coordinate $\kappa(t|a)$ equal to 0, i.e., if it is on the border of $\mathscr{K}(S,\mathcal{T})$.

Blahut-Arimoto algorithm

Here, we denote by $\mathscr{K}_+(S,\mathcal{T})$ the subset of $\mathscr{K}(S,\mathcal{T})$ made of channels with only positive entries, by $\Delta_{\mathcal{T},+}$ the open simplex of full-support probabilities on $\mathcal{T}$, and by $\mathbb{R}_+$ the positive real numbers. We define, for $\kappa \in \mathscr{K}(S,\mathcal{T})$, a probability $r(T) \in \Delta_{\mathcal{T}}$ on $\mathcal{T}$, and some $m(T) \in (\mathbb{R}_+)^{|\mathcal{T}|}$,

$$\begin{aligned} &F(\kappa, r(T), m(T)) \\ &:= \sum_{a,t} \mu(a)\kappa(t|a)\log\left(\frac{\kappa(t|a)}{r(t)}\right) - \beta\sum_{a,t}\mu(a)\kappa(t|a)\left(\log\left(m(t)\frac{\tilde{\mu}(a)}{\mu(a)}\right) - m(t)\frac{\tilde{\mu}(a)}{\mu(a)} + 1\right). \end{aligned}$$

The function F is thus defined on the open and convex set

$$\mathsf{Dom}_F := \mathscr{K}_+(S,\mathcal{T}) \times \Delta_{\mathcal{T},+} \times (\mathbb{R}_+)^{|\mathcal{T}|}.$$

The next proposition defines the Blahut-Arimoto (BA) algorithm adapted to our problem, and describes its properties.

Proposition B.4.2. *The function F is convex in each of its coordinates. Moreover, for $\kappa_i(T|A) \in \mathscr{K}_+(S,\mathcal{T})$, defining*

$$\begin{aligned} r_{i+1}(t) &:= \sum_a \mu(a)\kappa_i(t|a), \\ m_{i+1}(t) &:= \frac{\sum_a \mu(a)\kappa_i(t|a)}{\sum_a \tilde{\mu}(a)\kappa_i(t|a)}, \\ q_{i+1}(t|a) &:= \frac{1}{Z(a,\beta)} r_{i+1}(t)\exp\left[-\beta\left(m_{i+1}(t)\frac{\tilde{\mu}(a)}{\mu(a)} - \log\left(m_{i+1}(t)\frac{\tilde{\mu}(a)}{\mu(a)}\right) - 1\right)\right], \end{aligned} \tag{B.4.12}$$

where $Z(a,\beta)$ is a positive normaliser, we have:

(i) All quantities in (B.4.12) are well-defined, and $(q_{i+1}(T|A), r_{i+1}(T), m_{i+1}(T)) \in \mathsf{Dom}_F$.

(ii) $F(q_i(T|A), r_i(T), m_i(T)) = \mathcal{L}_\beta(q_i(T|A)) + K$, where the Lagrangian $\mathcal{L}_\beta$ is defined in (B.4.1) and K is a constant that does not depend on i.

(iii) At each update of $\kappa_i(T|A)$, $r_i(T)$ and $m_i(T)$, the function F is minimised w.r.t. the corresponding coordinate. In particular,

$$F(q_{i+1}(T|A), r_{i+1}(T), m_{i+1}(T)) \leq F(\kappa_i(T|A), r_i(T), m_i(T)).$$

Before proving it, let us first draw the consequences of Proposition B.4.2. Define some $q_0(T|A) \in \mathscr{K}_+(\mathcal{S},\mathcal{T})$, and the corresponding sequence $(\kappa_i(T|A), r_i(T), m_i(T))_{i\geq 1}$ from (B.4.12). From point (i), the sequence is included in Dom_F, and from point (ii), we have, for all i,

$$\mathcal{L}_\beta(q_i(T|A)) = F((q_i(T|A), r_i(T), m_i(T))) - K.$$

From point (iii), this yields a non-increasing sequence of images $(\mathcal{L}_\beta(q_i(T|A)))_i$. As $\mathcal{L}_\beta$ is bounded from below, this implies that this sequence converges. Moreover, as the closure $\overline{\mathscr{K}_+(\mathcal{S},\mathcal{T})} = \mathscr{K}(\mathcal{S},\mathcal{T})$ of $\mathscr{K}_+(\mathcal{S},\mathcal{T})$ is compact, we can, up to extracting a subsequence, assume that $(\kappa_i(T|A))_i$ converges to a point $q_*(T|A) \in \mathscr{K}(\mathcal{S},\mathcal{T})$.[3] From the definition of $(\kappa_i(T|A))_i$ through (B.4.12) and from the continuity of this iterative equation, we obtain that the limit $q_*(T|A)$ satisfies the fixed-point equation (B.4.11). Hence we proved the claims made the beginning of Appendix B.4.2 about this BA algorithm. Note that even though F is convex in each coordinate, we did not prove that F is convex as a whole. Thus we cannot apply the classic BA arguments (Yeung, 2008) to prove that the sequence $(\mathcal{L}_\beta(q_i(T|A)))_i$ converges to a global minimum of $\mathcal{L}_\beta$. However, the statements proved here match exactly the corresponding statements proven for the BA algorithm in the classic IB case (Tishby et al., 2000).

Proof of Proposition B.4.2. The convexity of F in each coordinate is straightforward. Point (i) comes from the fact that $\kappa \in \mathscr{K}(\mathcal{S},\mathcal{T})$, where $\mathcal{S}$ is the support of $\mu(A)$, which contains that of $\tilde{\mu}(A)$ (because $D(\mu||\tilde{\mu}) < +\infty$). Point (ii) is a direct computation. Let us now prove point (iii).

For fixed $(r(T), m(T))$, we know that the function $F(\cdot, r(T), m(T))$ is convex on $\mathscr{K}_+(\mathcal{S},\mathcal{T})$, so that the minimum is achieved at points κ such that $\nabla_\kappa F(\kappa, r(T), m(T) = 0$. A direct computation shows that the latter is equivalent to, for all $a \in \mathcal{S}, t \in \mathcal{T}$,

$$\kappa(t|a) = \frac{1}{Z(a,\beta)} r(t) \exp\left[-\beta\left(m(t)\frac{\tilde{\mu}(a)}{\mu(a)} - \log\left(m(t)\frac{\tilde{\mu}(a)}{\mu(a)}\right) - 1\right)\right], \tag{B.4.13}$$

where $Z(a,\beta)$ is a positive normaliser. Moreover, it is standard (Yeung, 2008) to prove that, for fixed $(\kappa, m(T)) \in \mathscr{K}_+(\mathcal{S},\mathcal{T}) \times (\mathbb{R}_+)^{|\mathcal{T}|}$, the minimimum of F w.r.t to $r(T)$ is achieved for

$$r(T) = q(T) := \sum_a \mu(a)\kappa(T|a) \tag{B.4.14}$$

[3] In practice, in numerical implementations, we always observed the convergence of $(\kappa_i(T|A))_i$, without any subsequence extraction.

Eventually, for fixed $(\kappa, r(T)) \in \mathscr{K}_+(\mathcal{S},\mathcal{T}) \times \Delta_{\mathcal{T},+}$ the minimum of F w.r.t. $m(T)$ is, again by convexity, achieved if and only if the corresponding gradient vanishes. But we have

$$\begin{aligned}\partial_{m(t)} F(\kappa, r(T), m(T)) &= \sum_a \mu(a)\kappa(t|a)\left(\frac{1}{m(t)} - \frac{\tilde{\mu}(a)}{\mu(a)}\right) \\ &= \frac{q(t)}{m(t)} - \tilde{q}(t),\end{aligned}$$

so that the gradient w.r.t $m(T)$ cancels if and only if for all $t \in \mathcal{T}$,

$$m(t) = \frac{q(t)}{\tilde{q}(t)} = \frac{\sum_a \mu(a)\kappa(t|a)}{\sum_a \tilde{\mu}(a)\kappa(t|a)}. \tag{B.4.15}$$

This proves point (iii). □

B.4.3 Details on effective cardinality

Ref. (Zaslavsky et al., 2019) defines a concept of effective cardinality for the Lagrangian formulation of the classic IB. Here, we adapt this concept to the DIB framework in its primal formulation, i.e., problem (2.3.1), and in a way which also encompasses the case $\mathrm{supp}(\mu(A)) \subsetneq \mathcal{A}$. For $\kappa = \kappa \in \mathscr{K}(\mathcal{A},\mathcal{T})$, consider the "probabilistic image of $\mathcal{A}$ through κ", i.e.,

$$\kappa \cdot \mathcal{A} := \{t \in \mathcal{T} \ : \ \exists a \in \mathcal{A} \ : \ \kappa(t|a) > 0\}.$$

Note that this definition depends only on κ and not on $\mu(A)$. We then define the cardinality of κ as $K(\kappa) := |\kappa \cdot \mathcal{A}|$. However, for $\kappa \in \mathrm{DIB}(\lambda)$, the number $K(\kappa)$ does not necessarily carry any meaningful information about the DIB problem itself: e.g., it can be easily verified that composing any $\kappa \in \mathrm{DIB}(\lambda)$ with a congruent channel γ (which can arbitrarily increase the cardinality $K(\kappa)$) still yields a solution $\gamma\circ\kappa \in \mathrm{DIB}(\lambda)$. This motivates the definition of the minimum number of symbols t necessary to describe the output of a bottleneck encoder κ. Formally:

Definition B.4.3. The *effective cardinality* of a DIB solution $\kappa \in \mathrm{DIB}(\lambda)$ is

$$k(\kappa) := \min_{\gamma \in \mathscr{K}(\mathcal{T}):\ \gamma\circ\kappa \in \mathrm{DIB}(\lambda)} K(\gamma\circ\kappa),$$

i.e., it is the minimum bottleneck cardinality obtained from a post-processing of κ that still produces a DIB solution for the same parameter λ.

Let us fix an arbitrary $\kappa \in \mathrm{DIB}(\lambda)$, write $q(A,T) := \mu\kappa$ where we use the hook-up notation (see Section 2.1), and assume first that μ is full support. It can be shown, using the log-sum inequality, that $k(\kappa)$ is the cardinality of the partition $\bar{\mathcal{T}}$ of $\mathrm{supp}(q(T))$ defined by the equivalence relation $t \sim t' \Leftrightarrow q(A|t) = q(A|t')$.

For the non full support case, denote by $\tilde{\mu}(A)$ the unique distribution satisfying $D(\mu||\tilde{\mu}) = D(\mu||\mathcal{E})$ (Ay et al., 2017), and note that $D(\mu||\mathcal{E}) < \infty$ implies $\mathcal{S} \subseteq \tilde{\mathcal{S}}$, where $\mathcal{S} := \mathrm{supp}(\mu(A))$ and $\tilde{\mathcal{S}} := \mathrm{supp}(\tilde{\mu}(A))$. It can be easily verified that the value of $\kappa(T|a)$ for $a \in \tilde{\mathcal{S}}^{\complement}$ affects neither the target nor the constraint function of the DIB problem (2.3.1). However, a direct consequence of Proposition B.4.1 is that if $\tilde{\mathcal{S}} \setminus \mathcal{S} \neq \emptyset$, the image $\kappa \cdot \mathcal{A}$ of $\mathcal{A}$ through a solution $\kappa \in \mathrm{DIB}(\lambda)$ must contain at least one symbol $t_0 \notin \kappa \cdot \mathcal{S}$, on which to send the elements of $\tilde{\mathcal{S}} \setminus \mathcal{S}$. Here we denoted by $\kappa \cdot \mathcal{S}$ the "probabilistic image of $\mathcal{S}$ through the channel $\kappa = \kappa$", i.e.,

$$\kappa \cdot \mathcal{S} := \{t \in \mathcal{T} \ : \ \exists a \in \mathcal{S} \ : \ \kappa(t|a) > 0\}.$$

Note that $\kappa \cdot \mathcal{S} = \mathrm{supp}(q_\kappa(T))$, for $q_\kappa(T) := \sum_{a\in\mathcal{S}} \mu(a)\kappa(T|a)$.

It can be easily verified that the previous paragraph implies that in the non full support case, the effective cardinality becomes $|\bar{\mathcal{T}}| + 1$. Note that this is the situation we encounter in our numerical experiments (Section 2.4).

We use the above to numerically compute the effective cardinality. Note that the choice of the threshold for rounding $|\kappa(t|a) - \kappa(t|a')|$ to 0 is here important. We choose 10^{-3}.

B.4.4 Computable form of $D_\mu(\kappa||\mathscr{K}_\mathcal{G})$

Here we provide more details on the divergence introduced in Section 2.3.4, and prove that it can be computed directly as a divergence between two channels. Let $\mathcal{S} := \mathrm{supp}(\mu(A))$. For two channels κ, γ in either $\mathscr{K}(\mathcal{A}, \mathcal{T})$ or $\mathscr{K}(\mathcal{S}, \mathcal{T})$, we define their Kullback-Leibler divergence $D_\mu(\kappa||\gamma)$ with respect to $\mu = \mu(A)$, as (Ay et al., 2017)

$$D_\mu(\kappa||\gamma) := \sum_{a\in\mathcal{S}} \mu(a) D(\kappa(T|a)||\gamma(T|a)).$$

We also define, for a group $\mathcal{G}$ acting on $\mathcal{A}$, the set of input-symmetric channels w.r.t. $\mathcal{G}$, i.e.,

$$\mathscr{K}_\mathcal{G} := \{\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{T}) : \quad \forall \Phi \in \mathcal{G}, \ \ \gamma \circ \Phi = \gamma \ \},$$

and the corresponding divergence of some $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$ from $\mathscr{K}_\mathcal{G}$ with respect to μ as (Ay, 2015)

$$D_\mu(\kappa||\mathscr{K}_\mathcal{G}) := \min_{\gamma\in\mathscr{K}_\mathcal{G}} D_\mu(\kappa||\gamma).$$

For all purposes relevant to this chapter's scope, we have $D_\mu(\kappa||\mathscr{K}_\mathcal{G})$ if and only if $\kappa \circ \Phi = \kappa$ for all $\Phi \in \mathcal{G}$. More precisely:

Assume first that $\mu(A)$ is full support. Then, from the continuity of the KL divergence and the fact that $\mathscr{K}_\mathcal{G}$ is a closed subset of $\mathscr{K}(\mathcal{A}, \mathcal{T})$, we have $D(\kappa||\mathscr{K}_\mathcal{G}) = 0$ if and only if $\kappa \in \mathscr{K}_\mathcal{G}$, i.e., $\kappa \circ \Phi = \kappa$ for all $\Phi \in \mathcal{G}$.

Let us now drop the full support assumption on $\mu(A)$, but assume instead that (i) the group $\mathcal{G}$ leaves $\mathcal{S}$ invariant, and (ii) the channel $\kappa = \kappa$ is as $\kappa'(T|A)$ in equation (B.4.5), i.e., it sends $\mathcal{S}^\complement$ on a single symbol outside the image of $\mathcal{S}$ through κ. From point (i), the action of $\mathcal{G}$ on $\mathcal{A}$ induces an action on $\mathcal{S}$, and a corresponding set $\mathscr{K}_{\mathcal{G},\mathcal{S}}$. Denote by $\kappa_\mathcal{S}$ the restriction of a channel $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$ to $\mathcal{S}$. Using both points (i) and (ii), it can be easily verified that for all $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$, we have $\kappa \in \mathscr{K}_\mathcal{G}$ if and only if $\kappa_\mathcal{S} \in \mathscr{K}_{\mathcal{G},\mathcal{S}}$, and that $D_\mu(\kappa||\mathscr{K}_\mathcal{G}) = D_\mu(\kappa_\mathcal{S}||\mathscr{K}_{\mathcal{G},\mathcal{S}})$. From that we can conclude, using the full support case described above, that here we also have $D_\mu(\kappa||\mathscr{K}_\mathcal{G})$ if and only if $\kappa \circ \Phi = \kappa$ for all $\Phi \in \mathcal{G}$.

Note that points (i) and (ii) are satisfied in our numerical experiments in Section 2.4, and that they are also automatically satisfied if $\mu(A)$ is full support.

Let us now provide a form of $D_\mu(\kappa||\mathscr{K}_\mathcal{G})$ which is easier to compute.

Proposition B.4.4. *Fix $\mu(A) \in \Delta_\mathcal{A}$, a finite group $\mathcal{G}$ acting on $\mathcal{A}$ and leaving $\mathcal{S}$ invariant, and $\kappa \in \mathscr{K}(\mathcal{A}, \mathcal{T})$. Then*

$$D_\mu(\kappa||\mathscr{K}_\mathcal{G}) = D_\mu(\kappa||\kappa_\mathcal{G})$$

where $\kappa_\mathcal{G} \in \mathscr{K}(\mathcal{S}, \mathcal{T})$ is defined through, for $a \in \mathcal{S}$ and $t \in \mathcal{T}$,

$$\kappa_\mathcal{G}(t|a) := \frac{\sum_{a'\in[a]} \mu(a')\kappa(t|a')}{\mu([a])},$$

with $[a]$ *the orbit of* a *under the action of* $\mathcal{G}$.

Intuitively, $\kappa_{\mathcal{G}}$ is the average of the channel κ over the group $\mathcal{G}$ acting on its input, computed using the distribution μ on the input.

Proof. It is enough to prove that for all $\gamma \in \mathscr{K}_{\mathcal{G}}$,

$$D_\mu(\kappa||\kappa_{\mathcal{G}}) \leq D_\mu(\kappa||\gamma).$$

For $a \in \mathcal{S}$, we have $[a] \subseteq \mathcal{S}$ (because $\mathcal{G}$ leaves $\mathcal{S}$ invariant), and $\kappa_{\mathcal{G}}(T|a')$ is well-defined and constant for $a' \in [a]$. Moreover, for $\gamma \in \mathscr{K}_{\mathcal{G}}$, it is straightforward to verify that $\gamma(T|a')$ is also constant for $a' \in [a]$, so that

$$\sum_{a' \in [a]} \gamma(t|a')\mu(a') = \gamma(t|a)\mu([a]).$$

Thus, for $\gamma \in \mathscr{K}_{\mathcal{G}}$, and $a_1, \ldots, a_n$ a system of representatives of all the orbits included in $\mathcal{S}$,

$$\begin{aligned}
D_\mu(\kappa||\gamma) - D_\mu(\kappa||\kappa_{\mathcal{G}}) &= \sum_{a \in \mathcal{S}, t \in \mathcal{T}} \mu(a)\kappa(t|a) \log\left(\frac{\kappa_{\mathcal{G}}(t|a)}{\gamma(t|a)}\right) \\
&= \sum_{i=1}^{n} \sum_{t} \log\left(\frac{\kappa_{\mathcal{G}}(t|a_i)}{\gamma(t|a_i)}\right) \sum_{a \in [a_i]} \mu(a)\kappa(t|a) \\
&= \sum_{i=1}^{n} \sum_{t} \log\left(\frac{\sum_{a \in [a_i]} \kappa(t|a)\mu(a)}{\sum_{a \in [a_i]} \gamma(t|a)\mu(a)}\right) \sum_{a \in [a_i]} \mu(a)\kappa(t|a) \\
&= D(q_1||q_2) \geq 0,
\end{aligned}$$

where q_1 and q_2 are distributions defined on ${}^{\mathcal{S}}/_{\mathcal{G}} \times \mathcal{T}$, through

$$\begin{aligned}
q_1([a_i], t) &= \sum_{a \in [a_i]} \mu(a)\kappa(t|a), \\
q_2([a_i], t) &= \sum_{a \in [a_i]} \mu(a)t(t|a).
\end{aligned}$$

□

B.5 Proof of Proposition 2.5.1

Proposition 2.5.1. *Assume that* $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ *is full-support and* $\rho \in \mathscr{K}_{\mathcal{G}}^{\otimes}(\mathcal{X} \times \mathcal{Y})$. *Then*

$$I_\rho(X, Y; X', Y'|G) \leq H(X, Y)$$

with equality if and only if for all $g \in \mathcal{G}$, *the channels* $\phi_g \in \mathscr{K}(\mathcal{X})$ *and* $\psi_g \in \mathscr{K}(\mathcal{Y})$ *such that* $\rho_g = \phi_g \otimes \psi_g$ *are both defined by a bijective function.*

Proof. Let us first fix $g \in \mathcal{G}$. Then we have

$$I_\rho(X, Y; X', Y'|G = g) \leq H(X, Y|G = g) \tag{B.5.1}$$

with equality achieved if and only if, given $G = g$, the variable (X, Y) is a deterministic function of (X', Y'), i.e., if and only if the restriction of the conditional distribution ρ_g to the support of μ is a congruent channel (see Definition 2.2.2). But as μ is here full-support, the latter condition means that ρ_g is a congruent channel from $\mathcal{X}\times\mathcal{Y}$ to itself: i.e., that it is defined

by a bijective function. Moreover, as we assumed that $\rho_g = \phi_g \otimes \psi_g$ for some $\phi_g \in \mathscr{K}(\mathcal{X})$ and $\psi_g \in \mathscr{K}(\mathcal{Y})$, the channel ρ_g is defined by a bijective function if and only if both ϕ_g and ψ_g are. Summing (B.5.1) over $g \in \mathcal{G}$ with uniform weights $\frac{1}{|\mathcal{G}|}$, we obtain

$$I_\rho(X, Y; X', Y'|G) \leq H(X, Y|G) \tag{B.5.2}$$

with equality achieved in (B.5.2) if and only if it is achieved in (B.5.1) for all $g \in \mathcal{G}$, i.e., if and only if for all $g \in \mathcal{G}$, both ϕ_g and ψ_g are defined by a bijective function. But here (X, Y) and G are independent (see (2.5.1)), so that $H(X, Y|G) = H(X, Y)$.

This ends the proof of Proposition 2.5.1. □

Appendix C

Appendix for Chapter 3

C.1 Appendix for Section 3.1

C.1.1 Proof of Proposition 3.1.2

Proposition 3.1.2. *There exists a (set-theoretic) class-pose decomposition w.r.t. ρ if and only if the restricted actions ρ^c and $\rho^{c'}$ are isomorphic for all $c, c' \in C$.*

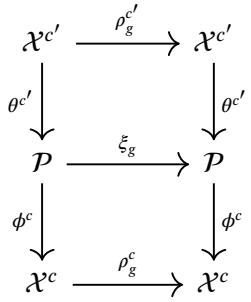
Proof. Assume that (κ, θ, ξ) is a class-pose decomposition w.r.t. ρ, and fix $c, c' \in C$. Denote by $\phi : C \times \mathcal{P} \to \mathcal{X}$ the inverse of (κ, θ). Note that

$$\phi(c, \mathcal{P}) = (\kappa, \theta)^{-1}(c, \mathcal{P}) = \kappa^{-1}(c) \cap \theta^{-1}(\mathcal{P}) = \kappa^{-1}(c) = \mathcal{X}^c,$$

where the last equality uses that κ is the projection on orbits. We can thus consider the map

$$\begin{aligned} \phi^c : \mathcal{P} &\to \mathcal{X}^c \\ p &\mapsto \phi^c(p) := \phi(c, p). \end{aligned}$$

As ϕ is a bijection, the map ϕ^c is a bijection between the pose space $\mathcal{P}$ and the orbit $\mathcal{X}^c$. More precisely, as $\phi^c(\theta(x)) = \phi(c, \theta(x)) = \phi(\kappa(x), \theta(x)) = x$ for all $x \in \mathcal{X}^c$, the map ϕ^c is the inverse of the restriction θ^c of θ to $\mathcal{X}^c$ — and similarly for c'. But by assumption, for all $g \in \mathcal{G}$ we have $\theta^c \circ \rho_g^c = \xi_g \circ \theta^c$ (see point $(ii)'$ before proposition 3.1.2), so which from the above is equivalent to $\rho_g^c \circ \phi^c = \phi^c \circ \xi_g$. Combined with the equality $\theta^{c'} \circ \rho_g^{c'} = \xi_g \circ \theta^{c'}$, this yields the following commutative diagram:



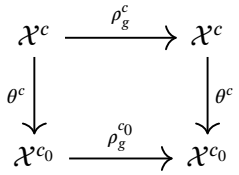
and, in particular,

$$\rho_g^c \circ \phi^c \circ \theta^{c'} = \phi^c \circ \circ \theta \circ \rho_g^{c'},$$

where $\phi^c \circ \theta^{c'}$ is a bijection from $\mathcal{X}^{c'}$ to $\mathcal{X}^c$: i.e., ρ^c and $\rho^{c'}$ are isomorphic.

Assume now that for all $c, c' \in C$, the group actions ρ^c and $\rho^{c'}$ are isomorphic. In particular, for fixed $c_0 \in C$ and all $c \in C$, there exists a bijective map $\theta^c : \mathcal{X}^c \to \mathcal{X}^{c_0}$ such that for

all $g \in \mathcal{G}$, the diagram

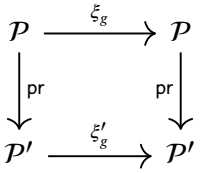


is commutative. From points $(ii)'$ and $(iii)'$ before Proposition 3.1.2, this proves that defining $\theta(x) := \theta^c(x)$ for all $x \in \mathcal{X}, c \in \mathcal{C}$ such that $x \in \mathcal{X}^c$, and $\xi := \rho^c$, the tuple (κ, θ, ξ) defines a class-pose decomposition of ρ, where κ is the projection on orbits. This ends the proof of Proposition 3.1.2. □

C.1.2 Proof of Proposition 3.1.5

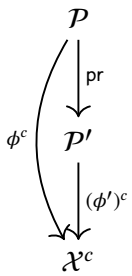
Proposition 3.1.5. *Let $(\rho^c)_{c\in C}$ be a family of actions of the same group $\mathcal{G}$ on resp. state spaces $(\mathcal{X}^c)_{c\in C}$. The j-factor relation between the joinings of $(\rho^c)_{c\in C}$ is a pre-order: i.e., it is reflexive and transitive.*

Proof. The reflexivity is straightforward (take pr equal to the identity map). Let us prove the transitivity. Assume that ξ, ξ' and ξ'' are each joinings of the family of group actions $(\rho^c)_{c\in C}$, on resp. state-spaces $\mathcal{P}$, $\mathcal{P}'$ and $\mathcal{P}''$. Assume that ξ' is a j-factor of ξ with factor map pr: i.e., the diagram



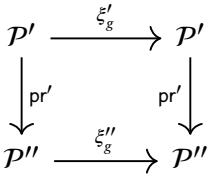
(C.1.1)

commutes for all $g \in \mathcal{G}$, and the diagram



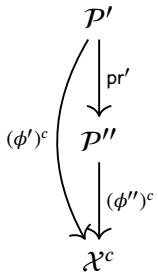
(C.1.2)

commutes for all $c \in C$; and that ξ'' is a j-factor of ξ' with factor map pr′: i.e., the diagram



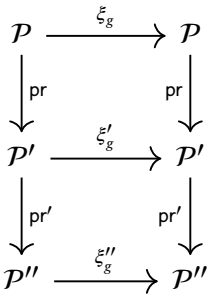
(C.1.3)

commutes for all $g \in \mathcal{G}$, and the diagram



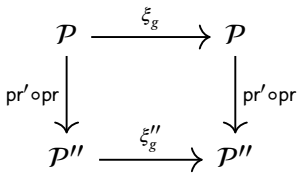
(C.1.4)

commutes for all $c \in \mathcal{C}$. By combining diagrams (C.1.1) and (C.1.3), we obtain, for all $g \in \mathcal{G}$, the commutativity of the diagram



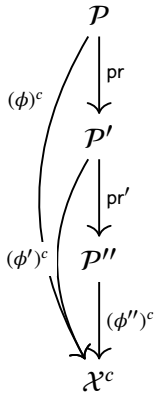
(C.1.5)

which implies the commutativity of the diagram



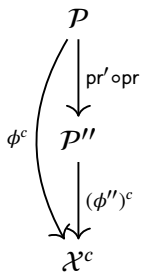
(C.1.6)

Moreover, by combining diagrams (C.1.2) and (C.1.4), we obtain, for all $c \in \mathcal{C}$, the commutativity of the diagram



(C.1.7)

which implies the commutativity of the diagram



(C.1.8)

Thus, from Definition 3.1.4, diagrams (C.1.6) and (C.1.8) imply that the joining (ξ'', ϕ'') is a j-factor of the joining (ξ, ϕ) with factor map $\mathsf{pr}'\circ\mathsf{pr}$. This ends the proof of Proposition 3.1.5. □

C.2 Measure-theoretic definitions

C.2.1 Measurable spaces (long version of Section 3.2.1)

The basic setting of measure theory comprises the objects that are measured, i.e., specific subsets of *measurable* spaces, and the set functions that measure them, i.e., *measures*. We start by recalling definitions relating to the former.

Notation. Sets are denoted with calligraphic letters, e.g., $\mathcal{A}$, and algebras or σ-algebras with gothic letters, e.g., $\mathfrak{A}$. A measurable space $(\mathcal{A}, \mathfrak{A})$ is only denoted by $\mathcal{A}$ when there is no ambiguity on the σ-algebra. The algebra, resp. σ-algebra induced

Definition C.2.1. Let $\mathcal{A}$ be a set. A *σ-algebra* on $\mathcal{A}$ is a collection $\mathfrak{A}$ of subsets of $\mathcal{A}$ such that (i) it contains $\mathcal{A}$, (ii) any complement of element of $\mathfrak{A}$ belongs to $\mathfrak{A}$, and (iii) any countable union of elements of $\mathfrak{A}$ belongs to $\mathfrak{A}$. An *algebra* on $\mathcal{A}$ is a collection $\mathfrak{A}$ of subsets of $\mathcal{A}$ satisfying (i) and (ii), but where we only require finite unions of elements of $\mathfrak{A}$ to belong to $\mathfrak{A}$. If $\mathfrak{A}$ is an arbitrary collection of subsets of $\mathcal{A}$, we denote by $\sigma(\mathfrak{A})$ the smallest σ-algebra containing $\mathfrak{A}$, and by $\mathsf{Alg}(\mathfrak{A})$ the smallest algebra containing $\mathfrak{A}$. A *measurable space* is a pair $(\mathcal{A}, \mathfrak{A})$ where $\mathcal{A}$ is a set and $\mathfrak{A}$ a σ-algebra on $\mathcal{A}$, and it is only denoted by $\mathcal{A}$ when there is no ambiguity on the σ-algebra. The elements of $\mathfrak{A}$ are then called *measurable sets*. For $(\mathcal{A}, \mathfrak{A})$ a measurable space, an algebra $\tilde{\mathfrak{A}} \subseteq \mathfrak{A}$ is called a *generating algebra* if $\sigma(\tilde{\mathfrak{A}}) = \mathfrak{A}$. For $E \subseteq \mathfrak{A}$, the *induced σ-algebra on E* is $\mathfrak{A}_E := \mathfrak{A} \cap E$, which makes $(E, \mathfrak{A}_E)$ into a measurable space. If not mentioned explicitly, any measurable subset of a measurable space, when regarded itself as a measurable space, is equipped with the induced σ-algebra. A map f between two measurable spaces $(\mathcal{A}, \mathfrak{A})$ and $(\mathcal{B}, \mathfrak{B})$ is called *measurable* if for all $E \in \mathfrak{B}$, we have $f^{-1}(E) \in \mathfrak{A}$.

We will focus our attention, as much as possible, on a class of "nice" measurable spaces:

Definition C.2.2. A *Polish space* is a set $\mathcal{A}$ equipped with a topology $\mathcal{T}$ which is:

- Separable: i.e., there exists a countable dense subset,
- Completely metrisable: i.e., there exists a metric that generates the topology of $\mathcal{A}$ and makes it into a complete space (in the sense that all Cauchy sequences converge).

For $(\mathcal{A}, \mathcal{T})$ a topological space, the *Borel σ-algebra* of $\mathcal{A}$ is the σ-algebra $\sigma(\mathcal{T})$ generated by its topology $\mathcal{T}$. It is also denoted by $\mathrm{Bor}_{\mathcal{A}}$ and its elements are called the *Borelians* of $\mathcal{A}$. A *Borel space* $(\mathcal{A}, \mathcal{T}, \sigma(\mathcal{T}))$ is a topological space $(\mathcal{A}, \mathcal{T})$ equipped with its σ-algebra of Borelians. A *standard Borel space* is a Borel space whose topology makes it a Polish space. We will often not refer explicitly to the underlying topology of a standard Borel space, and

denote it by $(\mathcal{A}, \mathfrak{A})$ where $\mathfrak{A} := \mathrm{Bor}_{\mathcal{A}}$, or just $\mathcal{A}$ when this yields no confusion. The set $\mathbb{R}$ of real numbers is always equipped with its usual topology (which is Polish) and corresponding Borelians, with measurable subsets equipped with the induced topology and σ-algebra.

Standard Borel spaces are "nice" because they encompass many important examples (e.g., countable spaces, Euclidean spaces, separable Banach spaces, differential manifolds[1]), but still have enough structure for many desirable properties to hold. Most importantly for us, one can take conditional probabilities (see Proposition C.2.16 below) and do ergodic and information theory (see (Gray, 2009, 2011) and Section 3.4 below). Another fact that will be useful to us it that the σ-algebras of these spaces are *countably generated*, in the following sense:

Proposition C.2.3 ((Gray, 2009) Theorem 4.2)**.** *Let $(\mathcal{A}, \mathfrak{A})$ be a standard Borel space. Then there exists a countable family of measurable subsets $\{E_n\}_{n\in\mathbb{N}} \subseteq \mathfrak{A}$ such that $\mathfrak{A} = \sigma(\{E_n\}_{n\in\mathbb{N}})$.*

Next, we need to define products of measurable spaces.

Definition C.2.4. Let $(\mathcal{A}_i, \mathfrak{A}_i)_{i\in\mathcal{I}}$ a family of measurable spaces. For subsets of indices $\mathcal{J}' \subseteq \mathcal{J} \subseteq \mathcal{I}$, we denote by

$$\begin{aligned} \mathsf{pr}_{\mathcal{J}\to\mathcal{J}'} : \bigtimes_{j\in\mathcal{J}} \mathcal{A}_j &\to \bigtimes_{j'\in\mathcal{J}'} \mathcal{A}_{j'} \\ (a_j)_{j\in\mathcal{J}} &\mapsto (a'_j)_{j'\in\mathcal{J}'} \end{aligned}$$

the corresponding projection between Cartesian products. The set of (finite-dimensional) *rectangles* is

$$\mathsf{Rect} := \bigcup_{\mathcal{J}\subseteq\mathcal{I} \text{ finite}} \mathsf{pr}^{-1}_{\mathcal{I}\to\mathcal{J}}\left(\left\{ \bigtimes_{j\in\mathcal{J}} F_j,\ F_j \in \mathfrak{A}_j \text{ for all } j \in \mathcal{J} \right\}\right).$$

The *product measurable space* of the family $(\mathcal{A}_i, \mathfrak{A}_i)_{i\in\mathcal{I}}$ is then defined as

$$\left(\bigtimes_{i\in\mathcal{I}} \mathcal{A}_c, \bigotimes_{i\in\mathcal{I}} \mathfrak{A}_i\right),$$

where $\bigotimes_{i\in\mathcal{I}} \mathfrak{A}_i := \sigma(\mathsf{Rect})$ is called the *product σ-algebra*. A Cartesian product of measurable space will, unless stated otherwise, always be equipped with the product σ-algebra. Moreover, when $\mathcal{I} = \mathbb{N}$ models time for $\mathcal{A}_n = \mathcal{A}$ for all $n \in \mathbb{N}$ and $\mathcal{A}$ a measurable space, we write $\overrightarrow{\mathcal{A}} := \mathcal{A}^{\mathbb{N}} = \bigtimes_{n\in\mathbb{N}} \mathcal{A}_n$.

We can build standard Borel spaces by taking products of standard Borel spaces. However, this is only true if the product is at most countable:

Proposition C.2.5. *If $(\mathcal{A}_i)_{i\in\mathcal{I}}$ is a family of standard Borel spaces with $\mathcal{I}$ at most countable, then the Cartesian product $\bigtimes_{i\in\mathcal{I}} \mathcal{A}_i$, equipped with the product topology and product σ-algebra, is standard Borel. However, for any family of measurable spaces $(\mathcal{A}_i)_{i\in\mathcal{I}}$ with $\mathcal{I}$ uncountable, where each $\mathcal{A}_i$ contains more than one point, the product topology cannot be Polish: in particular, the resulting Borel space cannot be standard Borel.*

Proof. The first part is Lemma 4.1 in (Gray, 2009). Consider an unncountable product of non-trivial spaces (i.e., with each more than one point), equiped with a topology $\mathcal{T}$ that contains the product topology. Moreover, Theorem 22.3 in (Willard, 1970) states that a unncountable product of non-trivial spaces (i.e., with more than one point), equipped with the product topology, is never metrisable. This proves the second part. □

[1] Differential manifolds are usually assumed separable. They are completely metrizable locally, and thus globally (use a partition of unity).

We will indeed need to consider uncountable products of standard Borel spaces — more precisely, of what we will define as "ergodic components of a Markov Decision Process". Proposition C.2.5 is thus the reason why we need to deal also with general measurable spaces, not only standard Borel ones. Moreover, the following will be useful to endow each of these "ergodic components" with its own standard Borel structure:

Theorem C.2.6 ((Kechris, 1995), Theorem 13.1 and Corollary 13.4)**.** *Let $(\mathcal{A}, \mathcal{T}, \mathfrak{A})$ be standard Borel with topology $\mathcal{T}$, and $E \in \mathfrak{A}$ a Borel subset. Denote by $\mathcal{T}_E := \{O \cap E, O \in \mathcal{T}\}$ the induced topology and $\mathfrak{A}_E := \{F \cap E,\ F \in \mathfrak{A}\}$ the induced σ-algebra. Then there exists a topology $\mathcal{T}'_E$ on E such that $\mathcal{T}_E \subseteq \mathcal{T}'_E$, $\mathrm{Bor}_{\mathcal{T}'_E} = \mathfrak{A}_E$ and $(E, \mathcal{T}'_E, \mathfrak{A}_E)$ is standard Borel.*

In other words, one can equip any Borel subset E of a standard Borel space $\mathcal{A}$ with its own standard Borel structure, whose topology is at least as fine as the one $\mathcal{T}_E$ induced by the ambient space, and whose Borelians coincide exactly with the σ-algebra $\mathfrak{A}_E$ induced by the ambient space.

C.2.2 Measures (long version of Section 3.2.2)

We now turn to measure-related definitions.

Notation. For a subset F of a set $\mathcal{A}$, the complement $\mathcal{A} \setminus F$ of F is denoted by $F^{\complement}$. This is not to be confused with superscripts of the form F^c, which we will heavily rely on and do *not* denote complements.

Definition C.2.7. Let $(\mathcal{A}, \mathfrak{A})$ be a measurable space. A set function $\mu : \mathfrak{A} \to \mathbb{R}$ is *non-negative* of $\mu(F) \geq 0$ for all $F \in \mathfrak{A}$, *σ-additive* if $\mu(\bigcup_{i\in\mathbb{N}} F_i) = \sum_{i\in\mathbb{N}} \mu(F_i)$ whenever the $(F_i)_{i\in\mathbb{N}}$ are disjoint, and *normalised* if $\mu(\mathcal{A}) = 1$. A set function $\mu : \mathfrak{A} \to \mathbb{R}$ is called a *signed measure* if it is σ-additive; it is a *finite positive measure*, or *measure* for short, if it is σ-additive and non-negative; and it is a *probability measure* if it is σ-additive, non-negative and normalised. The sets of signed measures, (finite positive) measures and probability measures on a measurable space $(\mathcal{A}, \mathfrak{A})$ are denoted by resp. $\mathscr{M}_{\mathcal{A}}$, $\mathscr{M}^+_{\mathcal{A}}$ and $\Delta_{\mathcal{A}}$. A *measure space* is a triplet $(\mathcal{A}, \mathfrak{A}, \mu)$ with $(\mathcal{A}, \mathfrak{A})$ a measurable space and $\mu \in \mathscr{M}^+_{\mathcal{A}}$; it is called a *probability space* if $\mu \in \Delta_{\mathcal{A}}$; and it is only denoted by $(\mathcal{A}, \mu)$ when there is no ambiguity on the σ-algebra. For another measurable space $(\mathcal{B}, \mathfrak{B})$ and a measurable map $f : \mathcal{A} \to \mathcal{B}$, the *push-forward* of a measure $\mu \in \mathscr{M}^+_{\mathcal{A}}$ is the measure $f \cdot \mu \in \mathscr{M}^+_{\mathcal{A}}$ defined by $(f \cdot \mu)(B) := \mu(f^{-1}(B))$ for all $B \in \mathfrak{B}$.

Let us stress that:

- Here, **if not specified otherwise, the term "measure" denotes a finite positive measure** on a measurable space $\mathcal{A}$ — i.e., an element of $\mathscr{M}^+_{\mathcal{A}}$. We always specify it explicitly when we are considering a signed measure and not just a (finite positive) measure.
- For all measurable space $\mathcal{A}$, we have the inclusions $\Delta_{\mathcal{A}} \subseteq \mathscr{M}^+_{\mathcal{A}} \subseteq \mathscr{M}_{\mathcal{A}}$.
- We are primarily interested in probability measures. Signed measures will only be necessary to deal with ergodic decomposition in Sections 3.3 and 3.4.

We use the following standard terminology, where $(\mathcal{A}, \mathfrak{A})$, $(\mathcal{B}, \mathfrak{B})$ are measurable spaces:

Definition C.2.8. For $a \in \mathcal{A}$, the *Dirac measure* on a, denoted by δ_a, is defined, for all $F \in \mathfrak{A}$, by $\delta_a(F) = 1$ if $a \in F$, and $\delta_a(F) = 0$ otherwise. For $\tilde{\mathcal{A}} \in \mathfrak{A}$, if $\mu \in \mathscr{M}^+_{\mathcal{A}}$ is such that $\mu(\tilde{\mathcal{A}}^{\complement}) = 0$, then it is said *concentrated on* $\tilde{\mathcal{A}}$, and the measure $\tilde{\mu} \in \mathscr{M}^+_{\tilde{\mathcal{A}}}$ defined by $\tilde{m}(A \cap \tilde{\mathcal{A}}) := \mu(A)$ for all $A \in \mathfrak{A}$ is called the *restriction* of μ to $\tilde{\mathcal{A}}$. For $\tilde{\mathcal{A}} \in \mathfrak{A}$ and $\tilde{\mu} \in \mathscr{M}^+_{\tilde{\mathcal{A}}}$, the *extension* of $\tilde{\mu}$ to $(\mathcal{A}, \mathfrak{A})$ is the measure μ defined by $\mu(F) := \tilde{\mu}(F \cap \tilde{\mathcal{A}})$ for all $F \in \mathfrak{A}$. With a slight abuse of notation, the restriction or extension of a measure μ will often

be denoted with the same symbol μ. Eventually, for two maps $f : \mathcal{A} \to \mathcal{B}$ and $g : \mathcal{A} \to \mathcal{B}$, and $\mu \in \mathscr{M}^+_{\mathcal{A}}$, the equality $f = g$ is said to hold μ-almost everywhere, or μ-a.e. for short, if there exists $\tilde{\mathcal{A}} \in \mathfrak{A}$ such that $\mu(\tilde{\mathcal{A}}^{\complement}) = 0$ and $f(a) = g(a)$ for all $a \in \tilde{\mathcal{A}}$.

The next result is fundamental to build probability distributions through only their values on a generating algebra.

Proposition C.2.9 ((Gray, 2009), Corollary 3.4 and Theorem 1.1)**.** *Let $(\mathcal{A}, \mathfrak{A})$ be measurable, and let $\mathfrak{A}' \subseteq \mathfrak{A}$ a generating algebra. Then:*

(i) If a set function $\mu : \mathfrak{A}' \to \mathbb{R}$ is non-negative, normalised and σ-additive, then there exists a unique extension of μ into a probability on $(\mathcal{A}, \mathfrak{A})$: i.e., there exists a unique probability $\tilde{q} : \mathfrak{A} \to \mathbb{R}$ such that $\tilde{q}$ coincides with μ on $\mathfrak{A}'$.

(ii) In particular, if two probability distributions $\mu, \mu' \in \Delta_{\mathcal{A}}$ coincide on the generating algebra $\mathfrak{A}'$, then they coincide on the full σ-algebra $\mathfrak{A}$.

Proposition C.2.9 is particularly useful for measures on product spaces (see Definition C.2.4). Indeed, it can be easily verified that in a product measurable space, the algebra generated by rectangles is made of the finite unions of rectangles, where the unions can always be chosen disjoint. If a set function is only defined on Rect, but σ-additive on Rect, we can thus extend it to a σ-additive function on $\sigma(\mathsf{Rect})$. This, combined with Proposition C.2.9, yields the following: specifying the values of a non-negative, normalised and σ-additive set function on rectangles uniquely defines a probability on the product space. It is thus common practice to define positive measures on product spaces only through their values on rectangles, where the non-negativity, normalisation and σ-additivity of the proposed set function on Rect is implicitly stated to indeed hold. We will also do this here.

A related tool is the Kolmogorov extension theorem. It states, in short, that given a family of compatible probability distributions on finite sets of coordinates, there exists a unique extension to a probability on the whole product space.

Theorem C.2.10 (Kolmogorov extension theorem for standard Borel spaces)**.** *Let $(\mathcal{A}_i, \mathcal{T}_i, \mathrm{Bor}(\mathcal{T}_i))_{i \in \mathcal{I}}$ a family of standard Borel spaces, with $\mathcal{I}$ an arbitrary set. For each finite $\mathcal{J} \subseteq \mathcal{I}$, let $\mu_{\mathcal{J}} \in \Delta_{\mathcal{A}_{\mathcal{J}}}$, where $\mathcal{A}_{\mathcal{J}}$ is the product measurable space of $(\mathcal{A}_j)_{j \in \mathcal{J}}$. Assume that for any $\mathcal{J}' \subseteq \mathcal{J}$ with $\mathcal{J}, \mathcal{J}'$ both finite, we have*

$$\mathsf{pr}_{\mathcal{J} \to \mathcal{J}'} \cdot \mu_{\mathcal{J}} = \mu_{\mathcal{J}'},$$

i.e., the marginal of $\mu_{\mathcal{J}}$ on $\mathcal{J}'$ coincides with $\mu_{\mathcal{J}'}$. Eventually, denote by $\boldsymbol{\mathcal{A}}$ the product measurable space of $(\mathcal{A}_i)_{i \in \mathcal{I}}$. Then there exists a unique probability measure $\boldsymbol{\mu} \in \Delta_{\boldsymbol{\mathcal{A}}}$ such that for all finite $\mathcal{J} \subseteq \mathcal{I}$,

$$\mathsf{pr}_{\mathcal{I} \to \mathcal{J}} \cdot \boldsymbol{\mu} = \mu_{\mathcal{J}},$$

i.e., such that the marginal of $\boldsymbol{\mu}$ on $\mathcal{J}$ coincides with $\mu_{\mathcal{J}}$.

Proof. Theorem 2.4.3 in (Tao, 2011) states that the conclusion holds if each measurable space $(\mathcal{A}_i, \mathfrak{A}_i)$ is equipped with a metric d_i such that (i) any compact set $K \subseteq \mathcal{A}_i$ (for the topology induced by d_i) is a measurable set $K \in \mathfrak{A}_i$, and (ii) the measure μ_i on $(\mathcal{A}_i, \mathfrak{A}_i)$ satisfies $\mu_i(F) = \sup_{K \subseteq F,\, K \text{ compact}} \mu_i(F)$ for all $F \in \mathfrak{A}_i$. Point (i) is clearly satisfied in a standard Borel space: compact subsets of Haussdorf, and a fortiori metric spaces are closed (see Theorem 17.5 in (Willard, 1970)), and therefore Borelian. The fact that standard Borel spaces satisfy point (ii) is the content of Theorem 18.2 of (Coudène, 2016). □

C.2.3 Lebesgue and Bochner integrals (long version of Section 3.2.3)

We assume familiarity with Lebesgue integration theory , and refer to, e.g., (Gray, 2009) for basic definitions and results. Let us just point out the following fact which we will use several times:

Lemma C.2.11 ((Rudin, 1987), Theorem 1.39)**.** *Let $(\mathcal{A}, \mathfrak{A}, \mu)$ be a measure space. If f is an integrable function satisfying $\int_E f\, d\mu = 0$ for all $E \in \mathfrak{A}$, then $f = 0$ μ-a.e. on $\mathcal{A}$.*

Moreover, the following notation will be convenient.

Notation. For $\mathcal{A}$ a measurable space, $f : \mathcal{A} \to \mathbb{R}$ a measurable function and $\mu \in \mathscr{M}_{\mathcal{A}}$, whenever the integral makes sense, we write

$$\langle \mu, f \rangle := \int_{\mathcal{A}} f\, d\mu(a).$$

In Section 3.4, we will need to consider integrals valued in probability spaces, for which *Bochner integrals* are a natural language. Indeed, the latter generalise the Lebesgue integrals, defined for scalar-valued functions, to maps valued in potentially infinite-dimensional spaces — more precisely, *Banach spaces*.[2] We refer to, e.g., Appendix E in (Cohn, 2013) for a complete presentation of Bochner integrals, and point out only the few facts that we will need. We assume familiarity with basic facts about Banach spaces.

First, while Lebesgue integration considers measurable scalar-valued functions, integrating Banach space-valued functions requires adding a requirement of separability:

Definition C.2.12. Let $(\mathcal{A}, \mathfrak{A})$ be a measurable space, and $(\mathcal{B}, \mathcal{T}, \mathrm{Bor}_{\mathcal{T}})$ a Borel space such that $\mathcal{B}$ is a Banach space with norm $\|\cdot\|$, and $\mathcal{T}$ the corresponding norm topology. A function $h : \mathcal{A} \to \mathcal{B}$ is said *strongly measurable* if it is measurable and $h(\mathcal{A})$ is separable. Given a measure $\mu \in \mathscr{M}_{\mathcal{A}}$, the map h is said *Bochner integrable* w.r.t. μ' if it is strongly measurable and such that the real-valued function $x \mapsto \|f(x)\|$ is Lebesgue-integrable.

One can then construct a notion of *Bochner integral* of any Bochner integrable function (Cohn, 2013), where the integral is valued in the same Banach space as the function's output space. Lebesgue integral's dominated convergence theorem generalises to Bochner integrals:

Theorem C.2.13 ((Cohn, 2013), Theorem E.6)**.** *Let $\mathcal{A}$ a measurable space, $\mathcal{B}$ a real Banach space, μ a positive measure on $\mathcal{A}$, and $g : \mathcal{A} \to [0, \infty]$ Lebesgue-integrable. Suppose that $h, h_0, h_1, \ldots$ are strongly measurable functions from $\mathcal{A}$ to $\mathcal{B}$ such that (i) $h(a) = \lim_{n\to\infty} h_n(a)$ for all $a \in \mathcal{A}$, and (ii) $\|h_n(a)\| \leq g(a)$ for μ-a.e. $a \in \mathcal{A}$ and all $n \in \mathbb{N}$. Then $h, h_0, h_1, \ldots$ are integrable, and $\lim_{n\to\infty} \int_{\mathcal{A}} h_n(a) d\mu(a) = \int_{\mathcal{A}} h(a) d\mu(a)$.*

C.2.4 Appendix for Section 3.2.4

The following is straightforward to verify. It shows that measured isomorphisms are "well-behaved": (i) mod 0 inverses of measured isomorphisms are also isomorphisms, (ii) the composition of two measured isomorphisms also is one, and (iii) almost everywhere equality with a measured isomorphism is the same as being a measured isomorphism.

Proposition C.2.14. *Let $(\mathcal{A}, \mu_{\mathcal{A}})$, $(\mathcal{B}, \mu_{\mathcal{B}})$, $(\mathcal{C}, \mu_{\mathcal{C}})$ be measure spaces, $f : (\mathcal{A}, \mu_{\mathcal{A}}) \to (\mathcal{B}, \mu_{\mathcal{B}})$, $g : (\mathcal{B}, \mu_{\mathcal{B}}) \to (\mathcal{C}, \mu_{\mathcal{C}})$ measured isomorphisms, f^{-1} a mod 0 inverse of f, and $f' : \mathcal{A} \to \mathcal{B}$ measurable. Then:*

[2] I.e., vector spaces equipped with a norm that induces a complete metric. Banach spaces can be seen as infinite-dimensional generalisations of Euclidean spaces, and play a central role in functional analysis.

(i) f^{-1} *is a measured isomorphism from* $(\mathcal{B}, \mu_{\mathcal{B}})$ *to* $(\mathcal{A}, \mu_{\mathcal{A}})$*, and* f *is a mod 0 inverse of* f^{-1}*.*

(ii) $g \circ f$ *is a measured isomorphism from* $(\mathcal{A}, \mu_{\mathcal{A}})$ *to* $(\mathcal{C}, \mu_{\mathcal{C}})$*.*

(iii) If the equality $f = f'$ *holds* $\mu_{\mathcal{A}}$*-a.e., then* f' *is a measured isomorphism from* $(\mathcal{A}, \mu_{\mathcal{A}})$ *to* $(\mathcal{B}, \mu_{\mathcal{B}})$*.*

C.2.5 Appendix for Section 3.2.5

There is a useful formula for evaluating hook-ups on arbitrary sets of the product space. For $F \in \mathfrak{A} \times \mathfrak{B}$, define the *section* of F at a as

$$F_a := \{b \in \mathcal{B} : \ (a, b) \in F\}.$$

We then have the following:

Proposition C.2.15. *For* $(\mathcal{A}, \mathfrak{A})$*,* $(\mathcal{B}, \mathfrak{B})$ *measurable spaces,* $\mu \in \Delta_{\mathcal{A}}$*,* $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$ *and any* $F \in \mathfrak{A} \otimes \mathfrak{B}$*, the bounded function* $a \mapsto \gamma(F_a|a)$ *is measurable, and*

$$\mu\gamma(F) = \int_{\mathcal{A}} \gamma(F_a|a) d\mu(a).$$

Proof. See, e.g., Section 1.5 of (Gray, 2011) — in particular equation (1.28) there. □

The following proposition shows that in standard Borel spaces, we can always define conditional distribution from joint distributions:

Proposition C.2.16. *Let* $\mathcal{A}, \mathcal{B}$ *be standard Borel, and* $\mu \in \Delta_{\mathcal{A}\times\mathcal{B}}$*. Then there exists a channel* $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$ *such that, writing* $\mu_{\mathcal{A}}$ *the marginal of* μ *on* $\mathcal{A}$*, we have* $\mu = \mu_{\mathcal{A}}\gamma$*.*

Proof. This is a direct consequence of the fact that in standard Borel spaces, there are *regular conditional probabilities* (see, e.g., Section 6.8 of (Gray, 2009) for a definition of regular conditional probabilites, and the corresponding statement). □

C.2.6 Appendix for Section 3.2.6

Let us provide more explanations on the fact that tensor product are well defined (see Definition 3.2.6). For $\mathcal{I}$ finite, this is as a consequence of our remarks after Definition C.2.4. For $\mathcal{I}$ arbitrary but each $\mathcal{A}_i$ and $\mathcal{B}_i$ standard Borel, the case $\mathcal{I}$ finite ensures that equations (3.2.1), (3.2.2) and (3.2.3) uniquely define probabilities on each finite product, and as these finite-dimensional probabilities are consistent, the Kolmogorov Extension Theorem C.2.10 ensures that they uniquely extend to probabilities on the full product indexed by $\mathcal{I}$. For tensor or output tensor products of channels, we also need to verify the measurability condition from the Definition 3.2.2 of channel, but this is straightforward once definitions are unpacked.

Let us also state the the following properties of tensor products (bilinearity and commutation with composition), mentioned in Section 3.2.6:

Proposition C.2.17. *For* $|\mathcal{I}| = 2$*, each of the tensor products defined above is bilinear. Moreover, let us also consider another family* $(\mathcal{C}_i)_{i\in\mathcal{I}}$ *of measurable spaces, and, for all* $i \in \mathcal{I}$*, a corresponding channel* $\gamma_{\mathcal{B}_i\to\mathcal{C}_i} \in \mathscr{K}(\mathcal{B}_i, \mathcal{C}_i)$*. Assume that* $\mathcal{I}$ *is finite or that for all* $i \in \mathcal{I}$*, the spaces* $\mathcal{A}_i$*,* $\mathcal{B}_i$ *and* $\mathcal{C}_i$ *are standard Borel. Then:*

$$\left(\bigotimes_{i\in\mathcal{I}} \gamma_{\mathcal{B}_i\to\mathcal{C}_i}\right) \circ \left(\bigotimes_{i\in\mathcal{I}} \gamma_{\mathcal{A}_i\to\mathcal{B}_i}\right) = \bigotimes_{i\in\mathcal{I}} \left(\gamma_{\mathcal{B}_i\to\mathcal{C}_i} \circ \gamma_{\mathcal{A}_i\to\mathcal{B}_i}\right).$$

Proof. The proof is straightforward on rectangles, which yields the result through Proposition C.2.9 and Theorem C.2.10. □

C.3 Some useful rules (details)

Here we state and prove the technical results mentioned in Section 3.2.8.

Lemma C.3.1. *Let $(\mathcal{A}, \mathfrak{A})$, $(\mathcal{A}', \mathfrak{A}')$ $(\mathcal{B}, \mathfrak{B})$, $(\mathcal{B}', \mathfrak{B}')$, $(\mathcal{C}, \mathfrak{C}')$ measurable spaces. Then:*

(i) For $\mu \in \Delta_{\mathcal{A}}$ and $\gamma, \gamma' \in \mathscr{K}(\mathcal{A}, \mathcal{B})$,

$$\gamma = \gamma' \quad \mu\text{-a.e.} \quad \Rightarrow \quad \mu\gamma = \mu\gamma'.$$

If moreover $(\mathcal{B}, \mathfrak{B})$ is standard Borel, then the implication is an equivalence.

(ii) For $\mu \in \Delta_{\mathcal{A}}$, $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$ and $\gamma' \in \mathscr{K}(\mathcal{B}, \mathcal{C})$,

$$(\gamma' \circ \gamma) \cdot \mu = \gamma' \cdot (\gamma \cdot \mu).$$

(iii) For $a_0 \in \mathcal{A}$ and $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$,

$$\delta_{a_0}\gamma = \delta_{a_0} \otimes (\gamma \cdot \delta_{a_0}).$$

(iv) For $\mu \in \Delta_{\mathcal{A}}$, $f : \mathcal{A} \to \mathcal{A}'$, $g : \mathcal{B} \to \mathcal{B}'$ measurable and $\gamma \in \mathscr{K}(\mathcal{A}', \mathcal{B})$,

$$(f \otimes g) \cdot \mu(\gamma \circ f) = (f \cdot \mu)(g \circ \gamma).$$

(v) For $\mu \in \Delta_{\mathcal{A}}$, $f : \mathcal{A} \to \mathcal{B}$ measurable, $\gamma, \gamma' \in \mathscr{K}(\mathcal{B}, \mathcal{C})$,

$$\mu(\gamma \circ f) = \mu(\gamma' \circ f) \quad \Leftrightarrow \quad (f \cdot \mu)\gamma = (f \cdot \mu)\gamma'$$

(vi) For $\mu \in \Delta_{\mathcal{A}}$, $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$, $\gamma' \in \mathscr{K}(\mathcal{A}', \mathcal{B}')$ and $f : \mathcal{A} \to \mathcal{A}'$, $g : \mathcal{B} \to \mathcal{B}'$ measurable,

$$\mu(g \circ \gamma) = \mu(\gamma' \circ f) \quad \Rightarrow \quad (f \otimes g) \cdot \mu\gamma = (f \cdot \mu)\gamma'.$$

If moreover $(\mathcal{B}', \mathfrak{B}')$ is standard Borel, then the implication is an equivalence.

(vii) Let $\mu \in \Delta_{\mathcal{A}}$, $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$, measured isomorphisms $f : (\mathcal{A}, \mu) \to (\mathcal{A}', f \cdot \mu)$, $g : (\mathcal{B}, \gamma \cdot \mu) \to (\mathcal{B}', g \cdot \gamma \cdot \mu)$, and f^{-1}, g^{-1} mod 0 inverses of resp. f and g. Then:

(a) $\mu(\gamma \circ f^{-1} \circ f) = \mu\gamma$, and in particular, $(f^{-1} \circ f) \cdot \mu = \mu$,

(b) $\mu(g^{-1} \circ g \circ \gamma) = \mu\gamma$,

(c) $f \otimes g$ is a measured isomorphism from $(\mathcal{A} \times \mathcal{B}, \mu\gamma)$ to $(\mathcal{A}' \times \mathcal{B}', (f \cdot \mu)(g \circ \gamma \circ f^{-1}))$.

(d) For any channel $\gamma' \in \mathscr{K}(\mathcal{A}', \mathcal{B})$,

$$\mu\gamma = \mu(\gamma' \circ f) \quad \Leftrightarrow \quad (f \cdot \mu)(\gamma \circ f^{-1}) = (f \cdot \mu)\gamma'.$$

(viii) Let $\mu \in \Delta_{\mathcal{A}}$, $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$, $\gamma' \in \mathscr{K}(\mathcal{A}', \mathcal{B}')$, measured isomorphisms $f : (\mathcal{A}, \mu) \to (\mathcal{A}', f \cdot \mu)$, $g : (\mathcal{B}, \gamma \cdot \mu) \to (\mathcal{B}', g \cdot \gamma \cdot \mu)$, and f^{-1}, g^{-1} mod 0 inverses of resp. f and g. Then

$$\mu(g \circ \gamma) = \mu(\gamma' \circ f) \quad \Leftrightarrow \quad \mu(g^{-1} \circ \gamma') = \mu(\gamma \circ f^{-1})$$

Proof. In this proof, we will repeatedly use Proposition C.2.9, which implies in particular that two probabilities on a product measurable space coincide if they coincide on the rectangles (see the remarks after Definition C.2.4).

(i). Assume that $\gamma = \gamma'$ holds μ-a.e.; let $\tilde{\mathcal{A}} \in \mathfrak{A}$ such that $\mu(\tilde{\mathcal{A}}) = 1$ and $\gamma(\cdot|a) = \gamma'(\cdot|a)$ for all $a \in \tilde{\mathcal{A}}$. Then for all $E_{\mathcal{A}} \in \mathfrak{A}$. $E_{\mathcal{B}} \in \mathfrak{B}$,

$$\begin{aligned}
\mu\gamma(E_{\mathcal{A}} \times E_{\mathcal{B}}) &= \int_{E_{\mathcal{A}}} \gamma(E_{\mathcal{B}}|a) d\mu(a) \\
&= \int_{E_{\mathcal{A}} \cap \tilde{\mathcal{A}}} \gamma(E_{\mathcal{B}}|a) d\mu(a) \\
&= \int_{E_{\mathcal{A}} \cap \tilde{\mathcal{A}}} \gamma'(E_{\mathcal{B}}|a) d\mu(a) \\
&= \int_{E_{\mathcal{A}}} \gamma'(E_{\mathcal{B}}|a) d\mu(a) \\
&= \mu'\gamma'(E_{\mathcal{A}} \times E_{\mathcal{B}}).
\end{aligned}$$

Thus $\mu\gamma = \mu'\gamma'$. Conversely, assume that $\mu\gamma = \mu\gamma'$. As $(\mathcal{B}, \mathfrak{B})$ is standard Borel, there exists a countable family of measurable subsets $\{F_n\}_{n\in\mathbb{N}} \subseteq \mathfrak{B}$ such that $\mathfrak{B} = \sigma(\{F_n\}_{n\in\mathbb{N}})$ (see Proposition C.2.3). Moreover, for fixed n and all $E \in \mathfrak{A}$,

$$\int_E \gamma(F_n|a) d\mu(a) = (\mu\gamma)(E \times F_n) = (\mu\gamma')(E \times F_n) = \int_E \gamma'(F_n|a) d\mu(a).$$

From Lemma C.2.11, this implies the existence of a set $\mathcal{A}_n \in \mathfrak{A}$ such that $\mu(\mathcal{A}_n) = 1$ and $\gamma(F_n|a) = \gamma'(F_n|a)$ for all $a \in \mathcal{A}_n$. Let us define $\tilde{\mathcal{A}} := \bigcap_{n\in\mathbb{N}} \mathcal{A}_n$, and let $a \in \tilde{\mathcal{A}}$. We have $\gamma(F_n|a) = \gamma'(F_n|a)$ for all $n \in \mathbb{N}$, which from Proposition C.2.9 implies that $\gamma(\cdot|a)$ and $\gamma'(\cdot|a)$ coincide on $\sigma(\{F_n\}_{n\in\mathbb{N}}) = \mathfrak{B}$. Eventually, note that $\mu(\tilde{\mathcal{A}}) = 1$, as $\tilde{\mathcal{A}}$ is a countable intersection of measurable sets of probability one.

(ii). This is straightforward.

(iii). The result holds on rectangles: if $E_{\mathcal{A}} \in \mathfrak{A}$, $E_{\mathcal{B}} \in \mathfrak{B}$,

$$\delta_{a_0}\gamma(E_{\mathcal{A}} \times E_{\mathcal{B}}) = \mathbb{1}_{a_0 \in E_{\mathcal{A}}} \gamma(E_{\mathcal{B}}|a_0) = \delta_{a_0}(E_{\mathcal{A}})(\gamma \cdot \delta_{a_0})(E_{\mathcal{B}}) = \big(\delta_{a_0} \otimes (\gamma \cdot \delta_{a_0})\big)(E_{\mathcal{A}} \times E_{\mathcal{B}}).$$

Therefore it is true for all $E \in \mathfrak{A} \otimes \mathfrak{B}$.

(iv). We will first prove that

$$(f \otimes \mathtt{Id}_{\mathcal{B}}) \cdot \mu(\gamma \circ f) = (f \cdot \mu)\gamma,$$

and then that

$$(\mathtt{Id}_{\mathcal{A}} \otimes g) \cdot \mu\gamma = \mu(g \circ \gamma),$$

which, using $f \otimes g = (\mathtt{Id}_{\mathcal{A}} \otimes g) \circ (f \otimes \mathtt{Id}_{\mathcal{B}})$ and point (ii), will imply the result.

For $E_{\mathcal{A}'} \times E_{\mathcal{B}} \in \mathfrak{A}' \times \mathfrak{B}$,

$$\begin{aligned}(f \otimes \mathtt{Id}_{\mathcal{B}}) \cdot \big(\mu(\gamma \circ f)\big)(E_{\mathcal{A}'} \times E_{\mathcal{B}}) &= \big(\mu(\gamma \circ f)\big)(f^{-1}(E_{\mathcal{A}'}) \times E_{\mathcal{B}}) \\ &= \int_{f^{-1}(E_{\mathcal{A}'})} (\gamma \circ f)(E_{\mathcal{B}}|a) d\mu(a) \\ &= \int_{\mathcal{A}} \mathbb{1}_{E_{\mathcal{A}'}}(f(a)) \gamma(E_{\mathcal{B}}|f(a)) d\mu(a) \\ &= \int_{\mathcal{A}} (h \circ f) d\mu,\end{aligned}$$

where $h : a' \mapsto \mathbb{1}_{E_{\mathcal{A}'}}(a')\gamma(E_{\mathcal{B}}|a')$ is measurable non-negative. Therefore

$$\begin{aligned}(f \otimes \mathtt{Id}_{\mathcal{B}}) \cdot \big(\mu(\gamma \circ f)\big)(E_{\mathcal{A}'} \times E_{\mathcal{B}}) &= \int_{\mathcal{A}'} h d(f \cdot \mu) \\ &= \int_{E_{\mathcal{A}'}} \gamma(E_{\mathcal{B}}|a') d(f \cdot \mu)(a') \\ &= \big((f \cdot \mu)\gamma\big)(E_{\mathcal{A}'} \times E_{\mathcal{B}}).\end{aligned}$$

As $E_{\mathcal{A}'} \times E_{\mathcal{B}}$ is an arbitrary rectangle, this proves that $f \otimes \mathtt{Id}_{\mathcal{B}} = (f \cdot \mu)\gamma$. Now for $E_{\mathcal{A}} \times E_{\mathcal{B}'} \in \mathfrak{A} \times \mathfrak{B}'$,

$$\begin{aligned}\big((\mathtt{Id}_{\mathcal{A}} \otimes g) \cdot \mu\gamma\big)(E_{\mathcal{A}} \times E_{\mathcal{B}'}) &= (\mu\gamma)(E_{\mathcal{A}} \times g^{-1}(E_{\mathcal{B}'})) \\ &= \int_{E_{\mathcal{A}}} \gamma(g^{-1}(E_{\mathcal{B}'})|a) d\mu(a) \\ &= \int_{E_{\mathcal{A}}} (g \circ \gamma)(E_{\mathcal{B}'}|a) d\mu(a) \\ &= \big(\mu(g \circ \gamma)\big)(E_{\mathcal{A}} \times E_{\mathcal{B}'}).\end{aligned}$$

(v). If $\mu(\gamma \circ f) = \mu(\gamma' \circ f)$, then from point (iii),

$$(f \cdot \mu)\gamma = (f \otimes \mathtt{Id}_{\mathcal{C}}) \cdot \mu(\gamma \circ f) = (f \otimes \mathtt{Id}_{\mathcal{C}}) \cdot \mu(\gamma' \circ f) = (f \cdot \mu)\gamma',$$

Conversely, assume $(f \cdot \mu)\gamma = (f \cdot \mu)\gamma'$. Then, for all $E_{\mathcal{B}} \in \mathfrak{B}$, $E_{\mathcal{C}} \in \mathfrak{C}$, defining the measurable function $h : \mathcal{B} \to \mathcal{C}$ by $h(b) := \gamma(E_{\mathcal{C}}|b)$, we have

$$\begin{aligned}
\big(\mu(\gamma\circ f)\big)\big(f^{-1}(E_{\mathcal{B}}) \times E_{\mathcal{C}}\big) &= \int_{f^{-1}(E_{\mathcal{B}})} (\gamma\circ f)(E_{\mathcal{C}}|a)d\mu(a) \\
&= \int_{f^{-1}(E_{\mathcal{B}})} \gamma(E_{\mathcal{C}}|f(a))d\mu(a) \\
&= \int_{f^{-1}(E_{\mathcal{B}})} (h\circ f)d\mu(a) \\
&= \int_{E_{\mathcal{B}}} hd(f \cdot \mu)(a) \\
&= \int_{E_{\mathcal{B}}} \gamma(E_{\mathcal{C}}|a)d(f \cdot \mu)(a) \\
&= \big((f \cdot \mu)\gamma\big)\big(E_{\mathcal{B}} \times E_{\mathcal{C}}\big) \\
&= \big((f \cdot \mu)\gamma'\big)\big(E_{\mathcal{B}} \times E_{\mathcal{C}}\big) \\
&= \int_{f^{-1}(E_{\mathcal{B}})} \gamma'(E_{\mathcal{C}}|f(a))d\mu(a) \\
&= \big(\mu(\gamma'\circ f)\big)\big(f^{-1}(E_{\mathcal{B}}) \times E_{\mathcal{C}}\big).
\end{aligned}$$

Note, however, that the rectangles of the form $f^{-1}(E_{\mathcal{B}}) \times E_{\mathcal{C}}$ only generate the σ-algebra $f^{-1}(\mathfrak{B}) \otimes \mathfrak{C}$, not the full σ-algebra $\mathfrak{A} \otimes \mathfrak{C}$. We thus need the following last step. For fixed $E_{\mathcal{C}} \in \mathfrak{C}$, the equality

$$\mu(\gamma\circ f)(f^{-1}(E_{\mathcal{B}}) \times E_{\mathcal{C}}) = \mu'(\gamma\circ f)(f^{-1}(E_{\mathcal{B}}) \times E_{\mathcal{C}})$$

holds for arbitrary $f^{-1}(E_{\mathcal{B}}) \in f^{-1}(\mathfrak{B})$. From Lemma C.2.11, this implies the existence of a set $\tilde{\mathcal{A}} \in f^{-1}(\mathfrak{B}) \subseteq \mathfrak{A}$ (which might depend on $E_{\mathcal{C}}$) such that $\mu(\tilde{\mathcal{A}}) = 1$ and $(\gamma\circ f)(E_{\mathcal{C}}|a) = (\gamma'\circ f)(E_{\mathcal{C}}|a)$ for all $a \in \tilde{\mathcal{A}}$. Let us now choose an arbitrary $E_{\mathcal{A}} \in \mathfrak{A}$. Then

$$\begin{aligned}
\mu(\gamma\circ f)(E_{\mathcal{A}} \times E_{\mathcal{C}}) &= \int_{E_{\mathcal{A}}} (\gamma\circ f)(E_{\mathcal{C}}|a)d\mu(a) \\
&= \int_{E_{\mathcal{A}}\cap\tilde{\mathcal{A}}} (\gamma\circ f)(E_{\mathcal{C}}|a)d\mu(a) \\
&= \int_{E_{\mathcal{A}}\cap\tilde{\mathcal{A}}} (\gamma'\circ f)(E_{\mathcal{C}}|a)d\mu(a) \\
&= \int_{E_{\mathcal{A}}} (\gamma'\circ f)(E_{\mathcal{C}}|a)d\mu(a) \\
&= \mu(\gamma.\circ f)(E_{\mathcal{A}} \times E_{\mathcal{C}}).
\end{aligned}$$

Apply now the reasoning above to all $E_{\mathcal{C}} \in \mathfrak{C}$, we obtain that $\mu(\gamma\circ f)$ and $\mu(\gamma'\circ f)$ coincide on all rectangles $E_{\mathcal{A}} \times E_{\mathcal{C}}$ of $\mathfrak{A} \otimes \mathfrak{C}$, and thus on the whole σ-algebra.

(vi). If $\mu(g\circ\gamma) = \mu(\gamma'\circ f)$, then

$$\begin{aligned}(f\otimes g)\cdot\mu\gamma &= \big((f\otimes \mathtt{Id}_{\mathcal{B}'})\circ(\mathtt{Id}_{\mathcal{A}}\otimes g)\big)\cdot\mu\gamma\\ &= (f\otimes \mathtt{Id}_{\mathcal{B}'})\cdot\big((\mathtt{Id}_{\mathcal{A}}\otimes g)\cdot\mu\gamma\big)\\ &= (f\otimes \mathtt{Id}_{\mathcal{B}'})\cdot\big(\mu(g\circ\gamma)\big)\\ &= (f\otimes \mathtt{Id}_{\mathcal{B}'})\cdot\big(\mu(\gamma'\circ f)\big)\\ &= (f\cdot\mu)\gamma',\end{aligned}$$

where we used points (ii) and (iv) above.

Assume now that $(\mathcal{B}',\mathfrak{B}')$ is standard Borel and $(f\otimes g)\cdot\mu\gamma = (f\cdot\mu)\gamma'$, and let us write $\mu' := f\cdot\mu$. From Proposition C.2.3, there exists a countable family of measurable subsets $\{F_n\}_{n\in\mathbb{N}}\subseteq\mathfrak{B}'$ such that $\mathfrak{B}' = \sigma(\{F_n\}_{n\in\mathbb{N}})$. Moreover, for fixed n and all $E\in\mathfrak{A}'$, by assumption

$$\mu\gamma(f^{-1}(E)\times g^{-1}(F_n)) = \mu'\gamma'(E\times F_n),$$

i.e., denoting by $h_n : \mathcal{A}'\to[0,1]$ the measurable function defined by $h_n(a') := \gamma'(g^{-1}(F_n)|a')$

$$\begin{aligned}\int_{f^{-1}(E)}\gamma(g^{-1}(F_n)|a)d\mu(a) &= \int_E \gamma'(F_n|a')d\mu(a)\\ &= \int_E h d(f\cdot\mu)\\ &= \int_{f^{-1}(E)} h\circ f\, d\mu(a)\\ &= \int_{f^{-1}(E)}\gamma(F_n|f(a))\,\mu(a).\end{aligned}$$

As this is true for all $E\in\mathfrak{A}'$ and as the sets of the form $f^{-1}(E)$ generate the sub-σ-algebra $f^{-1}(\mathfrak{A}')\subseteq\mathfrak{A}$, from Lemma C.2.11, there exists a set $\mathcal{A}_n\in f^{-1}(\mathfrak{A}')\subseteq\mathfrak{A}$ such that $\mu(\mathcal{A}_n)=1$ and $\gamma(g^{-1}(F_n)|a) = \gamma'(F_n|f(a))$ for all $a\in\mathcal{A}_n$. Defining $\tilde{\mathcal{A}} := \bigcap_{n\in\mathbb{N}}\mathcal{A}_n$, we have $\mu(\tilde{\mathcal{A}})=1$, as $\tilde{\mathcal{A}}$ is a countable intersection of measurable sets of probability one. Let $a\in\tilde{\mathcal{A}}$. Then for all $n\in\mathbb{N}$,

$$(g\circ\gamma)(F_n|a) = \int_{\mathcal{B}}\delta_{g(b)\in F_n}\,d\gamma(b|a) = \gamma(g^{-1}(F_n)|a) = \gamma'(F_n|f(a)) = (\gamma'\circ f)(F_n|a),$$

which from Proposition C.2.9 implies that $(g\circ\gamma)(\cdot|a)$ and $(\gamma'\circ f)(\cdot|a)$ coincide as probability measures on $\sigma(\{F_n\}_{n\in\mathbb{N}}) = \mathfrak{B}'$. Thus we proved that $g\circ\gamma$ and $\gamma'\circ f$ coincide μ-a.e., which from point (i) above implies that $\mu(g\circ\gamma) = \mu(\gamma'\circ f)$.

(*vii*)-(*a*). Let $\tilde{\mathcal{A}} \subseteq \mathcal{A}$, such that $\mu(\tilde{\mathcal{A}}) = 1$ and the restriction of $f^{-1}\circ f$ to $\tilde{\mathcal{A}}$ coincides with the identity on $\tilde{\mathcal{A}}$. For all $E_{\mathcal{A}} \in \mathfrak{A}$, $E_{\mathcal{B}} \in \mathfrak{B}$,

$$\begin{aligned}\Big(\mu(\gamma\circ f^{-1}\circ f)\Big)(E_{\mathcal{A}}\times E_{\mathcal{B}}) &= \int_{E_{\mathcal{A}}} (\gamma\circ f^{-1}\circ f)(E_{\mathcal{B}}|a)d\mu\\ &= \int_{E_{\mathcal{A}}\cap\tilde{\mathcal{A}}} \gamma(E_{\mathcal{B}}|f^{-1}\circ f(a))d\mu\\ &= \int_{E_{\mathcal{A}}\cap\tilde{\mathcal{A}}} \gamma(E_{\mathcal{B}}|a)d\mu\\ &= \int_{E_{\mathcal{A}}} \gamma(E_{\mathcal{B}}|a)d\mu\\ &= \mu\gamma(E_{\mathcal{A}}\times E_{\mathcal{B}}),\end{aligned}$$

so that $\mu(\gamma\circ f^{-1}\circ f) = \mu\gamma$.

(*vii*)-(*b*). Let $\tilde{\mathcal{B}} \subseteq \mathcal{B}$, such that $(\gamma\cdot\mu)(\tilde{\mathcal{B}}) = 1$ and the restriction of $g^{-1}\circ g$ to $\tilde{\mathcal{B}}$ coincides with the identity on $\tilde{\mathcal{B}}$. Then

$$(\mu\gamma)(\mathcal{A}\times\tilde{\mathcal{B}}) = (\gamma\cdot\mu)(\tilde{\mathcal{B}}) = 1,$$

and thus for $E_{\mathcal{A}} \in \mathfrak{A}$, $E_{\mathcal{B}} \in \mathfrak{B}$, using again point (*v*) above,

$$\begin{aligned}\Big(\mu(g^{-1}\circ g\circ\gamma)\Big)(E_{\mathcal{A}}\times E_{\mathcal{B}}) &= \Big(\big(\mathtt{Id}_{\mathcal{A}}\otimes(g^{-1}\circ g)\big)\cdot\mu\gamma\Big)(E_{\mathcal{A}}\times E_{\mathcal{B}})\\ &= (\mu\gamma)\Big(\big(\mathtt{Id}_{\mathcal{A}}\otimes(g^{-1}\circ g)\big)^{-1}(E_{\mathcal{A}}\times E_{\mathcal{B}})\Big)\\ &= (\mu\gamma)\Big(E_{\mathcal{A}}\times(g^{-1}\circ g)^{-1}(E_{\mathcal{B}})\Big)\\ &= (\mu\gamma)\Big(\big(E_{\mathcal{A}}\times(g^{-1}\circ g)^{-1}(E_{\mathcal{B}})\big)\cap\big(\mathcal{A}\times\tilde{\mathcal{B}}\big)\Big)\\ &= (\mu\gamma)\Big(E_{\mathcal{A}}\times\big((g^{-1}\circ g)^{-1}(E_{\mathcal{B}})\cap\tilde{\mathcal{B}}\big)\Big)\\ &= (\mu\gamma)\Big(E_{\mathcal{A}}\times(E_{\mathcal{B}}\cap\tilde{\mathcal{B}})\Big)\\ &= (\mu\gamma)\Big((E_{\mathcal{A}}\times E_{\mathcal{B}})\cap(\mathcal{A}\times\tilde{\mathcal{B}})\Big)\\ &= (\mu\gamma)\Big(E_{\mathcal{A}}\times E_{\mathcal{B}}\Big),\end{aligned}$$

so that $\mu(g^{-1}\circ g\circ\gamma) = \mu\gamma$.

(*vii*)-(*c*). Let $\tilde{\mathcal{A}} \subseteq \mathcal{A}$, $\tilde{\mathcal{B}} \subseteq \mathcal{B}$ such that $\mu(\tilde{\mathcal{A}}) = (\gamma\cdot\mu)(\tilde{\mathcal{B}}) = 1$ and f, resp. g, induces a measurable isomorphism on $\tilde{\mathcal{A}}$, resp. on $\tilde{\mathcal{B}}$, with set-theoretic inverse the map induced by f^{-1} on $f(\tilde{\mathcal{A}})$, resp. by g^{-1} on $g(\tilde{\mathcal{B}})$. Then the map $f\otimes g$ induces a measurable isomorphism from $\tilde{\mathcal{A}}\times\tilde{\mathcal{B}}$ to its image, whose set-theoretic inverse is the map induced by $(f^{-1}\otimes g^{-1})$ on $f(\tilde{\mathcal{A}})\times g(\tilde{\mathcal{B}})$. Moreover,

$$\begin{aligned}(\mu\gamma)\big((\tilde{\mathcal{A}}\times\tilde{\mathcal{B}})^{\complement}\big) &\leq (\mu\gamma)(\tilde{\mathcal{A}}^{\complement}\times\mathcal{B}) + (\mu\gamma)(\mathcal{A}\times\tilde{\mathcal{B}}^{\complement})\\ &= \mu(\tilde{\mathcal{A}}^{\complement}) + (\gamma\cdot\mu)(\tilde{\mathcal{B}}^{\complement}) = 0,\end{aligned}$$

i.e., $(\mu\gamma)(\tilde{\mathcal{A}}\times\tilde{\mathcal{B}}) = 1$. Thus $f\otimes g$ is a measured isomorphism from $(\mathcal{A}\times\mathcal{B},\mu\gamma)$ to $(\mathcal{A}'\times\mathcal{B}', (f\otimes g)\cdot(\mu\gamma))$. But, using points (*vii*)-(*a*) and (*iv*),

$$(f\otimes g)\cdot\mu\gamma = (f\otimes g)\cdot\mu(\gamma\circ f^{-1}\circ f) = (f\cdot\mu)(g\circ\gamma\circ f^{-1}),$$

which proves (vi)-(c).

(vii)-(d). It is enough to prove one implication, as the converse implication then follows by replacing μ by $f \cdot \mu$ and f by its mod 0 inverse f^{-1}. We have

$$\begin{aligned} \mu\gamma = \mu(\gamma' \circ f) \quad &\Rightarrow \quad \mu(\gamma \circ f^{-1} \circ f) = \mu(\gamma' \circ f) \\ &\Rightarrow \quad (f \otimes \mathtt{Id}_{\mathcal{B}}) \cdot \mu(\gamma \circ f^{-1} \circ f) = (f \otimes \mathtt{Id}_{\mathcal{B}}) \cdot \mu(\gamma' \circ f) \\ &\Rightarrow \quad (f \cdot \mu)(\gamma \circ f^{-1}) = (f \cdot \mu)\gamma', \end{aligned}$$

where the first line uses point (vii)-(a) above, and the last line uses point (iv) above.

$(viii)$. From point (vi)-(c), $f \otimes g^{-1}$ is a measured isomorphism from $(\mathcal{A} \times \mathcal{B}', \mu(g \circ \gamma))$ to $(\mathcal{A}' \times \mathcal{B}, m)$, where, using point (vi)-(b),

$$m := (f \cdot \mu)(g^{-1} \circ g \circ \gamma \circ f^{-1}) = (f \cdot \mu)(\gamma \circ f^{-1}),$$

and also from $(\mathcal{A} \times \mathcal{B}', \mu(\gamma' \circ f))$ to $(\mathcal{A}' \times \mathcal{B}, m')$, where, using point (vi)-(a),

$$m' := (f \cdot \mu)(g^{-1} \circ \gamma \circ f \circ f^{-1}) = (f \cdot \mu)(g^{-1} \circ \gamma).$$

Thus

$$\begin{aligned} \mu(\gamma' \circ f) = \mu(g \circ \gamma) \quad &\Leftrightarrow \quad (f \otimes g^{-1}) \cdot \mu(\gamma' \circ f) = (f \otimes g^{-1}) \cdot \mu(g \circ \gamma) \\ &\Leftrightarrow \quad (f \cdot \mu)(g^{-1} \circ \gamma') = (f \cdot \mu)(\gamma \circ f^{-1}). \end{aligned}$$

□

The following shows that measured isomorphisms "behave well" under countable tensor products:

Lemma C.3.2. *Let $(\mathcal{A}_i, \mathfrak{A}_i)_{i \in \mathcal{I}}$, $(\mathcal{B}_i, \mathfrak{B}_i)_{i \in \mathcal{I}}$ families of measurable spaces, such that either $\mathcal{I}$ is finite, or $\mathcal{I}$ is countable and $\mathcal{A}_i$, $\mathcal{B}_i$ are standard Borel for all $i \in \mathcal{I}$. Denote by $(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathfrak{A}})$, $(\boldsymbol{\mathcal{B}}, \boldsymbol{\mathfrak{B}})$ the corresponding product spaces. Let $f_i : \mathcal{A}_i \to \mathcal{B}_i$ measurable for all $i \in \mathcal{I}$, let $\boldsymbol{\mu} \in \Delta_{\boldsymbol{\mathcal{A}}}$, $\boldsymbol{\mu}' \in \Delta_{\boldsymbol{\mathcal{B}}}$, and denote by $\mu_i \in \Delta_{\mathcal{A}_i}$ the marginal of $\boldsymbol{\mu}$ on $\mathcal{A}_i$, resp. by $\mu_i' \in \Delta_{\mathcal{A}_i}$ the marginal of $\boldsymbol{\mu}'$ on $\mathcal{B}_i$. Then $\bigotimes_{i \in \mathcal{I}} f_i$ is a measured isomorphism from $(\boldsymbol{\mathcal{A}}, \boldsymbol{\mu})$ to $(\boldsymbol{\mathcal{B}}, \boldsymbol{\mu}')$ if and only if f_i is a measured isomorphism from $(\mathcal{A}_i, \mu_i)$ to $(\mathcal{B}_i, \mu_i')$ for all $i \in \mathcal{I}$ — in which case for $(g_i)_{i \in \mathcal{I}}$ a family of mod 0 inverses of each f_i, the tensor product $\bigotimes_{i \in \mathcal{I}} g_i$ is a mod 0 inverse of $\bigotimes_{i \in \mathcal{I}} f_i$.*

Note that in Proposition C.3.2 and its proof, the assumption "$\mathcal{I}$ is finite or it is countable and $\mathcal{A}_i$, $\mathcal{B}_i$ are standard Borel for all $i \in \mathcal{I}$" ensures, in particular, that the tensor products defined there are well-defined. Moreover by "countable" we mean — following usual terminology — that the set is either finite or countably infinite.

Proof. If $\bigotimes_{i \in \mathcal{I}} f_i$ is a measured isomorphism from $(\boldsymbol{\mathcal{A}}, \boldsymbol{\mu})$ to $(\boldsymbol{\mathcal{B}}, \boldsymbol{\mu}')$, then it is straightforward to verify, by marginalisation on each coordinate $i \in \mathcal{I}$, that f_i is a measured isomorphism from $(\mathcal{A}_i, \mu_i)$ to $(\mathcal{B}_i, \mu_i')$ for all $i \in \mathcal{I}$.

Conversely, assume that f_i is a measured isomorphism from $(\mathcal{A}_i, \mu_i)$ to $(\mathcal{B}_i, \mu_i')$ for all $i \in \mathcal{I}$. For all $i \in \mathcal{I}$, by assumption, there exists $\tilde{\mathcal{A}}_i \subseteq \mathcal{A}_i$, $\tilde{\mathcal{B}}_i \subseteq \mathcal{B}_i$ such that $\mu_i(\tilde{\mathcal{A}}_i) = \mu_i'(\tilde{\mathcal{B}}_i) = 1$ and f_i restricts to a measurable isomorphism $\tilde{f}_i : \tilde{\mathcal{A}}_i \to \tilde{\mathcal{B}}_i$, whose set-theoretic inverse $\tilde{f}_i^{-1}$ coincides with the restriction $\tilde{g}_i$ to $\tilde{\mathcal{B}}_i$ of a mod 0 inverse $g_i : \mathcal{B}_i \to \mathcal{A}_i$ of f_i. Note that as $\mathcal{I}$

is countable, the sets $\bigtimes_{i\in\mathcal{I}} \tilde{\mathcal{A}}_i$ and $\bigtimes_{i\in\mathcal{I}} \tilde{\mathcal{B}}_i$ are measurable.[3] Then the measurable map

$$\bigotimes_{i\in\mathcal{I}} f_i : \bigtimes_{i\in\mathcal{I}} \mathcal{A}_i \to \bigtimes_{i\in\mathcal{I}} \mathcal{B}_i$$

restricts to the measurable isomorphism

$$\bigotimes_{i\in\mathcal{I}} \tilde{f}_i : \bigtimes_{i\in\mathcal{I}} \tilde{\mathcal{A}}_i \to \bigtimes_{i\in\mathcal{I}} \tilde{\mathcal{B}}_i,$$

whose set-theoretic inverse coincides with the restriction

$$\bigotimes_{i\in\mathcal{I}} \tilde{g}_i : \bigtimes_{i\in\mathcal{I}} \tilde{\mathcal{B}}_i \to \bigtimes_{i\in\mathcal{I}} \tilde{\mathcal{A}}_i$$

of the measurable map

$$\bigotimes_{i\in\mathcal{I}} g_i : \bigtimes_{i\in\mathcal{I}} \mathcal{B}_i \to \bigtimes_{i\in\mathcal{I}} \mathcal{A}_i.$$

Moreover, as $\mu_i(\tilde{\mathcal{A}}_i) = 0$ for all $i \in \mathcal{I}$,

$$\begin{aligned}\boldsymbol{\mu}\left(\left(\bigtimes_{i\in\mathcal{I}} \tilde{\mathcal{A}}_i\right)^{\complement}\right) &\leq \boldsymbol{\mu}\left(\bigcup_{i\in\mathcal{I}}\left(\tilde{\mathcal{A}}_i^{\complement} \times \bigtimes_{i'\neq i} \mathcal{A}_{i'}\right)\right)\\ &\leq \sum_{i\in\mathcal{I}} \boldsymbol{\mu}\left(\tilde{\mathcal{A}}_i^{\complement} \times \bigtimes_{i'\neq i} \mathcal{A}_{i'}\right) = \sum_{i\in\mathcal{I}} \mu_i\left(\tilde{\mathcal{A}}_i^{\complement}\right) = 0,\end{aligned}$$

where we used again the assumption that $\mathcal{I}$ is countable. I.e., $\boldsymbol{\mu}\left(\bigtimes_{i\in\mathcal{I}} \tilde{\mathcal{A}}_i\right) = 1$. Similarly, we get $\boldsymbol{\mu}'\left(\bigtimes_{i\in\mathcal{I}} \tilde{\mathcal{B}}_i\right) = 1$. Thus we proved that $\bigotimes_{i\in\mathcal{I}} f_i$ is a measured isomorphism with mod 0 inverse $\bigotimes_{i\in\mathcal{I}} g_i$. □

C.4 Appendix for Section 3.3

C.4.1 Additional results from previous work

Here, we quote additional technical results from (Worm et al., 2011), to be used for our own proofs.

Proposition C.4.1 ((Worm et al., 2011), Proposition 2.4)**.** *Let $\mathcal{A}$ be a measurable space, and $h : \mathcal{A} \to \Delta_{\mathcal{X}} \subseteq \mathcal{X}_{\mathsf{BL}}$. Then the following are equivalent:*

(i) h is strongly measurable.

(ii) For all $F \in \mathfrak{X}$, the map

$$\begin{aligned}\mathcal{A} &\to \mathbb{R}\\ a &\mapsto h(a)(F)\end{aligned}$$

is measurable.

Proposition C.4.2. *Let $(\mathcal{A}, \mathfrak{A})$ be a measurables space, $\mu \in \Delta_{\mathcal{A}}$ and $h : \mathcal{A} \to \Delta_{\mathcal{X}} \subseteq \mathcal{X}_{\mathsf{BL}}$ Bochner integrable w.r.t. μ. Define the Bochner integral $\nu := \int_{\mathcal{A}} h(a)d\mu(a) \in \Delta_{\mathcal{X}}$. Then:*

[3] This would not necessarily be true for $\mathcal{I}$ uncountable (see our Definition C.2.4 of the product σ-algebra and Exercise 2.4.1 in (Tao, 2011)).

(i) For all bounded measurable function $f : \mathcal{X} \to \mathbb{R}$, the function

$$\begin{aligned} \mathcal{A} &\to \mathbb{R} \\ a &\mapsto \langle h(a), f \rangle \end{aligned}$$

is μ-integrable on $\mathcal{A}$, and

$$\int_{\mathcal{X}} f(x) d\nu(x) = \int_{\mathcal{A}} \langle h(a), f \rangle d\mu(a).$$

(ii) In particular, for any $E \in \mathfrak{A}$, $F \in \mathfrak{X}$,

$$\left(\int_E h(a) d\mu(a) \right)(F) = \int_E h(a)(F) d\mu(a).$$

Proof. This is Proposition 2.5 in (Worm et al., 2011). The only difference is that there, point (ii) requires $E = \mathcal{A}$. But if point (ii) is true for $E = \mathcal{A}$, then it is also true for arbitrary $E \in \mathfrak{A}$ (replace h by $h\mathbb{1}_E$, which is still Bochner integrable if h is). □

Proposition C.4.3. *For all $\mu \in \Delta_{\mathcal{X}}$, the function $x \mapsto \delta_x$ is Bochner integrable w.r.t. μ, and*

$$\mu = \int_{\mathcal{X}} \delta_x d\mu$$

Proposition C.4.4 ((Worm et al., 2011), Corollary 2.11)**.** *Let $\gamma \in \mathscr{K}(\mathcal{X})$ a Makov chain with state-space $\mathcal{X}$. Then the map*

$$\begin{aligned} \mathcal{X} &\to \mathcal{X}_{\mathsf{BL}} \\ x &\mapsto \gamma \cdot \delta_x \end{aligned}$$

is strongly measurable, and for all $\mu \in \Delta_{\mathcal{X}}$,

$$\gamma \cdot \mu = \int_{\mathcal{X}} \gamma \cdot \delta_x d\mu(x),$$

where the integral is a Bochner integral in $\mathcal{X}_{\mathsf{BL}}$.

C.4.2 Appendix for Section 3.3.3

C.4.3 Proof of Proposition 3.3.8

Proposition 3.3.8. *For any $F \in \mathfrak{X}$, the set $\mathsf{Inv}(F)$ is measurable, and if $\mu(F) = 1$ for some $\mu \in \Delta_{\mathcal{X}}$ then $\mu(\mathsf{Inv}(F)) = 1$. In particular, for all τ-stationary measure $\mu \in \Delta_{\mathcal{X}}$, we have $\mu(\mathcal{X}_{\mathrm{erg,inv}}) = 1$*

Proof. Corollary 4.2 in (Worm et al., 2011) states that for any stationary probability $\mu \in \Delta_{\mathcal{X}}$ and measurable set $F \in \mathfrak{X}$ such that $\mu(F) = 1$, there exists a measurable set $F' \subseteq F$ which is invariant and satisfies $\mu(F') = 1$. An inspection of the proof of Corollary 4.2 in (Worm et al., 2011) (and that of Lemma 4.1 on which the corollary relies) shows that we can actually choose $F' := \mathsf{Inv}(F)$. Clearly, for all $n \in \mathbb{N}$, the set F_n does not depend on μ; thus, neither does $\mathsf{Inv}(F)$, and for any $\mu \in \Delta_{\mathcal{X}}$ such that $\mu(F) = 1$, we have $\mu(\mathsf{Inv}(F)) = 1$. Applying the above reasoning to $F = \mathcal{X}_{\mathrm{erg}}$ and using Theorem 3.3.2, we get that $\mathcal{X}_{\mathrm{erg,inv}}$ is measurable and $\mu(\mathcal{X}_{\mathrm{erg,inv}}) = \mu(\mathsf{Inv}(\mathcal{X}_{\mathrm{erg}})) = 1$ for all stationary probability μ. This ends the proof of Proposition 3.3.8 □

C.4.4 Proof of Proposition 3.3.9

Proposition 3.3.9. *Let $\mathcal{X}$ standard Borel, $\tau \in \mathscr{K}(\mathcal{X})$ and assume that there exists a stationary measure. Then*

(i) $c \cap \mathcal{X}_{\text{erg,inv}} = \mathcal{X}^c \neq \emptyset$.

(ii) In particular, $\bigsqcup_{c \in C} \mathcal{X}^c = \mathcal{X}_{\text{erg,inv}}$.

Proof. (*i*). Let $\tilde{c} := c \cap \mathcal{X}_{\text{erg,inv}}$, where c is fixed. Let us first prove that $\tilde{c} \subseteq \mathcal{X}^c$. As $\mathcal{X}^c := \mathsf{Inv}(c)$ is the largest invariant subset of c (see the beginning of Section 3.3.3), it is enough to prove that $\tilde{c}$ is invariant: i.e., that for any fixed $x_0 \in \tilde{c}$, we have $\mu(\tilde{c}) = 1$, where we define $\mu := \tau \cdot \delta_{x_0}$. On the one hand, $x_0 \in \tilde{c} \subseteq c \subseteq \mathcal{X}_{\text{erg}}$ implies that the sequence $(\tau^{(n)} \cdot \delta_{x_0})_{n \in \mathbb{N}}$ converges to the ergodic measure ϵ^c in $\mathcal{X}_{\mathsf{BL}}$, and thus so does $(\tau^{(n)} \cdot \mu)_{n \in \mathbb{N}}$, as for all $n \in \mathbb{N}$,

$$\begin{aligned}
\left\| \tau^{(n)} \cdot (\tau \cdot \delta_{x_0}) - \tau^{(n)} \cdot \delta_{x_0} \right\| &= \left\| \frac{1}{n} \sum_{i=0}^{n-1} \tau^n \cdot (\tau \cdot \delta_{x_0} - \delta_{x_0}) \right\| \\
&= \left\| \frac{1}{n} (\tau^{n+1} \cdot \delta_{x_0} - \delta_{x_0}) \right\| \\
&\leq \frac{1}{n} (\|\tau^{n+1} \cdot \delta_{x_0}\| + \|\delta_{x_0}\|) \leq \frac{2}{n}.
\end{aligned}$$

But on the other hand,

$$\begin{aligned}
\lim_{n \to \infty} \tau^{(n)} \cdot \mu &= \lim_{n \to \infty} \int_{\mathcal{X}} \tau^{(n)} \cdot \delta_x d\mu(x) && \text{(C.4.1)} \\
&= \int_{\mathcal{X}_{\text{erg,inv}}} \epsilon_x d\mu(x) && \text{(C.4.2)} \\
&= \int_{\tilde{c}} \epsilon_x d\mu(x) + \int_{\mathcal{X}_{\text{erg,inv}} \setminus \tilde{c}} \epsilon_x d\mu(x) \\
&= \mu(\tilde{c}) \epsilon^c + \int_{\mathcal{X}_{\text{erg,inv}} \setminus \tilde{c}} \epsilon_x d\mu(x), && \text{(C.4.3)}
\end{aligned}$$

where:

- (C.4.1) uses Proposition C.4.4.
- (C.4.2) uses the dominated convergence theorem for Bochner integrals (Theorem C.2.13). More precisely: as $\mathcal{X}_{\text{erg,inv}} \subseteq \mathcal{X}_{\text{erg}}$, the functions $h_n : \mathcal{X} \to \mathcal{X}_{\mathsf{BL}}$ defined by

$$h_n(x) = \begin{cases} \tau^{(n)} \cdot \delta_x & \text{if } x \in \mathcal{X}_{\text{erg,inv}} \\ 0 & \text{if } x \in \mathcal{X} \setminus \mathcal{X}_{\text{erg,inv}} \end{cases}$$

converge pointwise to $h : \mathcal{X} \to \Delta_{\mathcal{X}}$ defined by

$$h(x) = \begin{cases} \epsilon_x & \text{if } x \in \mathcal{X}_{\text{erg,inv}} \\ 0 & \text{if } x \in \mathcal{X} \setminus \mathcal{X}_{\text{erg,inv}} \end{cases}$$

Moreover, from Proposition C.4.4, each map $x \mapsto \tau^{(n)} \cdot \delta_x$ from $\mathcal{X}$ to $\mathcal{X}_{\mathsf{BL}}$ is measurable, thus as $\mathcal{X}_{\text{erg,inv}}$ is a measurable set, each $h_n = (\tau^{(n)} \cdot \delta_x) \mathbb{1}_{\mathcal{X}_{\text{erg,inv}}} + \mathbb{1}_{\mathcal{X}_{\text{erg,inv}}{}^{\complement}}$ is also measurable. Each h_n is actually strongly measurable, as $\Delta_{\mathcal{X}}$ is separable and thus each subset

of it is as well. Using the pointwise convergence $\lim_{n\to\infty} h_n(x) = h(x)$, the strong measurability of each h_n, and the bound $\|h_n(x)\| \leq 1$ for all $n \in \mathbb{N}$, the dominated convergence theorem for Bochner integrals (Theorem C.2.13) yields that

$$\lim_{n\to\infty} \int_{\mathcal{X}} h_n(x) d\mu(x) = \int_{\mathcal{X}} h(x) d\mu(x)$$

However, by invariance of $\mathcal{X}_{\text{erg,inv}}$, $x_0 \in \mathcal{X}_{\text{erg,inv}}$ implies $\mu(\mathcal{X}_{\text{erg,inv}}) = (\tau \cdot \delta_{x_0})(\mathcal{X}_{\text{erg,inv}}) = 1$,[4] which yields

$$\int_{\mathcal{X}} h(x) d\mu(x) = \int_{\mathcal{X}_{\text{erg,inv}}} h(x) d\mu(x) = \int_{\mathcal{X}_{\text{erg,inv}}} \epsilon_x d\mu(x),$$

but also

$$\int_{\mathcal{X}} h_n(x) d\mu(x) = \int_{\mathcal{X}_{\text{erg,inv}}} h_n(x) d\mu(x) = \int_{\mathcal{X}_{\text{erg,inv}}} \tau^{(n)} \cdot \delta_x d\mu(x) = \int_{\mathcal{X}} \tau^{(n)} \cdot \delta_x d\mu(x).$$

- (C.4.3) uses that $\epsilon_x = \epsilon^c$ for all $x \in \tilde{c} \subseteq c$.

By unicity of the limit in $\mathcal{X}_{\mathsf{BL}}$, we obtain

$$(1 - \mu(\tilde{c}))\epsilon^c = \int_{\mathcal{X}_{\text{erg,inv}} \setminus \tilde{c}} \epsilon_x d\mu(x).$$

In particular, using Proposition C.4.2,

$$(1 - \mu(\tilde{c}))\epsilon^c(\tilde{c}) = \int_{\mathcal{X}_{\text{erg,inv}} \setminus \tilde{c}} \epsilon_x(\tilde{c}) d\mu(x). \tag{C.4.4}$$

But as ϵ^c is stationary, we have $\epsilon^c(\mathcal{X}_{\text{erg,inv}}) = 1$ from Proposition 3.3.8. Thus $\epsilon^c(\tilde{c}) = \epsilon^c(c \cap \mathcal{X}_{\text{erg,inv}}) = \epsilon^c(c) = 1$, where the last equality uses Theorem 3.3.4. Similarly, for all $x \in \mathcal{X}_{\text{erg,inv}} \setminus \tilde{c}$, the distribution ϵ_x is stationary and thus $\epsilon_x(\tilde{c}) = \epsilon_x(c \cap \mathcal{X}_{\text{erg,inv}}) = \epsilon_x(c) = 0$, with the last equality using Theorem 3.3.4 and $\mathcal{X}_{\text{erg,inv}} \setminus \tilde{c} \subseteq \mathcal{X}_{\text{erg}} \setminus c$. Injecting this in (C.4.4) yields $\mu(\tilde{c}) = 1$: i.e., $\tilde{c}$ is invariant. As $\tilde{c} \subseteq c$ and $\mathcal{X}^c = \mathsf{Inv}(c)$ is the largest invariant set contained in c, this implies $\tilde{c} \subseteq \mathcal{X}^c$.

To prove $\tilde{c} = \mathcal{X}^c$, we now need to prove the converse inclusion $\mathcal{X}^c \subseteq \tilde{c}$. By definition,

$$\mathcal{X}^c := \mathsf{Inv}(c) := \bigcap_{n\in\mathbb{N}} c_n,$$
$$\mathcal{X}_{\text{erg,inv}} := \bigcap_{n\in\mathbb{N}} \mathcal{X}_{\text{erg},n},$$

where $c_0 := c$, $\mathcal{X}_{\text{erg},0} := \mathcal{X}_{\text{erg}}$, and for all $n \in \mathbb{N}$,

$$c_n := \{x \in c_{n-1} : (\tau \cdot \delta_x)(c_{n-1}) = 1\},$$
$$\mathcal{X}_{\text{erg},n} := \{x \in \mathcal{X}_{\text{erg},n-1} : (\tau \cdot \delta_x)(\mathcal{X}_{\text{erg},n-1}) = 1\}.$$

But as $c_0 = c \subseteq \mathcal{X}_{\text{erg}} = \mathcal{X}_{\text{erg}_0}$, it follows, by iteration, that $c_n \subseteq \mathcal{X}_{\text{erg},n}$ for all $n \in \mathbb{N}$, and thus $\mathcal{X}^c \subseteq \mathcal{X}_{\text{erg,inv}}$. In particular, as $\mathcal{X}^c = \mathsf{Inv}(c) = \mathsf{Inv}(c) \subseteq c$,

$$\mathcal{X}^c = \mathcal{X}^c \cap c \subseteq \mathcal{X}_{\text{erg,inv}} \cap c = \tilde{c}.$$

[4] Crucially, had we only assumed that $x_0 \in \mathcal{X}_{\text{erg}}$, the latter fact might not always hold if $\mathcal{X}_{\text{erg}} \setminus \mathcal{X}_{\text{erg,inv}} \neq \emptyset$.

Eventually, c, as an equivalence class of $\mathcal{X}_{\text{erg}}$, is by definition non-empty, and from Point (i) in Theorem 3.3.7, for a fixed $x \in c \neq \emptyset$, we have $\epsilon_x(\tilde{c}) = \epsilon_x(\mathcal{X}^c) = 1$, which implies that $\tilde{c} \neq \emptyset$.

(ii). Recall that each $c \in C$ is an equivalence class for the relation $\sim$ defined on $\mathcal{X}_{\text{erg}}$ (see equation (3.3.2)). Thus the result is a direct consequence of point (i), as each $\mathcal{X}^c = c \cap \mathcal{X}_{\text{erg,inv}}$ coincides with a non-empty equivalence class w.r.t. the relation $\sim$ restricted to $\mathcal{X}_{\text{erg}}$. This ends the proof of Proposition 3.3.9. □

C.4.5 Appendix for Section 3.3.4

Here we present the results described informally in Section 3.3.4. Let us recall that the push-forward operator τ_* defined by $\tau \in \mathscr{K}(\mathcal{X})$ sends each positive measure $\mu \in \mathscr{M}_{\mathcal{X}}^+$ to $\gamma_*\mu := \gamma \cdot \mu \in \mathscr{M}_{\mathcal{X}}^+$, and that $\mathscr{M}_{\mathcal{X}}^+ \subset \mathcal{X}_{\mathsf{BL}}$ is equipped with the topology induced by the restriction of the norm topology on $\mathcal{X}_{\mathsf{BL}}$ (see Definition 3.2.2 and Section 3.3.2). Let us also recall that $\mathcal{X}_{\mathsf{BL}}$ is a Polish space for the norm topology $\mathcal{T}_{\|\cdot\|}$ (see Section 3.3.2), and thus $(\mathcal{X}_{\mathsf{BL}}, \mathcal{T}_{\|\cdot\|}, \sigma(\mathcal{T}_{\|\cdot\|}))$ is a standard Borel space.

Theorem C.4.5. *Let $\tau \in \mathscr{K}(\mathcal{X})$ such that $\tau_* : \mathscr{M}_{\mathcal{X}}^+ \to \mathscr{M}_{\mathcal{X}}^+$ is continuous for the norm induced by $\mathcal{X}_{\mathsf{BL}}$. The following holds:*

(i) The space $\mathsf{Erg}(\tau) \subseteq \mathcal{X}_{\mathsf{BL}}$ is a countable intersection of open sets for the induced topology on the set of stationary probability measures. In particular, it is measurable in the standard Borel space $\mathcal{X}_{\mathsf{BL}}$.

(ii) Denoting by resp. $\mathcal{T}$ and $\mathfrak{E}$ the topology and σ-algebra induced by $\mathcal{X}_{\mathsf{BL}}$ on $\mathsf{Erg}(\tau)$, there exists a topology $\mathcal{T}'$ on $\mathsf{Erg}(\tau)$ such that $\mathcal{T} \subseteq \mathcal{T}'$, $\text{Bor}_{\mathcal{T}'} = \mathfrak{E}$ and $(\mathsf{Erg}(\tau), \mathcal{T}', \mathfrak{E})$ is standard Borel.

(iii) This standard Borel structure on $\mathsf{Erg}(\tau)$ induces one on the set C that indexes the partition in ergodic components $(\mathcal{X}^c)_{c \in C}$ of the generic set $\mathcal{X}_{\text{erg,inv}} \subseteq \mathcal{X}$.

Point (i) is the core of the proof. Point (ii) is a consequence of (i) and Theorem C.2.6, which here yields, in short, that $\mathsf{Erg}(\tau)$ can be equipped with a standard Borel structure that looses no information about the induced topology or about the induced Borel σ-algebra in $\mathcal{X}_{\mathsf{BL}}$. Point (ii) then yields the most important result for us: point (iii), which means that under the continuity assumption, the space C — our candidate for the pose coordinate's space — indeed has the "nice" structure that we sought. Note that if τ is defined by a deterministic transformation on $\mathcal{X}$, then the continuity assumption means exactly that the transformation is continuous on $\mathcal{X}$, in the usual sense (see Section 2.3 in (Worm et al., 2011)).

The proof of Theorem C.4.5 will call on several results — either from (Worm et al., 2011) or from standard textbooks — that we need to introduce first. We denote by $\mathsf{BM}_{\mathcal{X}}$ and $\mathsf{BC}_{\mathcal{X}}$ the sets of resp. bounded measurable and bounded continuous functions from $\mathcal{X}$ to $\mathbb{R}$. We also recall the notation $\langle \mu, f \rangle := \int_{\mathcal{X}} f \, d\mu$, where the integral is well-defined for $\mu \in \mathscr{M}_{\mathcal{X}}^+$ and $f \in \mathsf{BM}_{\mathcal{X}}$. The following is a "dual" point of view on the continuity assumption in Theorem C.4.5:

Lemma C.4.6 ((Worm et al., 2011), Section 2.2 and Proposition 2.14)**.** *For any $\tau \in \mathscr{K}(\mathcal{X})$, the folloing holds:*

(i) There exists a unique bounded linear map τ^* from $\mathsf{BM}_{\mathcal{X}}$ to itself such that for all $\mu \in \mathscr{M}_{\mathcal{X}}$ and $f \in \mathsf{BM}_{\mathcal{X}}$, we have $\langle \tau_*\mu, f \rangle = \langle \mu, \tau^* f \rangle$.

(ii) $\tau_* : \mathscr{M}_{\mathcal{X}}^+ \to \mathscr{M}_{\mathcal{X}}^+$ is continuous if and only if $\tau^*(\mathsf{BC}_{\mathcal{X}}) \subseteq \mathsf{BC}_{\mathcal{X}}$.

(For readers interested in comparing the statement above to the original statements in (Worm et al., 2011), let us point out that there, channels $\tau \in \mathscr{K}(\mathcal{X})$ are called *transition probabilities*; push-forward operators τ_* — or more precisely, their extension to $\mathscr{M}^+_{\mathcal{X}}$ — are *regular Markov operators*; and τ^* is the *dual* of τ_*.)

Point (i) of the next result is cited in (Worm et al., 2011), but requires a quick introduction. For $\mu \in \Delta_{\mathcal{X}}$, the map $f \mapsto \langle \mu, f \rangle$ defines not only a bounded linear form on $\mathsf{BL}_{\mathcal{X}}$ (see Section 3.3.2), but also one on the Banach space $(\mathsf{BC}_{\mathcal{X}}, \|\cdot\|_{\mathsf{BC}})$ of bounded continous functions, with the sup norm $\|\cdot\|_{\mathsf{BC}} := \|\cdot\|_\infty$. As $\mathsf{BL}_{\mathcal{X}} \subseteq \mathsf{BC}_{\mathcal{X}}$, the operator on $\mathsf{BC}_{\mathcal{X}}$ restricts to the one on $\mathsf{BL}_{\mathcal{X}}$, which from Section 3.3.2 corresponds to a unique measure in $\mathscr{M}^+_{\mathcal{X}}$. Thus each $\mu \in \mathscr{M}^+_{\mathcal{X}}$ defines a *unique* bounded linear form on $\mathsf{BC}_{\mathcal{X}}$: i.e., $\mathscr{M}^+_{\mathcal{X}}$ identifies not only to a subset of $\mathsf{BL}^*_{\mathcal{X}}$, but also to a subset of $\mathsf{BC}^*_{\mathcal{X}}$. The interest of this new identification is to bring in a new, well-studied topology on $\mathscr{M}^+$: that induced by the weak-$*$ topology on $\mathsf{BC}^*_{\mathcal{X}}$. This new topology happens to coincide with the one previously defined (i.e., the one induced by the norm topology on $\mathcal{X}_{\mathsf{BL}} \subseteq \mathsf{BL}^*_{\mathcal{X}}$):

Lemma C.4.7. *The following holds:*

(i) The restriction to $\mathscr{M}^+_{\mathcal{X}}$ of the norm topology on $\mathcal{X}_{\mathsf{BL}} \subseteq \mathsf{BL}^*_{\mathcal{X}}$ coincides with the restriction to $\mathscr{M}^+_{\mathcal{X}}$ of the weak-$*$ topology on $\mathsf{BC}^*_{\mathcal{X}}$.

(ii) In particular, for $\mu, \mu_1, \mu_2, \ldots \in \mathscr{M}^+_{\mathcal{X}} \subseteq \mathcal{X}_{\mathsf{BL}}$,

$$\lim_{n\to\infty} \mu_n = \mu \text{ in } \mathcal{X}_{\mathsf{BL}} \quad \Leftrightarrow \quad \forall f \in \mathsf{BC}_{\mathcal{X}}, \ \lim_{n\to\infty} \langle \mu_n, f \rangle = \langle \mu, f \rangle.$$

and for $\mu, \mu' \in \mathscr{M}^+_{\mathcal{X}}$,

$$\mu' = \mu \quad \Leftrightarrow \quad \forall f \in \mathsf{BC}_{\mathcal{X}}, \ \langle \mu', f \rangle = \langle \mu, f \rangle.$$

(iii) $\|\mu\|_{\mathsf{BC}^*} = \|\mu\|_{\mathsf{BL}^*}$ for all $\mu \in \mathscr{M}^+_{\mathcal{X}}$, where $\|\cdot\|_{\mathsf{BC}^*}$ and $\|\cdot\|_{\mathsf{BL}^*}$ are the usual dual norms.

(iv) $\mathscr{M}^+_{\mathcal{X}}$ is closed for the weak-$*$ topology in $\mathsf{BC}^*_{\mathcal{X}}$.

Proof. Point (i) is Theorem 18 in (Dudley, 1966), and point (ii) is a direct corollary using the definition of the weak-$*$ topology. Point (iii) is straightforward: from the definitions of the norms $\|\cdot\|_{\mathsf{BC}} = \|\cdot\|_\infty$ and $\|\cdot\|_{\mathsf{BL}} = \|\cdot\|_\infty + \|\cdot\|_{\mathsf{Lip}}$, we clearly have $\|\mu\|_{\mathsf{BC}^*} \leq \|\mu\|_{\mathsf{BL}^*} \leq \mu(\mathcal{X})$ for $\mu \in \mathscr{M}^+$, and the inequalities become equalities by considering $f = \mathbb{1}_{\mathcal{X}}$. Point (iv) is straightforward as well. □

The following is the Banach-Alaoglu theorem, a classic theorem in functional analysis.

Lemma C.4.8 ((Brezis, 2011), Theorem 3.16)**.** *Let $\mathcal{B}$ be a Banach space with norm $\|\cdot\|$, $\mathcal{B}^*$ its dual, and $\|\cdot\|^*$ the usual dual norm. Then the unit ball of $\mathcal{B}^*$, i.e., the set*

$$\mathsf{Ball}_{\mathcal{B}^*}(1) := \{l \in \mathcal{B}^* \ : \ \|l\|^* \leq 1\},$$

is compact for the weak-$*$ topology.

Eventually, let us point out that the set of extreme points of a convex set is not necessarily a Borel set. However, we have the following (see Lemma 7.63 in (Aliprantis et al., 2006), and references therein for counter-examples):

Lemma C.4.9. *If K is a metrisable compact subset of a topological vector space, then the set of extreme points of K is a countable intersection of open sets in K, i.e., it is a countable intersection of open sets in K.*

While we did not explicitly define topological vector spaces, here we only need to know that any Banach space — in particular, $\mathcal{X}_{\mathsf{BL}}$ — is one.

Proof of Theorem C.4.5. Given point (i), point (ii) is an application of Theorem C.2.6 to the standard Borel space $\mathcal{A} := \mathcal{X}_{BL}$. Moreover, from point (iv) in Theorem 3.3.4, there is a bijection between the space of ergodic probabilities $\mathsf{Erg}(\tau)$ and the set $\mathcal{C}$, which we can use to transfer the standard Borel structure on $\mathsf{Erg}(\tau)$ to one on $\mathcal{C}$. So that we only need to prove point (i).

Let $\mu_1, \mu_2, ... \in \Delta_{\mathcal{X}}$ a sequence of stationary probabilities that converges to $\mu \in \Delta_{\mathcal{X}}$ in $\mathcal{X}_{\mathsf{BL}}$. From point (ii) in Lemma C.4.7, this is equivalent to:

$$\forall h \in \mathsf{BC}_{\mathcal{X}}, \quad \langle \mu, h \rangle = \lim_{n\to\infty} \langle \mu_n, h \rangle \tag{C.4.5}$$

Then for all $f \in \mathsf{BC}_{\mathcal{X}}$,

$$\begin{aligned}
\langle \tau_* \mu, f \rangle &= \langle \mu, \tau^* f \rangle && \text{(C.4.6)}\\
&= \lim_{n\to\infty} \langle \mu_n, \tau^* f \rangle && \text{(C.4.7)}\\
&= \lim_{n\to\infty} \langle \tau_* \mu_n, f \rangle && \text{(C.4.8)}\\
&= \lim_{n\to\infty} \langle \mu_n, f \rangle && \text{(C.4.9)}\\
&= \langle \mu, f \rangle, && \text{(C.4.10)}
\end{aligned}$$

where equations (C.4.6) and (C.4.8) use point (i) in Lemma C.4.6; equation (C.4.7) uses Theorem C.4.5's assumption that $\tau_* : \mathscr{M}^+_{\mathcal{X}} \to \mathscr{M}^+_{\mathcal{X}}$ is continuous, point (ii) in Lemma C.4.6 which yields $\tau^* f \in \mathsf{BC}_{\mathcal{X}}$, and equation (C.4.5); equation (C.4.9) uses the stationarity of each μ_n; and equation (C.4.10) uses equation (C.4.5) again. Using point (ii) in Lemma C.4.7, this proves that the set of invariant probabilities is closed for the norm topology in $\mathcal{X}_{\mathsf{BL}}$.

On the other hand, from Lemma C.4.8, the unit ball $\mathsf{Ball}_{\mathsf{BC}^*}(1)$ of $\mathsf{BC}^*_{\mathcal{X}}$ is compact for the weak-$*$ topology. But as $\mathscr{M}^+_{\mathcal{X}}$ is a closed subset of $\mathsf{BC}^*_{\mathcal{X}}$ (point (iv) in Lemma C.4.7), the set $\mathscr{M}^+_{\mathcal{X}} \cap \mathsf{Ball}_{\mathsf{BC}^*_{\mathcal{X}}}(1)$ is also compact for the weak-$*$ topology in $\mathsf{BC}^*_{\mathcal{X}}$. Yet from point (iii) in Lemma C.4.7, we have

$$\mathscr{M}^+_{\mathcal{X}} \cap \mathsf{Ball}_{\mathsf{BC}^*_{\mathcal{X}}}(1) = \mathscr{M}^+_{\mathcal{X}} \cap \mathsf{Ball}_{\mathsf{BL}^*_{\mathcal{X}}}(1),$$

and combining this with point (i) in the same Lemma, we get that $\mathscr{M}^+_{\mathcal{X}} \cap \mathsf{Ball}_{\mathsf{BL}^*_{\mathcal{X}}}(1)$ is compact for the restriction of the norm topology on $\mathcal{X}_{\mathsf{BL}}$.

Therefore, the set of stationary probability measures is a closed subset of a compact subset in $\mathcal{X}_{\mathsf{BL}}$, so that it is compact. Moreover, as a closed subset of $\mathcal{X}_{\mathsf{BL}}$, it is metrisable (take the restriction of the norm metric in $\mathcal{X}_{\mathsf{BL}}$). We can now apply Lemma C.4.9 to conclude that the corresponding set of extreme points is a countable intersection of open sets, for the topology induced by $\mathcal{X}_{\mathsf{BL}}$ on the set of stationary probabilities. Yet, from point (i) in Proposition 3.3.4, this set of extreme points is exactly $\mathsf{Erg}(\tau)$, which ends the proof of Theorem C.4.5. □

C.4.6 Appendix for Section 3.3.5

Proof of Proposition 3.3.12

Proposition 3.3.12. *Let τ be a measurable Markov chain with standard Borel state-space $\mathcal{X}$. Then for any probability $\mu \in \Delta_{\mathcal{X}}$ such that $\mu(\mathcal{X}_{\mathsf{erg,inv}}) = 1$, denoting by $q := q(\overrightarrow{X})$ the*

corresponding process distribution, we have, in the Banach space $(\mathcal{X}\times\mathcal{X})_{\mathsf{BL}}$*, the convergence*

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{i=0}^{n-1} q(X_0, X_i) = \int_{\mathcal{X}} \epsilon_x \otimes \epsilon_x \, d\mu(x), \tag{3.3.4}$$

where we used the tensor product notation (see Definition 3.2.6).

Our proof will use the following lemma, which is intuitive but has, to our knowledge, not been proven before. It states that the Bochner integral commutes with hook-ups:

Lemma C.4.10. *Let* $(\mathcal{A}, \mathfrak{A})$ *and* $(\mathcal{B}, \mathfrak{B})$ *be standard Borel spaces, and* $\mu \in \mathcal{A}_{\mathsf{BL}}$*. Assume that there exists a Bochner integrable map*

$$\begin{aligned} \mathcal{A} &\to \Delta_{\mathcal{A}} \subseteq \mathcal{X}_{\mathsf{BL}} \\ a &\mapsto \xi_a \end{aligned}$$

such that

$$\mu = \int_{\mathcal{A}} \xi_a d\mu(a). \tag{C.4.11}$$

Then, for any channel $\gamma \in \mathscr{K}(\mathcal{A}, \mathcal{B})$*, the map*

$$\begin{aligned} \mathcal{A} &\to \Delta_{\mathcal{A}\times\mathcal{B}} \subseteq (\mathcal{A}\times\mathcal{B})_{\mathsf{BL}} \\ a &\mapsto \xi_a\gamma \end{aligned}$$

is Bochner integrable, and we have, in the Banach space $(\mathcal{A}\times\mathcal{B})_{\mathsf{BL}}$*, the equality*

$$\mu\gamma = \int_{\mathcal{A}} \xi_a\gamma \, d\mu(a),$$

where we used the hook-up notation (see Definition 3.2.3).

Proof. Let $E \in \mathfrak{A}\otimes\mathfrak{B}$. Define $f : \mathcal{A} \to [0,1]$ by $f(a_0) := \gamma(E_{a_0}|a_0)$, where E_{a_0} is the section of E at a_0, i.e.,

$$E_{a_0} := \{b \in \mathcal{B} : \ (a_0, b) \in E\}.$$

From Proposition C.2.15, the bounded function f is measurable and for all $a \in \mathcal{A}$,

$$\langle \xi_a, f\rangle := \int_{\mathcal{A}} f(a_0) d\xi_a(a_0) = \xi_a\gamma(E) \tag{C.4.12}$$

In particular, using point (i) in Proposition C.4.2, the bounded function

$$\begin{aligned} \mathcal{A} &\to \mathbb{R} \\ a &\mapsto \langle \xi_a, f\rangle = \xi_a\gamma(E) \end{aligned}$$

is measurable as well, and

$$\mu\gamma(E) = \int_{\mathcal{A}} \gamma(E_{a_0}|a_0) d\mu(a_0) \tag{C.4.13}$$
$$= \int_{\mathcal{A}} f(a_0) d\mu(a_0)$$
$$= \int_{\mathcal{A}} \langle \xi_a, f \rangle d\mu(a), \tag{C.4.14}$$

where line (C.4.13) uses Proposition C.2.15; while line (C.4.14) uses the assumption (C.4.11) and point (i) in Proposition C.4.2. Using equation (C.4.12), this yields

$$\mu\gamma(E) = \int_{\mathcal{A}} \xi_a \gamma(E) d\mu(a). \tag{C.4.15}$$

But as the measurability of the function $a \mapsto \xi_a\gamma(E)$ holds for all $E \in \mathfrak{X} \otimes \mathfrak{X}$, Proposition C.4.1 shows that the map

$$\mathcal{A} \to \Delta_{\mathcal{A}\times\mathcal{B}} \subseteq (\mathcal{A} \times \mathcal{B})_{\mathsf{BL}}$$
$$a \mapsto \xi_a\gamma$$

is strongly measurable. As it is also bounded in norm and μ is a probability, this map is Bochner integrable w.r.t. μ (see Definition C.2.12). We can thus apply point (ii) from Proposition C.4.2 in the space $(\mathcal{A} \times \mathcal{B})_{\mathsf{BL}}$, and obtain, for all $E \in \mathfrak{A} \otimes \mathfrak{B}$,

$$\int_{\mathcal{A}} \xi_a\gamma(E) d\mu(a) = \left(\int_{\mathcal{A}} \xi_a\gamma \, d\mu(a) \right)(E),$$

i.e., from equation (C.4.15),

$$\mu\gamma(E) = \left(\int_{\mathcal{A}} \xi_a\gamma \, d\mu(a) \right)(E).$$

□

Proof of Proposition 3.3.12. For all $x_0 \in \mathcal{X}_{\text{erg,inv}}$, using point (iii) in Lemma C.3.1,

$$\delta_{x_0}\tau^{(n)} = \delta_{x_0} \otimes (\tau^{(n)} \cdot \delta_{x_0}). \tag{C.4.16}$$

But as $x_0 \in \mathcal{X}_{\text{erg,inv}} \subseteq \mathcal{X}_{\text{erg}}$, the sequence $(\tau^{(n)} \cdot \delta_{x_0})_n$ converges, in $\mathcal{X}_{\mathsf{BL}}$, to the ergodic measure ϵ_{x_0}. This implies that, for all $x_0 \in \mathcal{X}_{\text{erg,inv}}$,

$$\lim_{n\to\infty} \left(\delta_{x_0}\tau^{(n)} \right) = \delta_{x_0} \otimes \epsilon_{x_0} \quad \text{in } (\mathcal{X} \times \mathcal{X})_{\mathsf{BL}}. \tag{C.4.17}$$

Indeed, from point (ii) in Lemma C.4.7, the convergence, in $\mathcal{X}_{\mathsf{BL}}$, of $(\tau^{(n)} \cdot \delta_{x_0})_n$ to ϵ_{x_0} implies

$$\lim_{n\to\infty} \langle \tau^{(n)} \cdot \delta_{x_0}, f \rangle = \langle \epsilon_{x_0}, f \rangle$$

for all bounded continuous function $f : \mathcal{X} \to \mathbb{R}$. Thus, if now we have a bounded continuous function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defining $f_{x_0} : \mathcal{X} \to \mathbb{R}$ as $f_{x_0}(x) := f(x_0, x)$, we obtain a bounded

continuous function on $\mathcal{X}$, so that

$$\begin{aligned}\lim_{n\to\infty} \langle \delta_{x_0} \otimes (\tau^{(n)} \cdot \delta_{x_0}), f \rangle &= \lim_{n\to\infty} \langle \tau^{(n)} \cdot \delta_{x_0}, f_{x_0} \rangle \\ &= \langle \epsilon_{x_0}, f_{x_0} \rangle \\ &= \langle \delta_{x_0} \otimes \epsilon_{x_0}, f \rangle.\end{aligned}$$

Thus, using point (ii) in Lemma C.4.7 again, but now in the space $(\mathcal{X} \times \mathcal{X})_{\mathsf{BL}}$ (see Remark 3.3.11), we obtain

$$\lim_{n\to\infty} \left(\delta_{x_0} \otimes (\tau^{(n)} \cdot \delta_{x_0}) \right) \;=\; \delta_{x_0} \otimes \epsilon_{x_0} \quad \text{in } (\mathcal{X} \times \mathcal{X})_{\mathsf{BL}},$$

which, combined with equation (C.4.16), does yield (C.4.17). Now fix $x_0 \in \mathcal{X}_{\mathsf{erg,inv}}$, and let us prove that

$$\lim_{n\to\infty} \left(\epsilon_{x_0} \tau^{(n)} \right) \;=\; \epsilon_{x_0} \otimes \epsilon_{x_0} \quad \text{in } (\mathcal{X} \times \mathcal{X})_{\mathsf{BL}}. \tag{C.4.18}$$

We have

$$\begin{aligned}\epsilon_{x_0} \tau^{(n)} &= \left(\int_{\mathcal{X}} \delta_x d\epsilon_{x_0}(x) \right) \tau^{(n)} \\ &= \int_{\mathcal{X}} \delta_x \tau^{(n)} \, d\epsilon_{x_0}(x)\end{aligned}$$

where the first line uses Proposition C.4.3, and the second line Lemma C.4.10. Thus for all $x_0 \in \mathcal{X}_{\mathsf{erg,inv}}$ and $c \in \mathcal{C}$ such that $x_0 \in \mathcal{X}^c$,

$$\begin{aligned}\lim_{n\to\infty} \left(\epsilon_{x_0} \tau^{(n)} \right) &= \int_{\mathcal{X}} \delta_x \otimes \epsilon_x \, d\epsilon_{x_0}(x) \\ &= \int_{\mathcal{X}^c} \delta_x \otimes \epsilon_x \, d\epsilon_{x_0}(x) \\ &= \int_{\mathcal{X}^c} \delta_x \otimes \epsilon_{x_0} \, d\epsilon_{x_0}(x) \\ &= \left(\int_{\mathcal{X}^c} \delta_x \, d\epsilon_{x_0}(x) \right) \otimes \epsilon_{x_0} \\ &= \epsilon_{x_0} \otimes \epsilon_{x_0},\end{aligned}$$

where the first line uses the convergence in (C.4.17) and the dominated convergence theorem for Bochner integrals (Theorem C.2.13); the second line uses $\epsilon_{x_0}(\mathcal{X}^c) = 1$ (point (ii) in Theorem 3.3.7); the third line uses $\epsilon_x = \epsilon_{x_0}$ for all $x, x_0 \in \mathcal{X}^c$ (by definition of the relation $\sim$ in (3.3.2)), the fourth line uses Lemma C.4.10 again, and the last line Proposition C.4.3 again. Thus we proved the convergence in (C.4.18).

Eventually, for all $\mu \in \Delta_{\mathcal{X}}$,

$$\begin{aligned}\mu \tau^{(n)} &= \left(\int_{\mathcal{X}} \epsilon_x d\mu(x) \right) \tau^{(n)} \\ &= \int_{\mathcal{X}} \epsilon_x \tau^{(n)} \, d\mu(x),\end{aligned}$$

where the first line uses Theorem 3.3.5, and the second line Lemma C.4.10. Therefore, using the convergence in (C.4.18) and dominated convergence for Bochner integrals again, if

$\mu(\mathcal{X}_{\text{erg,inv}}) = 1$, then we have, in $(\mathcal{X} \times \mathcal{X})_{\mathsf{BL}}$, the convergence

$$\lim_{n\to\infty} \left(\mu\tau^{(n)}\right) = \int_{\mathcal{X}} \epsilon_x \otimes \epsilon_x d\mu(x).$$

Eventually, note that for all $n \in \mathbb{N}$, by linearity of hook-ups w.r.t. the channel argument,

$$\mu\tau^{(n)} = \frac{1}{n}\sum_{i=0}^{n-1} \mu\tau^i = \frac{1}{n}\sum_{i=0}^{n-1} q(X_0, X_i).$$

This ends the proof of Proposition 3.3.12. □

Proof of Proposition 3.3.13

Proposition 3.3.13. *Assume that $\mathcal{X}$ is countable, $\tau \in \mathscr{K}(\mathcal{X})$, fix a probability $\mu \in \Delta_{\mathcal{X}}$ such that $\mu(\mathcal{X}_{\text{erg,inv}}) = 1$, denote by $q = q(\overrightarrow{X})$ the corresponding process distribution, and write also*

$$\bar{q}^n(X, X') := \frac{1}{n}\sum_{i=0}^{n-1} q(X_0, X_i) \in \Delta_{\mathcal{X}\times\mathcal{X}},$$

Then, for all $x, x' \in \mathcal{X}_{\text{erg,inv}}$,

$$\lim_{n\to\infty} \bar{q}^n(x, x') = \bar{q}(x, x'),$$

where

$$\bar{q}(x, x') := \sum_{c\in C} \mu(\mathcal{X}^c)\, \epsilon^c(x)\epsilon^c(x')\delta_{x,x'\in\mathcal{X}^c}. \tag{3.3.5}$$

Proof. For $\mathcal{A}$ countable, using the definition of $\mathcal{A}_{\mathsf{BL}}$ as a subset of the dual of the space of bounded Lipschitz functions, it is straightforward to verify that the topology induced by $\mathcal{A}_{\mathsf{BL}}$ on $\Delta_{\mathcal{A}}$ coincides with the usual topology, i.e., that induced by the ambient space $\mathbb{R}^{|\mathcal{A}|+1}$. In particular, the convergence in $(\mathcal{X} \times \mathcal{X})_{\mathsf{BL}}$ is equivalent to the symbol-wise convergence of probabilities. Thus, for all $x, x' \in \mathcal{X}_{\text{erg,inv}}$,

$$\begin{aligned}
\lim_{n\to\infty} \bar{q}^n(x, x') &= \sum_{x''\in\mathcal{X}} \mu(x'')\epsilon_{x''}(x)\epsilon_{x''}(x') \\
&= \sum_{x''\in\mathcal{X}_{\text{erg,inv}}} \mu(x'')\epsilon_{x''}(x)\epsilon_{x''}(x') \\
&= \sum_{c\in C}\sum_{x''\in\mathcal{X}^c} \mu(x'')\epsilon_{x''}(x)\epsilon_{x''}(x') \\
&= \sum_{c\in C} \mu(\mathcal{X}^c)\epsilon^c(x)\epsilon^c(x') \\
&= \sum_{c\in C} \mu(\mathcal{X}^c)\epsilon^c(x)\epsilon^c(x')\delta_{x,x'\in\mathcal{X}^c} = \bar{q}(x, x'),
\end{aligned}$$

where the last line recalls that each ϵ^c is concentrated on $\mathcal{X}^c$ (see Theorem 3.3.7). This proves equation (3.3.5). Now, marginalising over X',

$$\mu(x) = \bar{q}(x) = \sum_{c\in C} \mu(\mathcal{X}^c)\epsilon^c(x)\delta_{x\in\mathcal{X}^c},$$

so that, continuing from the previous equation,

$$\begin{aligned}\bar{q}(x,x') &= \sum_{c\in C}\mu(x)\epsilon^c(x')\delta_{x,x'\in\mathcal{X}^c}\\ &= \mu(x)\sum_{c\in C}\delta_{x\in\mathcal{X}^c}\epsilon^c(x')\\ &= \mu(x)\sum_{c\in C}\delta_{\mathsf{pr}(x)=c}\epsilon(x'|c)\\ &= \mu(x)\,(\epsilon\circ\mathsf{pr})(x'|x)\\ &= \big[\mu(\epsilon\circ\mathsf{pr})\big](x,x'),\end{aligned}$$

where the second line uses that $\epsilon^c(x) = \epsilon^c(x)\delta_{x'\in\mathcal{X}^c}$. This ends the proof of Proposition 3.3.13. □

Proof of Proposition 3.3.15

Proposition 3.3.15. *Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ countable sets,let $q_{AB}\in\Delta_{\mathcal{A}\times\mathcal{B}}$ and $f:\mathcal{A}\to\mathcal{C}$. Then f is a sufficient statistic of A w.r.t. B if and only if there exists a channel $\gamma\in\mathscr{K}(\mathcal{C},\mathcal{B})$ such that, denoting by q_A the marginal of q_{AB} on $\mathcal{A}$ and using the hook-up notation (see Definition 3.2.3), we have $q_{AB} = q_A(\gamma\circ f)$.*

Proof. Assume that f is a sufficient statistic of A w.r.t. B. From the Markov chain $A-C-B$, there exists a channel $\gamma\in\mathscr{K}(\mathcal{C},\mathcal{B})$ such that, for all $(a,b,c)\in\mathcal{A}\times\mathcal{B}\times\mathcal{C}$,

$$q_{AB}(a,b)\delta_{f(a)=c} = q_A(a)\delta_{f(a)=c}\gamma(b|c).$$

Marginalising to $\mathcal{A}\times\mathcal{B}$, we obtain $q_{AB} = q_A(\gamma\circ f)$. Conversely, assume that there exists $\gamma\in\mathscr{K}(\mathcal{C},\mathcal{B})$ such that $q_{AB} = q_A(\gamma\circ f)$. Then for all $(a,b,c)\in\mathcal{A}\times\mathcal{B}\times\mathcal{C}$,

$$\begin{aligned}q_{AB}(a,b)\delta_{f(a)=c} &= q_A(a)\left(\sum_{c'\in C}\delta_{f(a)=c'}\gamma(b|c')\right)\delta_{f(a)=c}\\ &= q_A(a)\delta_{f(a)=c}\gamma(b|c),\end{aligned}$$

which proves the Markov chain $A-C-B$. This ends the proof of Proposition 3.3.15. □

Proof of Proposition 3.3.17

Proposition 3.3.17. *Let $\mu\in\Delta_{\mathcal{X}}$ such that $\mu(\mathcal{X}_{\text{erg,inv}}) = 1$, define $\bar{q}\in\Delta_{\mathcal{X}\times\mathcal{X}}$ as in (3.3.6), and let $\mathcal{T}$ be a countable space. Then for all $\kappa\in\mathscr{K}(\mathcal{X},\mathcal{T})$ such that $\bar{q} = \mu(\gamma\circ\kappa)$ for some channel $\gamma\in\mathscr{K}(\mathcal{T},\mathcal{X})$, there exists a function $h:\mathcal{T}\to\mathcal{C}$ such that $\mu\mathsf{pr} = \mu(h\circ\kappa)$.*

Proof. Using Proposition 3.3.13, we have

$$\mu(\epsilon\circ\mathsf{pr}) = \bar{q} = \mu(\gamma\circ\kappa),$$

and therefore, from point (iv) in Lemma C.3.1,

$$\mu(\mathsf{pr}\circ\epsilon\circ\mathsf{pr}) = \mu(\mathsf{pr}\circ\gamma\circ\kappa).$$

But as for all $c\in\mathcal{C}$, the ergodic distribution $\epsilon(\cdot|c) = \epsilon^c$ is concentrated on the ergodic component $\mathcal{X}^c$ (see Theorem 3.3.7), we have $\mathsf{pr}\circ\epsilon = \mathsf{Id}_{\mathcal{C}}$. This yields

$$\mu\mathsf{pr} = \mu(\mathsf{pr}\circ\gamma\circ\kappa).$$

But as pr is deterministic, this implies that the restriction of the channel $\mathsf{pr}\circ\gamma\circ\kappa$ to the support $\mathrm{supp}(\mu)$ of μ is deterministic. It can easily be verified that this implies that the restriction of $\mathsf{pr}\circ\gamma$ to the support $\mathrm{supp}(\kappa\cdot\mu)$ of $\kappa\cdot\mu\in\Delta_{\mathcal{T}}$ is a deterministic function $h_+ : \mathrm{supp}(\kappa\cdot\mu)\to\mathcal{X}$. Let us now fix an arbitrary function $h_0 : \mathrm{supp}(\kappa\cdot\mu)^{\complement}\to\mathcal{X}$ defined on the complement $\mathrm{supp}(\kappa\cdot\mu)^{\complement}$ of $\mathrm{supp}(\kappa\cdot\mu)$, and define the function h by

$$h : \mathcal{T}\to\mathcal{X}$$
$$t\mapsto\begin{cases} h_+(t) & \text{if } t\in\mathrm{supp}(\kappa\cdot\mu),\\ h_0(t) & \text{if } t\in\mathrm{supp}(\kappa\cdot\mu)^{\complement}.\end{cases}$$

As $\mathsf{pr}\circ\gamma$ and h coincide on $\mathrm{supp}(\kappa\cdot\mu)$, the composed channels $\mathsf{pr}\circ\gamma\circ\kappa$ and $h\circ\kappa$ coincide on $\mathrm{supp}(\mu)$. Therefore,

$$\mu(h\circ\kappa)=\mu(\mathsf{pr}\circ\gamma\circ\kappa)=\mu\mathsf{pr}.$$

This ends the proof of Proposition 3.3.17. □

Proof of Theorem 3.3.18

Theorem 3.3.18. *Under the mean-asympotic distribution $\bar{q}(X,X')$, the projection on ergodic components pr is a minimal sufficient statistic of X w.r.t. X', and of X' w.r.t. X.*

Proof. As already mentioned, the equality $\bar{q}=\mu(\epsilon\circ\mathsf{pr})$ from Proposition 3.3.13 shows that pr is a sufficient statistic of X w.r.t. X'. From Proposition 3.3.15, any other sufficient statistic $f : \mathcal{X}\to\mathcal{T}$ satisfies $\bar{q}=\mu(\gamma\circ f)$ for some channel $\gamma\in\mathscr{K}(\mathcal{T},\mathcal{X})$. From Proposition 3.3.17, this implies $\mu\mathsf{pr}=\mu(h\circ f)$ for some $h : \mathcal{T}\to\mathcal{X}$, i.e., that pr and $h\circ f$ coincide on $\mathrm{supp}(\mu)$. This proves that pr is a minimal sufficient statistic of X w.r.t. X'. Eventually, using the symmetry $\bar{q}=\bar{q}^{\mathsf{T}}$, we obtain that pr is also a minimal sufficient statistic of X' w.r.t. X. This ends the proof of Theorem 3.3.18. □

C.5 Proofs for Section 3.4

C.5.1 Proof of Theorem 3.4.1

Theorem 3.4.1. *Let (π,ρ) be a standard Borel measurable MDP, such that Assumption 1 holds. Denote by $\overrightarrow{\pi\rho}\in\mathscr{K}(\mathcal{X},\overrightarrow{\mathcal{G}\times\mathcal{X}})$ the corresponding process channel (see Definition 3.2.12). Let $(\mathcal{X}^c)_{c\in C}$ be the decomposition into ergodic components, $\kappa : \mathcal{X}\to C$ the corresponding projection, and $(\pi^c)_{c\in C}$ the family of restrictions of π to each $\mathcal{X}^c$. Then there exists a family $(\rho^c)_{c\in C}$, where $\rho^c\in\mathscr{K}(\mathcal{X}^c\times\mathcal{G},\mathcal{X}^c)$ for all $c\in C$, such that:*[5]

(i) Denoting by $\overrightarrow{\pi^c\rho^c}\in\mathscr{K}(\mathcal{X}^c,\overrightarrow{\mathcal{G}\times\mathcal{X}^c})$ the corresponding process channel of each measurable MDP (π^c,ρ^c), the map

$$\begin{aligned}\mathcal{X} &\to \Delta_{\overrightarrow{\mathcal{X}\times\mathcal{G}}}\subseteq(\overrightarrow{\mathcal{X}\times\mathcal{G}})_{\mathsf{BL}}\\ x &\mapsto \epsilon^{\kappa(x)}\,\overrightarrow{\pi^{\kappa(x)}\rho^{\kappa(x)}}\end{aligned}$$

is Bochner integrable, and for all stationary $\mu_0\in\Delta_{\mathcal{X}}$,

$$\mu_0\,\overrightarrow{\pi\rho}=\int_{\mathcal{X}}\epsilon^{\kappa(x)}\,\overrightarrow{\pi^{\kappa(x)}\rho^{\kappa(x)}}\,d\mu_0(x). \tag{3.4.1}$$

[5] Here, we identify distributions and channels to their restriction or extension to the relevant space (see Definitions C.2.8 and 3.2.2).

In particular, for all $c \in C$, we have $\epsilon^c \overrightarrow{\pi\rho} = \epsilon^c \overrightarrow{\pi^c \rho^c}$.

(ii) For all $c \in C$, denote by $\overline{\rho^c} := \rho^c \circ (\mathrm{Id}_{\mathcal{X}^c} \between \pi^c)$ the update channel of each measurable MDP (π^c, ρ^c), and recall that $\bar{\rho}^c$ is the restriction of $\bar{\rho} \in \mathscr{K}(\mathcal{X})$ to $\mathcal{X}^c$. Then $\overline{\rho^c} = \bar{\rho}^c$ holds ϵ^c-a.e.. In particular, $\epsilon^c \in \Delta_{\mathcal{X}^c}$ is the unique stationary distribution w.r.t. $\overline{\rho^c}$, and it is ergodic w.r.t. $\overline{\rho^c}$.

Moreover, an arbitrary family $(\tilde{\rho}^c)_{c\in C}$ satisfies points (i) and (ii) above if and only if for all $c \in C$, we have $\tilde{\rho}^c \in \mathscr{K}(\mathcal{X}^c \times \mathcal{G}, \mathcal{X}^c)$ and $\tilde{\rho}^c = \rho$ holds $\epsilon^c \pi^c$-a.e..

The proof of Theorem 3.4.1 will involve the following lemmas. Let us recall (see Definition 3.2.10 and remarks above) that for q the process distribution of a measured MDP and all $m, n \in \mathbb{N}$,

$$q_m^{n+1} = q(X_m, G_m, \dots, X_n, G_n, X_{n+1}) = (q_m^n \pi)\rho = q_m^n(\pi\rho), \tag{C.5.1}$$

where π, resp. ρ, is here seen as a channel from $\mathcal{X}_n$ to $\mathcal{G}_n$, resp. from $\mathcal{X}_n \times \mathcal{G}_n$ to $\mathcal{X}_{n+1}$; while $\pi\rho \in \mathscr{K}(\mathcal{X}_n, \mathcal{G}_n \times \mathcal{X}_{n+1})$ is a hook-up of channels and the last two terms in (C.5.1) also use the hook-up notation (see Definition 3.2.3). The following is intuitive but requires a verification in our general standard Borel setting:

Lemma C.5.1. *The process distribution q of a standard Borel measured MDP (μ_0, π, ρ) is uniquely defined by the sequence of marginals $(q_n^{n+1})_{n\in\mathbb{N}}$. In particular, if μ_0 is $\bar{\rho}$-stationary, then the process distribution q is uniquely determined by q_0^1.*

Proof. Let us prove iteratively that for all $n \in \mathbb{N}$, the distribution q_0^n depends only on $\{q_i^{i+1}\}_{0\le i\le n}$; from the Kolmogorov extension theorem (Theorem C.2.10), this will yield point (ii). The result holds for $n = 0$; assume that it holds for some $n \ge 1$, and let us show that $\{q_i^{i+1}\}_{0\le i\le n+1}$ uniquely defines q_0^{n+1}. Intuitively, we want to prove that if we take the conditional law of (G_n, X_{n+1}) given X_n (using the joint distribution $q(X_n, G_n, X_{n+1})$, without any direct reference to π or ρ), then q_0^{n+1} is defined by the hook-up of that conditional law with q_0^n. The remaining of the proof formalises this intuition in our standard Borel setting.

From Proposition C.2.16, there exists some $\gamma_q \in \mathscr{K}(\mathcal{X}_n, \mathcal{G}_n \times \mathcal{X}_{n+1})$ such that $q_n^{n+1} = q_n \gamma_q$, where $q_n := q(X_n)$. Importantly, here γ_q is uniquely defined by q_n^{n+1}, without any direct reference to π or ρ. From equation (C.5.1), we have $q_n\gamma_q = q_n^{n+1} = q_n(\pi\gamma)$. Thus, if we fix $\overrightarrow{F} \in \mathfrak{G}_n \otimes \mathfrak{X}_{n+1}$, then for all $F_n \in \mathfrak{X}_n$,

$$\int_{F_n} \pi\rho(\overrightarrow{F}|x_n) dq_n(x_n) = \int_{F_n} \gamma_q(\overrightarrow{F}|x_n) dq_n(x_n)$$

From Lemma C.2.11, this implies the existence of a set $\tilde{\mathcal{X}}_n$ such that $q_n(\tilde{\mathcal{X}}_n) = 1$ and $\pi\rho(\overrightarrow{F}|x_n) = \gamma_q(\overrightarrow{F}|x_n)$ for all $x_n \in \tilde{\mathcal{X}}_n$. Moreover, as q_n is a marginal of q_0^n, defining the measurable set

$$\widetilde{\mathsf{Past}} := \mathcal{X}_0^{n-1} \times \tilde{\mathcal{X}}_n \times \mathcal{G}_0^{n-1} \subseteq \mathcal{X}_0^n \times \mathcal{G}_0^{n-1}$$

we have $q_0^n(\widetilde{\mathsf{Past}}) = 1$. Thus, for all $\overleftarrow{F} \in \mathfrak{X}_0^n \otimes \mathfrak{G}_0^{n-1}$,

$$\begin{aligned} q_0^n \pi\rho(\overleftarrow{F} \times \overrightarrow{F}) &= \int_{\overleftarrow{F}} \pi\rho(\overrightarrow{F}|x_n) d q_0^n(x_0^n, g_0^{n-1}) \\ &= \int_{\overleftarrow{F} \cap \widetilde{\mathsf{Past}}} \pi\rho(\overrightarrow{F}|x_n) d q_0^n(x_0^n, g_0^{n-1}) \\ &= \int_{\overleftarrow{F} \cap \widetilde{\mathsf{Past}}} \gamma_q(\overrightarrow{F}|x_n) d q_0^n(x_0^n, g_0^{n-1}) \\ &= \int_{\overleftarrow{F}} \gamma_q(\overrightarrow{F}|x_n) d q_0^n(x_0^n, g_0^{n-1}) \\ &= q_0^n \gamma_q(\overleftarrow{F} \times \overrightarrow{F}). \end{aligned}$$

Note that here the set $\widetilde{\mathsf{Past}}$ might depend on $\overrightarrow{F}$, but the reasoning as a whole still allows to choose both $\overrightarrow{F} \in \mathfrak{G}_n \otimes \mathfrak{X}_{n+1}$ and $\overleftarrow{F} \in \mathfrak{X}_0^n \otimes \mathfrak{G}_0^{n-1}$ arbitrarily. Thus we proved that $q_0^n \pi\rho = q_0^n \gamma_q$, i.e., $q_0^{n+1} = q_0^n \gamma_q$. But from our recurrence assumption, q_0^n is uniquely defined by $(q_i^{i+1})_{0 \le i \le n}$, while γ_q has been uniquely defined by q_n^{n+1}. Thus q_0^{n+1} is uniquely defined by $\{q_i^{i+1}\}_{0 \le i \le n+1}$. I.e., we showed by iteration that the process distribution q is uniquely defined by the sequence of marginals $(q_n^{n+1})_{n \in \mathbb{N}}$. If moreover μ_0 is $\bar{\rho}$-stationary, then it is straightforward to verify that $q_n^{n+1} = q_m^{m+1}$ for all $m, n \in \mathbb{N}$, hence the second part of the result. □

As noted in Section 3.3.5, the results quoted in Section 3.3.2 about the space $\mathcal{X}_{\mathsf{BL}}$ actually only assumed that $\mathcal{X}$ is standard Borel, and thus $\mathcal{X}$ can be replaced by any other standard Borel space. In particular, we can consider the Banach space $(\overrightarrow{\mathcal{X} \times \mathcal{G}})_{\mathsf{BL}}$, as $\overrightarrow{\mathcal{X} \times \mathcal{G}}$ is a countable product of standard Borel spaces and thus standard Borel itself (see Proposition C.2.5).

Lemma C.5.2. *Under the same assumptions and notations as Theorem 3.4.1, the map*

$$\begin{aligned} \mathcal{X} &\to \Delta_{\overrightarrow{\mathcal{X} \times \mathcal{G}}} \subseteq (\overrightarrow{\mathcal{X} \times \mathcal{G}})_{\mathsf{BL}} \\ x &\mapsto \epsilon^{\kappa(x)} \overrightarrow{\pi\rho} \end{aligned}$$

is Bochner integrable, and for all stationary $\mu_0 \in \Delta_{\mathcal{X}}$, we have

$$\mu_0 \overrightarrow{\pi\rho} = \int_{\mathcal{X}} \epsilon^{c(x)} \overrightarrow{\pi\rho} d\mu_0(x).$$

Proof. From Theorem 3.3.5, the map $x \mapsto \epsilon_x \in \Delta_{\mathcal{X}}$ is Bochner integrable and

$$\mu_0 = \int_{\mathcal{X}} \epsilon_x d\mu_0(x).$$

The result is thus a consequence of the commutation of Bochner integrals with hook-ups (see Lemma C.4.10). □

Proof of Theorem 3.4.1. (*i*). Let $c \in C$. We know that $\mathcal{X}^c$ is $\bar{\rho}$-invariant: i.e., for all $x \in \mathcal{X}^c$, we have $\bar{\rho}(\mathcal{X} \setminus \mathcal{X}^c | x) = 0$. This implies that

$$\int_{\mathcal{X}^c \times \mathcal{G}} \rho(\mathcal{X} \setminus \mathcal{X}^c | x, g) d(\epsilon^c \pi)(x, g) = \int_{\mathcal{X}^c} \bar{\rho}(\mathcal{X} \setminus \mathcal{X}^c | x) d\epsilon^c(x) = 0.$$

Thus, there exists a set $E^c \in \mathfrak{X} \otimes \mathfrak{G}$ such that $(\epsilon^c \pi)(E^c) = 1$ and $\rho(\mathcal{X} \setminus \mathcal{X}^c | x, g) = 0$ for all $(x, g) \in E^c$. Let us fix some arbitrary channel $\rho_0^c \in \mathscr{K}\left((\mathcal{X}^c \times \mathcal{G}) \setminus E^c, \mathcal{X}^c\right)$, and define, for

all F in the induced σ-algebra $\mathfrak{X}^c := \mathfrak{X} \cap \mathcal{X}^c$, for all $x \in \mathcal{X}^c$, $g \in \mathcal{G}$,

$$\rho^c(F|x,g) := \begin{cases} \rho(F|x,g) & \text{if } (x.g) \in E^c, \\ \rho_0^c(F|x,g) & \text{if } (x.g) \in (\mathcal{X}^c \times \mathcal{G}) \setminus E^c. \end{cases} \tag{C.5.2}$$

As E^c is measurable and ρ is a channel, we obtain a channel $\rho^c \in \mathscr{K}(\mathcal{X}^c \times \mathcal{G}, \mathcal{X}^c)$. As ϵ^c is concentrated on $\mathcal{X}^c$ and $\mathcal{X}^c$ is τ-invariant where here $\tau = \bar{\rho}$, the distribution $\epsilon^c \bar{\rho} \in \Delta_{\mathcal{X}\times\mathcal{X}}$ is concentrated on $\mathcal{X}^c \times \mathcal{X}^c$. But as $\epsilon^c \bar{\rho}$ is the marginal on $\mathcal{X} \times \mathcal{X}$ of $\epsilon^c \pi\rho \in \Delta_{\mathcal{X}\times\mathcal{G}\times\mathcal{X}}$, this implies that $\epsilon^c \, \pi\rho$ is concentrated on $\mathcal{X}^c \times \mathcal{G} \times \mathcal{X}^c$. Thus for all $F_{\mathcal{X}}, F'_{\mathcal{X}} \in \mathfrak{X}$, and $F_{\mathcal{G}} \in \mathfrak{G}$,

$$\begin{aligned}
(\epsilon^c \pi\rho)(F_{\mathcal{X}} \times F_{\mathcal{G}} \times F'_{\mathcal{X}}) &= \left(\epsilon^c \overrightarrow{\pi\rho}\right)\left((F_{\mathcal{X}} \cap \mathcal{X}^c) \times F_{\mathcal{G}} \times (F'_{\mathcal{X}} \cap \mathcal{X}^c)\right) \\
&= \int_{(F_{\mathcal{X}} \cap \mathcal{X}^c)\times F_{\mathcal{G}}} \rho(F'_{\mathcal{X}} \cap \mathcal{X}^c|x,g) d(\epsilon^c \pi)(x,g) \\
&= \int_{\left((F_{\mathcal{X}} \cap \mathcal{X}^c)\times F_{\mathcal{G}}\right) \cap E^c} \rho(F'_{\mathcal{X}} \cap \mathcal{X}^c|x,g) d(\epsilon^c \pi)(x,g) \\
&= \int_{\left((F_{\mathcal{X}} \cap \mathcal{X}^c)\times F_{\mathcal{G}}\right) \cap E^c} \rho^c(F'_{\mathcal{X}} \cap \mathcal{X}^c|x,g) d(\epsilon^c \pi)(x,g) \\
&= \int_{\left((F_{\mathcal{X}} \cap \mathcal{X}^c)\times F_{\mathcal{G}}\right)} \rho^c(F'_{\mathcal{X}} \cap \mathcal{X}^c|x,g) d(\epsilon^c \pi)(x,g) \\
&= (\epsilon^c \pi^c \rho^c)\left((F_{\mathcal{X}} \cap \mathcal{X}^c) \times F_{\mathcal{G}} \times (F'_{\mathcal{X}} \cap \mathcal{X}^c)\right).
\end{aligned}$$

Therefore, if we still denote by $\epsilon^c \overrightarrow{\pi^c \rho^c}$ the extension of the process distribution $\epsilon^c \overrightarrow{\pi^c \rho^c} \in \Delta_{\overrightarrow{\mathcal{X}^c \times \mathcal{G}}}$ to $\overrightarrow{\mathcal{X} \times \mathcal{G}}$, the above proves that $\left(\epsilon^c \overrightarrow{\pi\rho}\right)_0^1 = \left(\epsilon^c \overrightarrow{\pi^c \rho^c}\right)_0^1$. But as ϵ^c is $\bar{\rho}$-stationary, this means, from Lemma C.5.1, that $\epsilon^c \overrightarrow{\pi\rho} = \epsilon^c \overrightarrow{\pi^c \rho^c}$. As $c \in C$ is arbitrary, we can combine the latter with Proposition C.5.2, yielding that the map

$$\begin{aligned}
\mathcal{X} &\to \Delta_{\overrightarrow{\mathcal{X}\times\mathcal{G}}} \\
x &\mapsto \epsilon^{\kappa(x)} \overrightarrow{\pi^{\kappa(x)} \rho^{\kappa(x)}}
\end{aligned}$$

is Bochner integrable in the Banach space $(\overrightarrow{\mathcal{X} \times \mathcal{G}})_{\mathsf{BL}}$, and for all stationary $\mu_0 \in \Delta_{\mathcal{X}}$,

$$\mu_0 \overrightarrow{\pi\rho} = \int_{\mathcal{X}} \epsilon^{\kappa(x)} \overrightarrow{\pi^{\kappa(x)} \rho^{\kappa(x)}} d\mu_0(x).$$

As $\kappa(x) = c$ for all $x \in \mathcal{X}^c$, this clearly implies that $\epsilon^c \overrightarrow{\pi\rho} = \epsilon^c \overrightarrow{\pi^c \rho^c}$ for all $c \in C$.

(*ii*). For all $c \in \mathcal{C}$, $x \in \mathcal{X}^c$, $F, F' \in \mathfrak{X}$, $n \in \mathbb{N}$,

$$\begin{aligned}
\epsilon^c \bar{\rho}^c(F \times F') &= \int_F \bar{\rho}^c(F_n|x) d\epsilon^c(x) \\
&= \int_{F\times\mathcal{G}} \rho(F_n|x,g) d(\epsilon^c \pi^c)(x,g) \\
&= \int_{(F\times\mathcal{G})\cap E^c} \rho(F_n|x,g) d(\epsilon^c \pi^c)(x,g) \\
&= \int_{(F\times\mathcal{G})\cap E^c} \rho^c(F_n|x,g) d(\epsilon^c \pi^c)(x,g) \\
&= \int_{F\times\mathcal{G}} \rho^c(F_n|x,g) d(\epsilon^c \pi^c)(x,g) \\
&= \int_F \overline{\rho^c}(F_n|x) d\epsilon^c(x) \\
&= \epsilon^c \overline{\rho^c}(F \times F'),
\end{aligned}$$

where the set E^c has been defined in the proof of point (*ii*). Thus $\epsilon^c \bar{\rho}^c = \epsilon^c \overline{\rho^c}$, and as the spaces are here standard Borel, point (*i*) in Lemma C.3.1 implies that $\bar{\rho}^c = \overline{\rho^c}$ holds ϵ^c-a.e.. The second part of the statement is a consequence of point (*iii*) in Theorem 3.3.7.

For the last part of the statement, observe than the family $(\rho^c)_{c\in C}$ that we constructed does satisfy, for all $c \in \mathcal{C}$, that $\rho^c \in \mathscr{K}(\mathcal{X}^c \times \mathcal{G}, \mathcal{X}^c)$ and that $\rho^c = \rho$ holds $\epsilon^c \pi^c$-ae.. Moreover, an inspection of the proof shows that these are the only properties that we need to carry out the proof (observe that, in equation (C.5.2), the channel ρ_0^c is chosen arbitrarily).

This ends the proof of Theorem 3.4.1. □

C.5.2 Proof of Theorem 3.4.6

Theorem 3.4.6. *Let $\mathcal{X}$, $\mathcal{G}$ standard Borel spaces such that $\mathcal{G}$ is a measurable group with a group-stationary probability $\upsilon \in \Delta_{\mathcal{G}}$, let $\rho \in \mathscr{K}(\mathcal{X}\times\mathcal{G}, \mathcal{X})$ a measurable action, and consider the measurable MDP (π_υ, ρ). Then, for ϵ_x defined in (3.4.2):*

(i) For $x \in \mathcal{X}$ and all $n \geq 1$, we have $\bar{\rho}^n \cdot \delta_x = \epsilon_x$; in particular, $\lim_{n\to\infty} \bar{\rho}^{(n)} \cdot \delta_x = \epsilon_x$ in $\mathcal{X}_{\mathsf{BL}}$.

(ii) $\mathcal{X} = \mathcal{X}_{\mathrm{erg,inv}}$,

(iii) The partition in ergodic components $\{\mathcal{X}^c\}_{c\in\mathcal{C}}$ coincides with the partition in orbits of $\mathcal{X}$ w.r.t. the action of $\mathcal{G}$

(iv) For all $x \in \mathcal{X}^c = [x]$, the probability ϵ_x is the unique $\bar{\rho}$-stationary probability such that $\epsilon_x(\mathcal{X}^c) = 1$, and it is ergodic w.r.t. $\bar{\rho}$.

(v) A probability μ on $\mathcal{X}$ is is $\bar{\rho}$-stationary if and only if it is ρ_g-stationary for all $g \in \mathcal{G}$.

(vi) The probability υ is the unique group-stationary probability on $\mathcal{G}$.

(vii) Denoting by ρ^c the restriction of the group action ρ to each orbit $\mathcal{X}^c$, the family of MDPs $(\epsilon^c, \pi_\upsilon^c, \rho^c)$ is an ergodic decomposition of the measurable MDP (π_υ, ρ).

Let us start with the following lemma:

Lemma C.5.3. *For $x, x' \in \mathcal{X}$, $\epsilon_x([x]) = 1$ and*

$$\epsilon_x = \epsilon_{x'} \quad \Leftrightarrow \quad [x] = [x'].$$

Proof. For all $g \in \mathcal{G}$, $F \in \mathfrak{X}$,

$$\mathcal{G}_{g\cdot x\to F} = \{g' \in \mathcal{G} : \; g'g \cdot x \in F\} = \{g' \in \mathcal{G} : \; g' \cdot x \in F\}\, g^{-1} = \mathcal{G}_{x\to F} g^{-1},$$

As υ is here stationary, this implies $\upsilon(\mathcal{G}_{g\cdot x\to F}) = \upsilon(\mathcal{G}_{x\to F})$. Thus $[x] = [x']$ implies $\epsilon_x = \epsilon_{x'}$. But if $x' \notin [x]$, then

$$\epsilon_{x'}([x]) = \upsilon(\mathcal{G}_{x'\to[x]}) = \upsilon(\emptyset) = 0,$$

while

$$\epsilon_x([x]) = \upsilon(\mathcal{G}_{x\to[x]}) = \upsilon(\mathcal{G}) = 1.$$

Thus $[x'] \neq [x]$ implies $\epsilon_{x'} \neq \epsilon_x$. □

Proof of Theorem 3.4.6. (*i*). Let $x \in \mathcal{X}$. Equation (3.4.2) implies that

$$(\bar{\rho} \cdot \delta_x) = \bar{\rho}(\cdot|x) = \epsilon_x.$$

Moreover, for all $F \in \mathfrak{X}$,

$$\begin{aligned}(\bar{\rho} \cdot \epsilon_x)(F) &= \int_{\mathcal{X}} \bar{\rho}(F|x')d\epsilon_x(x') \\ &= \int_{\mathcal{X}} \epsilon_{x'}(F)d\epsilon_x(x') \\ &= \int_{[x]} \epsilon_{x'}(F)d\epsilon_x(x') \\ &= \epsilon_x(F),\end{aligned}$$

where the third equality uses $\epsilon_x([x]) = 1$ and the last one uses $\epsilon_{x'} = \epsilon_x$ for $x' \in [x]$ (both proved in Lemma C.5.3 above). Thus, by iteration, $\bar{\rho}^n \cdot \epsilon_x = \epsilon_x$ for all $n \in \mathbb{N}$. In particular, $\lim_{n\to\infty} \bar{\rho}^n \cdot \delta_x = \epsilon_x$ in $\mathcal{X}_{\mathsf{BL}}$. But in any Banach space, if a sequence converges to a limit, then the Césaro means of that sequence converge to the same limit. Thus $\lim_{n\to\infty} \bar{\rho}^{(n)} \cdot \delta_x = \epsilon_x$ in $\mathcal{X}_{\mathsf{BL}}$.

(*ii*). Recall that $\mathcal{X}_{\text{erg,inv}} = \mathsf{Inv}(\mathcal{X}_{\text{erg}})$, where $x \in \mathcal{X}_{\text{erg}}$ if and only if $\bar{\rho}^{(n)} \cdot \delta_x$ converges, in $\mathcal{X}_{\mathsf{BL}}$, to an ergodic probability w.r.t. $\bar{\rho}$. Let us prove that $\mathcal{X}_{\text{erg}} = \mathcal{X}$, which will also prove that $\mathcal{X}_{\text{erg}}$ is invariant and thus $\mathcal{X}_{\text{erg,inv}} = \mathsf{Inv}(\mathcal{X}_{\text{erg}}) = \mathcal{X}_{\text{erg}} = \mathcal{X}$. Let $x_0 \in \mathcal{X}$; from point (*ii*) above, we need to prove that ϵ_{x_0} is ergodic w.r.t. $\bar{\rho}$. First, from point (*ii*) again, we know that ϵ_{x_0} is $\bar{\rho}$-invariant. Let F be a $\bar{\rho}$-invariant set: i.e., $(\bar{\rho} \cdot \delta_x)(F) = 1$ for all $x \in F$; and let us prove that $\epsilon_{x_0}(F) = 0$ or $\epsilon_{x_0}(F) = 1$. We saw in the proof of point (*ii*) that $\bar{\rho} \cdot \delta_x = \epsilon_x$, so that we have

$$\forall x \in F, \quad \epsilon_x(F) = (\bar{\rho} \cdot \delta_x)(F) = 1. \tag{C.5.3}$$

On the other hand, from Lemma C.5.3, we know that $\epsilon_{x_0}([x_0]) = 1$, so that $\epsilon_{x_0}(F) = \epsilon_{x_0}(F \cap [x_0])$. Thus if $F \cap [x_0] = \emptyset$, then $\epsilon_{x_0}(F) = 0$. Otherwise, there exists $x_0' \in [x_0] \cap F$. But then $x_0' \in F$ implies, from (C.5.3), that $\epsilon_{x_0'}(F) = 1$, while Lemma C.5.3 above implies $\epsilon_{x_0}(F) = \epsilon_{x_0'}(F)$, i.e., $\epsilon_{x_0}(F) = 1$. Thus we proved that for any $\bar{\rho}$-invariant set F, the quantity $\epsilon_{x_0}(F)$ is either 0 or 1: i.e., that ϵ_{x_0} is ergodic w.r.t. $\bar{\rho}$.

(*iii*). Point (*i*) and (*ii*) together show that for all $x \in \mathcal{X}_{\text{erg,inv}} = \mathcal{X}$, we have

$$[x] = [x'] \quad \Leftrightarrow \quad \lim_{n\to\infty} \bar{\rho}^{(n)} \cdot \delta_x = \lim_{n\to\infty} \bar{\rho}^{(n)} \cdot \delta_x.$$

Thus the equivalence classes $[x]$ are defined by the same equivalence relation as that defining those of the partition $(\mathcal{X}^c)_{c\in C}$ of $\mathcal{X}_{\text{erg,inv}} = \mathcal{X}_{\text{erg}} = \mathcal{X}$ (see equation (3.3.2) and Proposition 3.3.9).

(iv). We saw in the proof of point (iii) that each ϵ_x is stationary, and, from Lemma C.5.3, we have $\epsilon_x([x]) = 1$. However, from point (ii) in Theorem 3.3.7, there is only one such probability. We also proved in the proof of point (iii) that ϵ_x is ergodic for all $x \in \mathcal{X}$.

(v). Clearly, if μ is stationary under each $g \in \mathcal{G}$ then it is stationary under the update channel $\bar{\rho}$, which averages the actions over the group $\mathcal{G}$. Assume now that μ is $\bar{\rho}$-stationary. Then for all $F \in \mathfrak{X}$, $g \in \mathcal{G}$,

$$\begin{aligned}
\mu(g^{-1}\cdot F) = (\bar{\rho}\cdot\mu)(g^{-1}\cdot F) &= \int_{\mathcal{X}} \bar{\rho}(g^{-1}\cdot F|x)d\mu(x)\\
&= \int_{\mathcal{X}} \upsilon(\mathcal{G}_{x\to g^{-1}\cdot F})d\mu(x)\\
&= \int_{\mathcal{X}} \upsilon(g^{-1}\mathcal{G}_{x\to F})d\mu(x)\\
&= \int_{\mathcal{X}} \upsilon(\mathcal{G}_{x\to F})d\mu(x)\\
&= \int_{\mathcal{X}} \bar{\rho}(F|x)d\mu(x)\\
&= (\bar{\rho}\cdot\mu)(F) = \mu(F),
\end{aligned}$$

where the fourth equality uses the left-stationarity of υ w.r.t. the action of $\mathcal{G}$.

(vi). Let us apply the results above to $\mathcal{X} = \mathcal{G}$, and ρ defined by the action of $\mathcal{G}$ on itself by mutliplication on the right. As there is only one orbit under this action, from point (iv), there is only one ergodic component. Moreover, for all $g \in \mathcal{G}$, $F \in \mathfrak{G}$,

$$\epsilon_g(F) = \upsilon(\{g' \in \mathcal{G} \,:\, g'g \in F\}) = \upsilon(g^{-1}(F)) = \upsilon(F),$$

where we used the stationarity. Therefore, from point (v), the probability υ is the unique $\bar{\rho}$-stationary probability such that $\upsilon(\mathcal{G}) = 1$, i.e., the unique $\bar{\rho}$-stationary probability. From point (vi), this is equivalent to υ being the unique probability such that $\upsilon(Fg) = \upsilon(F)$, for all $F \in \mathfrak{G}$. Proceeding similarly with the action of $\mathcal{G}$ on itself on the left, we see that υ is the unique probability such that $\upsilon(gF) = \upsilon(F)$ for all $g \in \mathcal{G}$, $F \in \mathfrak{G}$. Thus υ is the unique stationary probability.

(vii). This is a consequence of the last part of the statement of Theorem 3.4.1.

This ends the proof of Theorem 3.4.6. □

C.6 Appendix for Section 3.5

C.6.1 Appendix for Section 3.5.1

Proof of Proposition 3.5.2

Proposition 3.5.2. *Let (μ_0, π, ρ) and (μ_0', π', ρ') be stationary MDPs, with resp. state-spaces $\mathcal{X}$, $\mathcal{X}'$ and resp. action spaces $\mathcal{G}$, $\mathcal{G}'$. Let $\phi : \mathcal{X} \to \mathcal{X}'$ and $\psi : \mathcal{G} \to \mathcal{G}'$ be measurable maps, and consider the following statements:*

(i) (μ_0', π', ρ') is a factor of (μ_0, π, ρ) with factor maps (ϕ, ψ).

(ii) $\phi \otimes \psi \otimes \phi$ is a measured morphism from $(\mathcal{X}\times\mathcal{G}\times\mathcal{X}, \mu_0\pi\rho)$ to $(\mathcal{X}'\times\mathcal{G}'\times\mathcal{X}', \mu_0'\pi'\rho')$.

Then $(i) \Rightarrow (ii)$, and if morever $\mathcal{X}'$, $\mathcal{G}'$ are standard Borel, then $(i) \Leftrightarrow (ii)$.

Proof.

$$
\begin{aligned}
&(\mu_0', \pi', \rho') \text{ factor of } (\mu_0, \pi, \rho) \\
\Leftrightarrow\quad &\begin{cases} \phi \cdot \mu_0 = \mu_0' \\ \mu_0(\psi \circ \pi) = \mu_0(\pi' \circ \phi) \\ \mu_0\pi\big(\phi \circ \rho\big) = \mu_0\pi\big(\rho' \circ (\phi \otimes \psi)\big) \end{cases} \\
\Rightarrow\quad &\begin{cases} \phi \cdot \mu_0 = \mu_0' \\ (\phi \otimes \psi) \cdot (\mu_0\pi) = (\phi \cdot \mu_0)\pi' \\ \mu_0\pi\big(\phi \circ \rho\big) = \mu_0\pi\big(\rho' \circ (\phi \otimes \psi)\big) \end{cases} && \text{(C.6.1)} \\
\Leftrightarrow\quad &\begin{cases} \phi \cdot \mu_0 = \mu_0' \\ (\phi \otimes \psi) \cdot (\mu_0\pi) = \mu_0'\pi' \\ \mu_0\pi\big(\phi \circ \rho\big) = \mu_0\pi\big(\rho' \circ (\phi \otimes \psi)\big) \end{cases} \\
\Rightarrow\quad &\begin{cases} \phi \cdot \mu_0 = \mu_0' \\ (\phi \otimes \psi) \cdot (\mu_0\pi) = \mu_0'\pi' \\ (\phi \otimes \psi \otimes \phi) \cdot (\mu_0\pi\rho) = \big((\phi \otimes \psi) \cdot \mu_0\pi\big)\rho' \end{cases} && \text{(C.6.2)} \\
\Leftrightarrow\quad &\begin{cases} \phi \cdot \mu_0 = \mu_0' \\ (\phi \otimes \psi) \cdot (\mu_0\pi) = \mu_0'\pi' \\ (\phi \otimes \psi \otimes \phi) \cdot (\mu_0\pi\rho) = \mu_0'\pi'\rho' \end{cases} \\
\Leftrightarrow\quad &(\phi \otimes \psi \otimes \phi) \cdot (\mu_0\pi\rho) = \mu_0'\pi'\rho', && \text{(C.6.3)}
\end{aligned}
$$

where lines (C.6.1) and (C.6.2) use point (*vi*) in Lemma C.3.1, and line (C.6.3) uses that the first two conditions are implied by the third condition by marginalisation. If moreover $\mathcal{X}'$ and $\mathcal{G}'$ are standard Borel, then using point (*vi*) in Lemma C.3.1 again, we obtain that the implications of lines (C.6.1) and (C.6.2) are equivalences. This ends the proof of Proposition 3.5.2. □

Proof of Proposition 3.5.4

Proposition 3.5.4. *The relation defined on stationary MDPs by MDP factors is a pre-order: ie., it is reflexive and transitive. The relation defined by MDP isomorphisms is an equivalence relation: i.e., it is reflexive, symmetric, and transitive. In particular, the factor and isomorphism relations on stationary Markov chains are, resp., a pre-order and an equivalence relation.*

Proof. Both for the factor and the isomorphism relation, the reflexiveness is straightforward (take $\phi := \mathtt{Id}_{\mathcal{X}}$ and $\psi := \mathtt{Id}_{\mathcal{G}}$). The idea is then, essentially, to combine (for the transivity) or invert (for the symmetry) the commutation relations from Definition 3.5.1. However, because of the dependency on the channels' input distribution for each commutation relation, we have to carefully justify each algebraic manipulation.

Let us first prove the transitivity. Assume that (μ_0', π', ρ') is a factor of (μ_0, π, ρ) with factor maps (ϕ, ψ), and (μ_0'', π'', ρ'') a factor of (μ_0', π', ρ') with factor maps (ϕ', ψ'). Explicitly: we have

(i) $\phi \cdot \mu_0 = \mu_0'$,

(ii) $\mu_0(\pi' \circ \phi) = \mu_0(\psi \circ \pi)$,

(iii) $\mu_0\pi(\rho' \circ (\phi \otimes \psi)) = \mu_0\pi(\phi \circ \rho)$.

and

(i)' $\phi' \cdot \mu_0' = \mu_0''$,

(ii)' $\mu_0'(\pi'' \circ \phi') = \mu_0'(\psi' \circ \pi')$,

(iii)' $\mu_0'\pi(\rho'' \circ (\phi' \otimes \psi')) = \mu_0'\pi(\phi' \circ \rho')$.

This implies:

(i)" Using point (ii) in Lemma C.3.1 and points (i), (i)' above:

$$(\phi' \circ \phi) \cdot \mu_0 = \phi' \cdot (\phi \cdot \mu_0) = \phi' \cdot (\mu_0') = \mu_0''.$$

(ii)" We have,

$$\begin{aligned}\mu_0(\pi'' \circ \phi' \circ \phi) &= \mu_0(\psi' \circ \pi' \circ \phi) \\ &= (\mathtt{Id}_{\mathcal{X}} \otimes \psi') \cdot \big(\mu_0(\pi' \circ \phi)\big) \\ &= (\mathtt{Id}_{\mathcal{X}} \otimes \psi') \cdot \big(\mu_0(\psi \circ \pi)\big) \\ &= \mu_0(\psi' \circ \psi \circ \pi).\end{aligned}$$

where we used, in the same order: point (ii)' above with point (v) in Lemma C.3.1; point (iv) in Lemma C.3.1; point (ii) above; and point (iv) in Lemma C.3.1 again.

(iii)" We have

$$\begin{aligned}\mu_0\pi\Big(\rho'' \circ (\phi' \otimes \psi') \circ (\phi \otimes \psi)\Big) &= \mu_0\pi\Big(\phi'' \circ \rho' \circ (\phi \otimes \psi)\Big) \\ &= \Big(\mathtt{Id}_{\mathcal{X}\times\mathcal{G}} \otimes \phi''\Big) \cdot \Big(\mu_0\pi\Big(\rho' \circ (\phi \otimes \psi)\Big)\Big) \\ &= \Big(\mathtt{Id}_{\mathcal{X}\times\mathcal{G}} \otimes \phi''\Big) \cdot \Big(\mu_0\pi\Big(\phi' \circ \rho\Big)\Big) \\ &= \mu_0\pi\Big(\phi'' \circ \phi' \circ \rho\Big),\end{aligned}$$

where we used, in the same order: point (iii)' above with point (v) in Lemma C.3.1; point (iv) in Lemma C.3.1; point (iii) above; and point (iv) in Lemma C.3.1 again.

This proves that (μ_0'', π'', ρ'') is a factor of (μ_0, π, ρ) with factor maps $(\phi' \circ \phi, \psi' \circ \psi)$. Thus the factor relation is transitive. Moreover, if in addition the factor maps ϕ, ϕ', ψ, ψ' are measured isomorphisms on their respective spaces, then from Proposition C.2.14, the compositions $\phi' \circ \phi$ and $\psi' \circ \psi$ are measured isomorphisms, resp., from $(\mathcal{X}, \mu_0)$ to $(\mathcal{X}'', \mu_0'')$ and from $(\mathcal{G}, \pi \cdot \mu_0)$ to $(\mathcal{G}'', \pi'' \cdot \mu_0'')$. This proves the transitivity of the isomorphism relation.

Let us now prove the symmetry of the isomorphism relation. Assume that $(\mathcal{X}', \mathcal{G}', \mu_0', \pi', \rho')$ is a factor of $(\mathcal{X}, \mathcal{G}, \mu_0, \pi, \rho)$ with factor maps (ϕ, ψ) such that ϕ, resp. ψ, is a measured isomorphism from $(\mathcal{X}, \mu_0)$ to $(\mathcal{X}', \mu_0')$, resp. from $(\mathcal{G}, \pi \cdot \mu_0)$ to $(\mathcal{G}', \pi \cdot \mu_0')$. Denote by ϕ^{-1} and ψ^{-1} mod 0 inverses of resp. ϕ and ψ (recall that these mod 0 inverses might not be set-theoretic inverses). First, $\phi \cdot \mu_0 = q_0'$ implies

$$\phi^{-1} \cdot q_0' = \phi^{-1} \cdot (\phi \cdot \mu_0) = (\phi^{-1} \circ \phi) \cdot \mu_0 = \mu_0$$

where the second equality uses point (ii) in Lemma C.3.1, and the last one uses point (vii)-(a) from the same lemma. Using now the lemma's point $(viii)$ with $f = \phi, g = \psi, q = \mu_0, \gamma = \pi$

and $\gamma' = \pi'$, we get

$$\mu_0(\pi' \circ \phi) = \mu_0(\psi \circ \pi) \quad \Leftrightarrow \quad \mu_0'(\psi^{-1} \circ \pi') = \mu_0'(\pi \circ \phi^{-1}),$$

Moreover, using, in this order, the same lemma's point (vii)-(c), point (v), and point (vii)-(a), we get

$$\begin{aligned}(\phi \otimes \psi) \cdot \mu_0\pi &= (\phi \cdot \mu_0)(\psi \circ \pi \circ \phi^{-1}) \\ &= (\phi \cdot \mu_0)(\pi' \circ \phi \circ \phi^{-1}) \\ &= \mu_0'\pi'.\end{aligned}$$

Using the latter and the lemma's point $(viii)$ with now $f = \phi \otimes \psi$, $g = \phi$, $q = \mu_0\pi$, $\gamma = \rho$ and $\gamma' = \rho'$ proves that

$$\mu_0\pi(\rho' \circ (\phi \otimes \psi)) = \mu_0\pi(\phi \circ \rho) \quad \Leftrightarrow \quad \mu_0'\pi'(\phi^{-1} \circ \rho') = \mu_0'\pi'(\rho \circ (\phi^{-1} \otimes \psi^{-1})).$$

Thus the stationary MDP isomorphism relation is an equivalence relation. Eventually, these results directly imply that stationary Markov chain factors, resp. isomorphisms, define a preorder, resp. an equivalence relation: indeed, stationary Markov chains can be seen as stationary MDPs with only one action. This ends the proof of Proposition 3.5.4.

□

C.6.2 Appendix for Section 3.5.2

Proof of Proposition 3.5.7

Proposition 3.5.7. *Let $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$ be a canonical joining of stationary standard Borel MDPs $(\mu_0^c, \pi^c, \rho^c)_{c \in C}$. Define the tensor products (see Definition 3.2.6)*

$$\begin{aligned}\boldsymbol{\pi}^{\otimes} &:= \bigotimes_{c \in C} \pi^c \in \mathscr{K}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{G}}), \\ \boldsymbol{\rho}^{\otimes} &:= \bigotimes_{c \in C} \rho^c \in \mathscr{K}(\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{X}}).\end{aligned} \tag{3.5.5}$$

Then:

- *If π^c is deterministic for all $c \in C$, then $\boldsymbol{\pi} = \boldsymbol{\pi}^{\otimes}$ holds $\boldsymbol{\mu}_0$-a.e.; but for a general family $(\pi^c)_{c \in C}$, the latter equality might not hold.*
- *If ρ^c is deterministic for all $c \in C$, then $\boldsymbol{\rho} = \boldsymbol{\rho}^{\otimes}$ holds $\boldsymbol{\mu}_0\boldsymbol{\pi}$-a.e.; but for a general family $(\rho^c)_{c \in C}$, the latter equality might not hold.*

Proof. Consider two families of standard Borel spaces $(\mathcal{A}^c)_{c \in C}$, $(\mathcal{B}^c)_{c \in C}$, with corresponding product spaces $\boldsymbol{\mathcal{A}}$, resp. $\boldsymbol{\mathcal{B}}$, together with for each $c \in C$, a measure $\mu^c \in \Delta_{\mathcal{A}}$ and a channel $\gamma^c \in \mathscr{K}(\mathcal{A}^c, \mathcal{B}^c)$. Let $\boldsymbol{\mu} \in \Delta_{\boldsymbol{\mathcal{A}}}$ and $\boldsymbol{\gamma} \in \mathscr{K}(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}})$ such that for all $c \in C$

$$\boldsymbol{\mu}(\gamma^c \circ \mathsf{pr}_{\boldsymbol{\mathcal{A}}}^c) = \boldsymbol{\mu}(\mathsf{pr}_{\boldsymbol{\mathcal{B}}}^c \circ \boldsymbol{\gamma}), \tag{C.6.4}$$

where $\mathsf{pr}_{\boldsymbol{\mathcal{A}}}^c$ and $\mathsf{pr}_{\boldsymbol{\mathcal{B}}}^c$ are the projection on the coordinate $\mathcal{A}^c$ in $\boldsymbol{\mathcal{A}}$, resp. on the coordinate $\mathcal{B}^c$ in $\boldsymbol{\mathcal{B}}$. Eventually, define

$$\boldsymbol{\gamma}^{\otimes} := \bigotimes_{c \in C} \gamma^c \in \mathscr{K}(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}),$$

To prove the result, it is enough to prove that

(i) If γ^c is deterministic for all $c \in \mathcal{C}$, then $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{\otimes}$ holds $\boldsymbol{\mu}$-a.e.,

(ii) In general, there exists $\boldsymbol{\gamma}$ satisfying the above conditions such that the latter equality does not hold.

Note first that as here the spaces are standard Borel, from point (i) in Lemma B.1.2, equation (C.6.4) is equivalent to

$$\gamma^c \circ \mathsf{pr}^c_{\mathcal{A}} = \mathsf{pr}^c_{\mathcal{B}} \circ \boldsymbol{\gamma} \quad \boldsymbol{\mu}\text{-a.e.} \tag{C.6.5}$$

To prove (i), assume that for all $c \in \mathcal{C}$, the channel γ^c is deterministic, and write $f^c : \mathcal{A} \to \mathcal{B}^c$ the function defined by the deterministic channel $\gamma^c \circ \mathsf{pr}^c_{\mathcal{A}}$. This implies, using equation (C.6.5), the existence of a set $\tilde{\mathcal{A}}$ such that $\boldsymbol{\mu}(\tilde{\mathcal{A}}) = 1$ and for all $\boldsymbol{a} \in \tilde{\mathcal{A}}$,

$$\delta_{f^c(\boldsymbol{a})} = (\gamma^c \circ \mathsf{pr}^c_{\mathcal{A}})(\cdot|\boldsymbol{a}) = (\mathsf{pr}^c_{\mathcal{B}} \circ \boldsymbol{\gamma})(\cdot|\boldsymbol{a}).$$

It is easy to verify that this implies,

$$\forall \boldsymbol{a} \in \tilde{\mathcal{A}}, \quad \delta_{\boldsymbol{f}(\boldsymbol{a})} = \boldsymbol{\gamma}(\cdot|\boldsymbol{a}), \tag{C.6.6}$$

where we defined the measurable function $\boldsymbol{f} : \boldsymbol{\mathcal{A}} \to \boldsymbol{\mathcal{B}}$ by $\boldsymbol{f}(\boldsymbol{a}) := (f^c(\boldsymbol{a}))_{c \in \mathcal{C}}$. Now, equation (C.6.5) and $\boldsymbol{\mu}(\tilde{\mathcal{A}}) = 1$ clearly implies

$$\boldsymbol{\gamma}_{\boldsymbol{f}} = \boldsymbol{\tau} \quad \boldsymbol{\mu}\text{-a.e.}$$

where $\boldsymbol{\gamma}_{\boldsymbol{f}}$ is the deterministic channel defined by $\boldsymbol{f}$. But by construction, we actually have $\boldsymbol{\gamma}_{\boldsymbol{f}} = \boldsymbol{\gamma}^{\otimes}$.

Let us now prove point (ii) above by exhibiting a counter-example. Choose $\mathcal{C} = \{1, 2\}$ and for all $c \in \{1, 2\}$, the spaces $\mathcal{A}^c = \mathcal{B}^c = \{a^c_1, a^c_2\}$, and the binary symmetric channel

$$\gamma^c(a^c_{i'}|a^c_i) = \begin{cases} 1 - r & \text{if } i' = i, \\ r & \text{if } i' \neq i, \end{cases}$$

where $0 < r < 1$ does not depend on $c \in \mathcal{C}$. Consider the distribution $\boldsymbol{\mu} \in \Delta_{\{a^1_1, a^1_2\} \times \{a^2_1, a^2_2\}}$ defined by $\boldsymbol{\mu}(a^1_1) = \boldsymbol{\mu}(a^2_2) = \frac{1}{2}$ and $\boldsymbol{\mu}(a^1_2) = \boldsymbol{\mu}(a^2_1) = 0$. Note that the support of $\mathrm{supp}(\boldsymbol{\mu})$ is a two-elements "diagonal" in the four-elements "square" $\{a^1_1, a^1_2\} \times \{a^2_1, a^2_2\}$: the idea, here, is to join the binary symmetric channels γ^1 and γ^2 into a binary symmetric channel on this diagonal. It does not matter how we define $\boldsymbol{\gamma}$ on the other, null probability diagonal; here we choose it to implement there another binary symmetric channel. I.e., formally:

$$\boldsymbol{\gamma}((a^1_{i'}, a^2_{j'})|(a^1_i, a^2_j)) = \begin{cases} 1 - r & \text{if } (i', j') = (i', j'), \\ r & \text{if } i \neq i' \text{ and } j \neq j', \\ 0 & \text{otherwise.} \end{cases}$$

Then the commutation relations from equation (C.6.5) are easily verified. But, on the other hand, here the "split" channel $\boldsymbol{\gamma}^{\otimes} = \gamma^1 \otimes \gamma^2$, starting from a given diagonal, "spills over" on the other diagonal. More precisely, it is defined by

$$\boldsymbol{\gamma}^{\otimes}((a^1_{i'}, a^2_{j'})|(a^1_i, a^2_j)) = \begin{cases} (1 - r)^2 & \text{if } (i', j') = (i', j'), \\ r^2 & \text{if } i \neq i' \text{ and } j \neq j', \\ r(1 - r) & \text{otherwise,} \end{cases}$$

which proves that $\boldsymbol{\mu\gamma} \neq \boldsymbol{\mu\gamma}^{\otimes}$. This ends the proof of Proposition 3.5.7. □

Proof of Proposition 3.5.11

Proposition 3.5.11. *The relation defined on the joinings of a given family of stationary MDPs* $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ *by j-factors is a pre-order: i.e., it is reflexive and transitive.*

Proof. The reflexivity of the j-factor relation is trivial: take Φ and Ψ equal to identity maps. Moreover, Proposition 3.5.4 proves that the MDP factor relation is a pre-order. Thus we only need to prove that the relation defined by equation (3.5.6) is transitive. The reasoning is similar in spirit to that of the proof of Proposition 3.5.4.

Let (ν_0, η, ξ), (ν_0', η', ξ') and (ν_0'', η'', ξ'') be three joinings of the same family of stationary MDPs $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$, with resp. marginalisation maps $(\phi^c, \psi^c)_{c\in C}$, $((\phi')^c, (\psi')^c)_{c\in C}$, and $((\phi'')^c, (\psi'')^c)_{c\in C}$. Assume that (ν_0', η', ξ') is a j-factor of (ν_0, η, ξ) with factor maps (Φ, Ψ), and (ν_0'', η'', ξ'') a j-factor of (ν_0', η', ξ') with factor maps (Φ', Ψ'). In particular,

$$(\nu_0\eta)\Big(\Big((\phi')^c \otimes (\psi')^c\Big)\circ\Big(\Phi \otimes \Psi\Big)\Big) = (\nu_0\eta)(\phi^c \otimes \psi^c) \tag{C.6.7}$$

and

$$(\nu_0'\eta')\Big(\Big((\phi'')^c \otimes (\psi'')^c\Big)\circ\Big(\Phi' \otimes \Psi'\Big)\Big) = (\nu_0'\eta')((\phi')^c \otimes (\psi')^c) \tag{C.6.8}$$

Moreover, from Proposition 3.5.2, we have

$$(\Phi \otimes \Psi \otimes \Phi) \cdot \nu_0\eta\rho = \nu_0'\eta'\rho',$$

which yields, by marginalisation,

$$(\Phi \otimes \Psi) \cdot \nu_0\eta = \nu_0'\eta'. \tag{C.6.9}$$

Combining equations (C.6.8) and (C.6.9), we have

$$\Big((\Phi \otimes \Psi) \cdot \nu_0\eta\Big)\Big(\Big((\phi'')^c \otimes (\psi'')^c\Big)\circ\Big(\Phi' \otimes \Psi'\Big)\Big) = \Big((\Phi \otimes \Psi) \cdot \nu_0\eta\Big)((\phi')^c \otimes (\psi')^c),$$

which from point (υ) in Lemma C.3.1 is equivalent to

$$(\nu_0\eta)\Big[\Big((\phi'')^c \otimes (\psi'')^c\Big)\circ(\Phi' \otimes \Psi')\circ(\Phi \otimes \Psi)\Big] = (\nu_0\eta)\Big[\Big((\phi')^c \otimes (\psi')^c\Big)\circ(\Phi \otimes \Psi)\Big] \tag{C.6.10}$$

We can now compute

$$\begin{aligned}
&(\nu_0\eta)\Big[\Big((\phi'')^c \otimes (\psi'')^c\Big)\circ\Big((\Phi'\circ\Phi) \otimes (\Psi\circ\Psi)\Big)\Big] \\
&= (\nu_0\eta)\Big[\Big((\phi'')^c \otimes (\psi'')^c\Big)\circ(\Phi' \otimes \Psi')\circ(\Phi \otimes \Psi)\Big] \\
&= (\nu_0\eta)\Big[\Big((\phi')^c \otimes (\psi')^c\Big)\circ(\Phi \otimes \Psi)\Big] && \text{(C.6.11)} \\
&= (\nu_0\eta)(\phi^c \otimes \psi^c), && \text{(C.6.12)}
\end{aligned}$$

where line (C.6.11), uses equation (C.6.10); and line (C.6.12) uses equation (C.6.7). This ends the proof of Proposition 3.5.11. □

Proof of Theorem 3.5.13

Theorem 3.5.13. *For a finite family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ of finite-alphabet, stationary MDPs, there always exists at least one minimal joining.*

The proof will use the following lemmas.

Lemma C.6.1. *Let (ν_0, η, ξ) and (ν_0', η', ξ') be two joinings of $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$, with resp. state-spaces $\mathcal{P}$, $\mathcal{P}'$ and resp. action spaces $\mathcal{K}$, $\mathcal{K}'$. Assume that (ν_0', η', ξ') is a j-factor of (ν_0, η, ξ) with factor maps (Φ, Ψ). If $\Phi \otimes \Psi$ is an isomorphism from $(\mathcal{P} \times \mathcal{K}, \nu_0\eta)$ to $(\mathcal{P}' \times \mathcal{K}', \nu_0'\eta')$, then (ν_0, η, ξ) is a j-factor of (ν_0', η', ξ').*

Proof. Assume that $\Phi \otimes \Psi$ is an isomorphism from $(\mathcal{P} \times \mathcal{K}, \nu_0\eta)$ to $(\mathcal{P}' \times \mathcal{K}', \nu_0'\eta')$. From Lemma C.3.2, the map Φ is an isomorphism from $(\mathcal{P}, \nu_0)$ to $(\mathcal{P}', \nu_0')$, and Ψ is an isomorphism from $(\mathcal{K}, \eta \cdot \nu_0)$ to $(\mathcal{K}', \eta' \cdot \nu_0')$. As by assumption, (ν_0', η', ξ') is a j-factor and in particular an MDP factor of (ν_0, η, ξ) with factor maps (Φ, Ψ), this implies that (ν_0', η', ξ') is MDP isomorphic to (ν_0, η, ξ) (see Definition 3.5.1). By symmetry of the isomorphism relation (see Proposition 3.5.4), this implies that (ν_0, η, ξ) is MDP isomorphic to (ν_0', η', ξ'), with factor maps (Φ^{-1}, Ψ^{-1}), where Φ^{-1} and Ψ^{-1} are mod 0 inverses of resp. Φ and Ψ. In particular, (ν_0, η, ξ) is an MDP factor of (ν_0', η', ξ') with factor maps (Φ^{-1}, Ψ^{-1}). Therefore, from the Definition 3.5.10 of j-factor, the only remaining thing to prove is the equality

$$(\nu_0'\eta')\Big(\big(\phi^c \otimes \psi^c\big)\circ\big(\Phi^{-1} \otimes \Psi^{-1}\big)\Big) = (\nu_0'\eta')\big((\phi')^c \otimes (\psi')^c\big). \tag{C.6.13}$$

where the maps $(\phi^c, \psi^c)_{c\in C}$, resp. $((\phi')^c, (\psi')^c)_{c\in C}$, are the marginalisation maps of the joining (ν_0, η, ξ), resp. of the joining (ν_0', η', ξ'). But from the assumption that $\Phi \otimes \Psi$ is an isomorphism, we have

$$\nu_0'\eta' = (\Phi \otimes \Psi) \cdot \nu_0\eta, \tag{C.6.14}$$

while from the assumption that (ν_0', η', ξ') is a j-factor of (ν_0, η, ξ), we have

$$(\nu_0\eta)\Big(\big((\phi')^c \otimes (\psi')^c\big)\circ\big(\Phi \otimes \Psi\big)\Big) = (\nu_0\eta)(\phi^c \otimes \psi^c). \tag{C.6.15}$$

Equation (C.6.13) is then a consequence of equation (C.6.14) and equation (C.6.15), combined with point (vii)-(d) in Lemma C.3.1. □

Lemma C.6.2. *Let $\mathcal{A}$, $\mathcal{B}$ be finite sets, $\mu \in \Delta_{\mathcal{A}}$, $\mu' \in \Delta_{\mathcal{B}}$, and $f : \mathcal{A} \to \mathcal{B}$ a measured morphism from $(\mathcal{A}, \mu)$ to $(\mathcal{B}, \mu')$ — i.e., $f \cdot \mu = \mu'$. Then:*[6]

(i) The restriction of f to $\mathrm{supp}(\mu)$ *is a surjective function to* $\mathrm{supp}(\mu')$.

(ii) f is also a measured isomorphism from $(\mathcal{A}, \mu)$ to $(\mathcal{B}, \mu')$ if and only if it induces a bijection between $\mathrm{supp}(\mu)$ *and* $\mathrm{supp}(\mu')$.

In particular, if the measured morphism f is not an isomorphism, then $|\mathrm{supp}(\mu')| < |\mathrm{supp}(\mu)|$.

Proof. This is a direct verification. □

[6] See Chapter 2, equation (2.1.1) for the definition of the support $\mathrm{supp}(\mu)$ of a finite distribution μ.

We are now ready to prove Theorem 3.5.13.

Proof of Theorem 3.5.13. We will use Zorn's lemma, which applies to partial orders, i.e., reflexive, transitive and anti-symmetric binary relations. From Proposition 3.5.11, the j-factor relation $\lesssim$ on $\mathsf{Join} := \mathsf{Join}((\mu_0^c, \pi^c, \rho^c)_{c\in C})$ is a pre-order, i.e., it is reflexive and transitive. Yet, it is not necessarily anti-symmetric: i.e., two joinings that are factor of one another are not necessarily equal. However, we can "make the relation anti-symmetric" by quotienting: i.e., we define the equivalence relation $\sim$ on Join by

$$(\nu_0, \eta, \xi) \sim (\nu_0', \eta', \xi') \quad \Leftrightarrow \quad (\nu_0, \eta, \xi) \lesssim (\nu_0', \eta', \xi') \text{ and } (\nu_0', \eta', \xi') \lesssim (\nu_0, \eta, \xi),$$

and we consider the set $[\mathsf{Join}]$ of equivalence classes w.r.t. this equivalence relation, where we denote by $[\mathsf{j}] \in [\mathsf{Join}]$ the class containing a joining $\mathsf{j} \in \mathsf{Join}$. The relation $\lesssim$ induces a relation on $[\mathsf{Join}]$, still denoted by $\lesssim$, and defined by

$$[\mathsf{j}] \lesssim [\mathsf{j}'] \quad \Leftrightarrow \quad \forall (\nu_0, \eta, \xi) \in [\mathsf{j}],\ \forall (\nu_0', \eta', \xi') \in [\mathsf{j}'],\ (\nu_0, \eta, \xi) \lesssim (\nu_0', \eta', \xi'). \tag{C.6.16}$$

This relation on $[\mathsf{Join}]$ not only inherits the reflexiveness and transitivity from the pre-order on Join, but is now also anti-symmetric, i.e., it is a partial order. We want to prove that this partial order has a minimal element, i.e., that

$$\exists\, [\mathsf{j}^*] \in [\mathsf{Join}] \ :\ \forall [\mathsf{j}],\ [\mathsf{j}^*] \lesssim [\mathsf{j}]. \tag{C.6.17}$$

From (C.6.16),, this will imply that any fixed element $(\nu_0^*, \eta^*, \xi^*) \in [\mathsf{j}^*]$ is a minimal joining. Now, Zorn's lemma states, in our context, that to prove (C.6.17), it is enough to prove that (i) the set $[\mathsf{Join}]$ is not empty, and (ii) any chain w.r.t. the partial order $\lesssim$ in $[\mathsf{Join}]$ has a lower bound.[7]

Point (i) is straightforward, as Join always contains the product joining (see Definition 3.5.6), and therefore $[\mathsf{Join}]$ contains the latter's class. To prove (ii), let us consider a chain Ch w.r.t. $\lesssim$. Fix an element $[\mathsf{j}^0] \in \mathsf{Ch}$. Let $[\mathsf{j}^1] \in \mathsf{Ch}$ such that $[\mathsf{j}^1] \lesssim [\mathsf{j}^0]$ with $[\mathsf{j}^1] \neq [\mathsf{j}^0]$, and $(\nu_0^0, \eta^0, \xi^0) \in [\mathsf{j}^0]$, $(\nu_0^1, \eta^1, \xi^1) \in [\mathsf{j}^1]$. The joining (ν_0^1, η^1, ξ^1) is a j-factor of the joining (ν_0^0, η^0, ξ^0); let $f : (\mathcal{P}^0, \nu_0^0) \to (\mathcal{P}^1, \nu_0^1)$ and $g : (\mathcal{K}^0, \eta^0 \cdot \nu_0^0) \to (\mathcal{K}^1, \eta^1 \cdot \nu_0^1)$ the corresponding factor maps. From Proposition 3.5.2, the map $f \otimes g$ is a measured morphism from $(\mathcal{P}^0 \times \mathcal{K}^0, \nu_0^0\eta^0)$ to $(\mathcal{P}^1 \times \mathcal{K}^1, \nu_0^1\eta^1)$. Moreover, as (ν_0^0, η^0, ξ^0) is not a j-factor of (ν_0^0, η^0, ξ^0), from Lemma C.6.1, the measured morphism $f \otimes g$ is not a measured isomorphism. Thus, from Lemma C.6.2, we have $|\mathrm{supp}(\nu_0^1\pi^1)| < |\mathrm{supp}(\nu_0^0\pi^0)|$. By iterating the argument, we obtain that for any finite sequence $([\mathsf{j}_i])_{1\leq i\leq n}$, included in the chain Ch, such that $[\mathsf{j}_{i+1}] \lesssim [\mathsf{j}_i]$ and $[\mathsf{j}_{i+1}] \neq [\mathsf{j}_i]$ for all $1 \leq i \leq n$, the sequence $(|\mathrm{supp}(\nu_0^i\eta^i)|)_{1\leq i\leq n}$ strictly decreases as i increases. As this sequence is bounded from above by $|\mathrm{supp}(\nu_0^0\eta^0)| < \infty$ and from below by 1, this implies that there exists $N \in \mathbb{N}$, that depends only on the first joining (ν_0^0, η^0, ξ^0) but not on the remaining joinings $(\nu_0^i, \eta^i, \xi^i)_{1\leq i\leq n}$, such that $n \leq N$. As, by definition, the chain Ch is totally ordered, this implies that there is only a finite number of distinct $[\mathsf{j}] \in \mathsf{Ch}$ such that $[\mathsf{j}] \leq [\mathsf{j}_0]$. Thus the chain Ch has a minimal element, which is in particular a lower bound. This proves property (ii) above, which ends the proof of Theorem 3.5.13. □

Note that in the proof above, the only point where we used the assumption that there are a finite number of finite-alphabet MDPs is to prove that each chain w.r.t. $\lesssim$ has a lower bound. Thus proving this fact for the non-finite case would be enough to generalise Theorem 3.5.13 to the non-finite case. Adaptation of arguments on *minimal σ-algebras* from (Ay et al., 2015) and *invariant σ-algebras* from (Pfante et al., 2015) might be relevant for that purpose.

[7] A chain is a totally ordered subset, i.e., here, a subset $\mathsf{Ch} \subseteq [\mathsf{Join}]$ s.t. for all $[\mathsf{j}], [\mathsf{j}'] \in \mathsf{Ch}$, we have either $[\mathsf{j}] \lesssim [\mathsf{j}']$ or $[\mathsf{j}'] \lesssim [\mathsf{j}]$. A lower bound of Ch is an element $[\mathsf{j}_{\mathsf{Ch}}] \in [\mathsf{Join}]$ s.t. $[\mathsf{j}_{\mathsf{Ch}}] \lesssim [\mathsf{j}]$ for all $[\mathsf{j}] \in \mathsf{Ch}$.

Proof of Proposition 3.5.14

Proposition 3.5.14. *Let $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ be a countable family of standard Borel stationary MDPs. Then for any joining (ν_0, η, ξ) of $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$, there exists a canonical joining of $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ which is a j-factor of (ν_0, η, ξ). In particular, the family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ has a minimal joining if and only if it has a canonical minimal joining.*

Proof. Let (ν_0, η, ξ) be a joining with state-space $(\mathcal{P}, \mathfrak{P})$ and action space $(\mathcal{K}, \mathfrak{K})$. From Proposition 3.5.2, for all $c \in C$, the marginalisation maps $\phi^c : \mathcal{P} \to \mathcal{X}^c$ and $\psi^c : \mathcal{K} \to \mathcal{G}^c$ are such that

$$(\phi^c \otimes \psi^c \otimes \phi^c) \cdot \nu_0\eta\xi = \mu_0^c\pi^c\rho^c.$$

Denote by $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{G}}$ the product measurable spaces of resp. $(\mathcal{X}^c)_{c\in C}$ and $(\mathcal{G}^c)_{c\in C}$. Define the measurable functions

$$\begin{aligned}\boldsymbol{\phi} : \ & \mathcal{P} \to \boldsymbol{\mathcal{X}} \\ & p \mapsto (\phi^c(p))_{c\in C}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\psi} : \ & \mathcal{K} \to \boldsymbol{\mathcal{G}} \\ & k \mapsto (\psi^c(k))_{c\in C}\end{aligned}$$

together with the distributtion

$$\boldsymbol{q}_0^1 := (\boldsymbol{\phi} \otimes \boldsymbol{\psi} \otimes \boldsymbol{\phi}) \cdot \nu_0\eta\xi \in \Delta_{\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{G}}\times\boldsymbol{\mathcal{X}}}. \tag{C.6.18}$$

As a countable Cartesian product of standard Borel spaces is standard Borel (see Proposition C.2.5), and as here C is countable, the spaces $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{G}}$, and $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{G}} \times \boldsymbol{\mathcal{X}}$ are standard Borel. Thus, using (twice) Proposition C.2.16, we obtain that there exist channels $\boldsymbol{\pi} \in \mathscr{K}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{G}})$ and $\boldsymbol{\rho} \in \mathscr{K}(\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{X}})$ such that $\boldsymbol{q}_0^1 = \boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho}$, where $\boldsymbol{\mu}_0$ the marginal of $\boldsymbol{q}_0^1 \in \Delta_{\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{G}}\times\boldsymbol{\mathcal{X}}}$ on the first state-space coordinate $\boldsymbol{\mathcal{X}}$. This yields a measured MDP $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$ which, from equation (C.6.18) and the stationarity of (ν_0, η, ξ), is stationary as well. Moreover, from equation (C.6.18) and Proposition 3.5.2, the stationary MDP $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$ is an MDP factor of (ν_0, η, ξ).

Now, denoting by $\mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c$, resp. $\mathsf{pr}_{\boldsymbol{\mathcal{G}}}^c$ the projection on the coordinate c in $\boldsymbol{\mathcal{X}}$, resp. $\boldsymbol{\mathcal{G}}$, we have, for all $c \in C$,

$$\begin{aligned}(\mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c \otimes \mathsf{pr}_{\boldsymbol{\mathcal{G}}}^c \otimes \mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c) \cdot \boldsymbol{\mu}_0\boldsymbol{\pi}\boldsymbol{\rho} &= (\mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c \otimes \mathsf{pr}_{\boldsymbol{\mathcal{G}}}^c \otimes \mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c) \cdot ((\boldsymbol{\phi} \otimes \boldsymbol{\psi} \otimes \boldsymbol{\phi}) \cdot \nu_0\eta\xi) \\ &= ((\mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c \otimes \mathsf{pr}_{\boldsymbol{\mathcal{G}}}^c \otimes \mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c) \circ (\boldsymbol{\phi} \otimes \boldsymbol{\psi} \otimes \boldsymbol{\phi})) \cdot \nu_0\eta\xi \\ &= ((\mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c \circ \boldsymbol{\phi}) \otimes (\mathsf{pr}_{\boldsymbol{\mathcal{G}}}^c \circ \boldsymbol{\psi}) \otimes (\mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c \circ \boldsymbol{\phi})) \cdot \nu_0\eta\xi \\ &= (\phi^c \otimes \psi^c \otimes \phi^c) \cdot \nu_0\eta\xi \\ &= \mu_0^c\pi^c\rho^c.\end{aligned}$$

Therefore, for all $c \in C$, from Proposition 3.5.2, the standard Borel stationary MDP (μ_0^c, π^c, ρ^c) is an MDP factor of the stationary MDP $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$ with factor maps $(\mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c, \mathsf{pr}_{\boldsymbol{\mathcal{G}}}^c)$: i.e., $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$ is a canonical joining of the family $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ (see Definition 3.5.6). Moreover, as we already saw that $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$ is an MDP factor of (ν_0, η, ξ) with factor maps (Φ, Ψ), the equalities

$$\begin{aligned}\mathsf{pr}_{\boldsymbol{\mathcal{X}}}^c \circ \boldsymbol{\phi} &= \phi^c \\ \mathsf{pr}_{\boldsymbol{\mathcal{G}}}^c \circ \boldsymbol{\psi} &= \psi^c\end{aligned}$$

show that $(\boldsymbol{\mu}_0, \boldsymbol{\pi}, \boldsymbol{\rho})$ is a j-factor of (ν_0, η, ξ), which ends the proof of Proposition 3.5.14. □

C.6.3 Proof of Theorem 3.5.17

Theorem 3.5.17. *Assume that there is a countable number of orbits. Then the following are equivalent:*

(i) The group action has a measured class-pose decomposition (κ, θ, ξ) with state-space $\mathcal{P}$.

(ii) The MDP (π_v, ρ) defined by the group action ρ has an isomorphic class-pose parametrisation whose isomorphic minimal joining is of the form (ν_0, η_v, ξ), with state-space $\mathcal{P}$ and action space $\mathcal{K} := \mathcal{G}$, where $\eta_v \in \mathscr{K}(\mathcal{P}, \mathcal{G})$ is the independent policy, and with marginalisation maps (ϕ, ψ) such that for all $c \in \mathcal{C}$, the map $\psi^c : \mathcal{K} = \mathcal{G} \to \mathcal{G}$ is trivial: i.e., it is the identity map $\mathtt{Id}_{\mathcal{G}}$.

Moreover, if any of the above holds, then the class-pose decomposition from (i) and the isometric class-pose parametrisation from (ii) can be chosen such that:

- *$\mathcal{P}$ is the same space in points (i) and (ii),*
- *ξ is the same transition channel in points (i) and (ii),*
- *For all $c \in \mathcal{C}$, the restriction θ^c of θ of point (i) — which is a measured isomorphism from $(\mathcal{X}^c, \epsilon^c)$ to $(\mathcal{P}, \nu_0)$ — and the restriction ϕ^c of ϕ of point (ii) — which is a measured isomorphism from $(\mathcal{P}, \nu_0)$ to $(\mathcal{X}^c, \epsilon^c)$ — are mod 0 inverses of each other.*

Proof. $(i) \Rightarrow (ii)$. Let $c \in \mathcal{C}$. Define $\phi^c : \mathcal{P} \to \mathcal{X}^c$ as a mod 0 inverse of the measured isomorphism θ^c from $(\mathcal{X}^c, \epsilon^c)$ to $(\mathcal{P}, \nu_0)$. We clearly have

$$\phi^c \cdot \nu_0 = \epsilon^c, \tag{C.6.19}$$

and for all $g \in \tilde{\mathcal{G}}$, the relation

$$\epsilon^c(\rho_g \circ \theta^c) = \epsilon^c(\theta^c \circ \xi_g)$$

implies, from point $(viii)$ in Lemma C.3.1, that

$$\nu_0(\xi_g \circ \phi^c) = \nu_0(\phi^c \circ \rho_g). \tag{C.6.20}$$

Define now $\phi : \mathcal{C} \times \mathcal{P} \to \mathcal{X}$ by $\phi(c, p) := \phi^c(p)$. For any $F \in \boldsymbol{\mathcal{X}}$ we have

$$\begin{aligned}
\phi^{-1}(F) &= \phi^{-1}\Big(\bigsqcup_{c \in \mathcal{C}} F \cap \mathcal{X}^c\Big) \\
&= \bigsqcup_{c \in \mathcal{C}} \phi^{-1}(F \cap \mathcal{X}^c) \\
&= \bigsqcup_{c \in \mathcal{C}} \{c\} \times (\phi^c)^{-1}(F \cap \mathcal{X}^c).
\end{aligned}$$

As each ϕ^c is measurable and $\mathcal{C}$ is here assumed countable, each $\phi^{-1}(F)$ is a countable union of measurable sets, and therefore measurable. Thus ϕ is measurable. We then define $\mathcal{K} := \mathcal{G}$, and the map $\psi : \mathcal{C} \times \mathcal{G} \to \mathcal{G}$ by $\psi(c, g) := \psi^c(g) := g$, which is clearly measurable. It can then be easily verified that for all $c \in \mathcal{C}$, equation (C.6.20) fo all $g \in \tilde{\mathcal{G}}$, with $\upsilon(\tilde{\mathcal{G}}) = 1$, implies

$$(\nu_0 \otimes \pi_v)(\xi \circ (\phi^c \otimes \psi^c)) = (\nu_0 \otimes \pi_v)(\phi^c \circ \rho),$$

i.e.,

$$\nu_0 \pi_\upsilon(\xi \circ (\phi^c \otimes \psi^c)) = \nu_0 \pi_\upsilon(\phi^c \circ \rho). \tag{C.6.21}$$

Let us now denote by $\eta_\upsilon \in \mathscr{K}(\mathcal{P}, \mathcal{G})$ the independent policy on $\mathcal{P}$. We have

$$\nu_0(\pi_\upsilon \circ \phi^c) = \nu_0 \otimes \upsilon = \nu_0 \eta_\upsilon = \nu_0(\mathtt{Id}_\mathcal{G} \circ \eta_\upsilon) = \nu_0(\psi \circ \eta_\upsilon). \tag{C.6.22}$$

Combining equations (C.6.19), (C.6.21) and (C.6.22), we obtain that $(\nu_0, \eta_\upsilon, \xi)$ is a joining of the family $(\epsilon^c, \pi_\upsilon^c, \rho^c)_{c \in \mathcal{C}}$, with factor maps ϕ and ψ. Moreover, by assumption, for all $c \in \mathcal{C}$ the map ϕ^c is an isomorphism from $(\mathcal{P}, \nu_0)$ to $(\mathcal{X}^c, \epsilon^c)$, and $\psi^c = \mathtt{Id}_\mathcal{G}$ is clearly an isomorphism from $(\mathcal{G}, \upsilon)$ to itself. Thus we have an isomorphic joining of the ergodic components $(\epsilon^c, \pi_\upsilon^c, \rho^c)$ of the measurable MDP (π_υ, ρ), with marginalisation maps ϕ^c and ψ^c such that ϕ and ψ are measurable.

$(ii) \Rightarrow (i)$. Let us now start from an isomorphic joining $(\nu_0, \eta_\upsilon, \xi)$ of the ergodic components $(\epsilon^c, \pi_\upsilon^c, \rho^c)_{c \in \mathcal{C}}$, with action space $\mathcal{K} := \mathcal{G}$, action space factor map ψ such that $\psi^c = \mathtt{Id}_\mathcal{G}$ for all $c \in \mathcal{C}$, and where η_υ is the independent policy. Note that from Proposition 3.5.9, the pose's state-space $\mathcal{P}$ can be chosen standard Borel, which we do.

Let $c \in \mathcal{C}$. As $(\nu_0, \eta_\upsilon, \xi)$ is a factor of $(\epsilon^c, \pi_\upsilon^c, \rho^c)$ with factor maps ϕ, ψ, we have (see Definition 3.5.1)

$$\phi^c \cdot \nu_0 = \epsilon^c, \tag{C.6.23}$$

and, noting that $\nu_0 \eta_\upsilon = \nu_0 \otimes \upsilon$ and $\psi^c = \mathtt{Id}_\mathcal{G}$ for all $c \in \mathcal{C}$,

$$(\nu_0 \otimes \upsilon)(\rho^c \circ (\phi^c \otimes \mathtt{Id}_\mathcal{G})) = (\nu_0 \otimes \upsilon)(\phi^c \circ \xi). \tag{C.6.24}$$

The space $(\mathcal{X}^c, \mathfrak{X}^c)$ is standard Borel, as well as $(\mathcal{P}, \mathfrak{P})$. Thus, from Proposition C.2.3, there exist countable families of measurable subsets $\{F_n^\mathcal{P}\}_{n \in \mathbb{N}} \subseteq \mathfrak{P}$, resp. $\{F_n^c\}_{n \in \mathbb{N}} \subseteq \mathfrak{X}^c$, such that $\{F_n^\mathcal{P}\}_{n \in \mathbb{N}}$ generates the σ-algebra $\mathfrak{P}$, resp. $\{F_n^c\}_{n \in \mathbb{N}}$ generates the σ-algebra $\mathfrak{X}^c$. Therefore the family $\{F_m^\mathcal{P} \times F_n^c\}_{m,n \in \mathbb{N}}$ generates $\mathfrak{P} \otimes \mathfrak{X}^c$ (it generates the rectangles, and thus the whole product σ-algebra). Let us now fix $m, n \in \mathbb{N}$. Define the measurable functions h, h' on $\mathcal{G}$ by

$$h(g) := \left[\nu_0(\phi^c \circ \xi_g)\right](F_m^\mathcal{P} \times F_n^c),$$
$$h'(g) := \left[\nu_0\left(\rho_g^c \circ \phi^c\right)\right](F_m^\mathcal{P} \times F_n^c).$$

For all $F^{\mathcal{G}} \in \mathfrak{G}$, we have

$$\begin{aligned}
\int_{F^{\mathcal{G}}} h(g) d\upsilon(g) &= \int_{F^{\mathcal{G}}} \left(\int_{F_m^{\mathcal{P}}} \left[(\phi^c \circ \xi)(F_n^c | p, g) \right] d\nu_0(p) \right) d\upsilon(g) \\
&= \int_{F_m^{\mathcal{P}} \times F^{\mathcal{G}}} \left[(\phi^c \circ \xi)(F_n^c | p, g) \right] d(\nu_0 \otimes \upsilon)(p, g) \\
&= \left[(\nu_0 \otimes \upsilon)(\phi^c \circ \xi) \right] (F_m^{\mathcal{P}} \times F^{\mathcal{G}} \times F_n^c) \\
&= \left[(\nu_0 \otimes \upsilon)(\rho^c \circ (\phi^c \otimes \mathtt{Id}_{\mathcal{G}})) \right] (F_m^{\mathcal{P}} \times F^{\mathcal{G}} \times F_n^c) \qquad \text{(C.6.25)} \\
&= \int_{F_m^{\mathcal{P}} \times F^{\mathcal{G}}} \left[(\rho_g^c \circ \phi^c)(F_n^c | p) \right] d(\nu_0 \otimes \upsilon)(p, g) \\
&= \int_{F^{\mathcal{G}}} \left(\int_{F_m^{\mathcal{P}}} \left[(\rho_g^c \circ \phi^c)(F_n^c | p) \right] d\nu_0(p) \right) d\upsilon(g) \\
&= \int_{F^{\mathcal{G}}} h'(g) d\upsilon(g),
\end{aligned}$$

where the line (C.6.25) uses equation (C.6.24). As this is true for all $F^{\mathcal{G}} \in \mathfrak{G}$, this implies (see Lemma C.2.11) that there exists some measurable $\mathcal{G}_{m,n}^c \in \mathfrak{G}$ such that $\upsilon(\mathcal{G}_{m,n}^c) = 1$ and for all $g \in \mathcal{G}_{m,n}^c$,

$$\left[\nu_0(\phi^c \circ \xi_g) \right] (F_m^{\mathcal{P}} \times F_n^c) = h(g) = h'(g) = \left[\nu_0\big(\rho^c \circ (\phi^c \otimes \psi_g)\big) \right] (F_m^{\mathcal{P}} \times F_n^c). \qquad \text{(C.6.26)}$$

Let us now define $\tilde{\mathcal{G}}^c := \bigcap_{m,n \in \mathbb{N}} \mathcal{G}_{m,n}^c$. As a countable intersection of measurable sets of probability one, $\tilde{\mathcal{G}}^c$ is measurable and $\upsilon(\tilde{\mathcal{G}}^c) = 1$. Moreover, for $g \in \tilde{\mathcal{G}}^c$, equation (C.6.26) holds for all $m, n \in \mathbb{N}$, and thus, using the fact that they generate $\mathfrak{P} \otimes \mathfrak{X}^c$ (and Proposition C.2.9), we get

$$\nu_0(\phi^c \circ \xi_g) = \nu_0(\rho_g^c \circ \phi^c). \qquad \text{(C.6.27)}$$

Let now θ^c be a mod 0 inverse of the measured isomorphism ϕ^c from $(\mathcal{P}, \nu_0)$ to $(\mathcal{X}^c, \epsilon^c)$. Note, first, that we clearly have

$$\theta^c \cdot \epsilon^c = \nu_0. \qquad \text{(C.6.28)}$$

Moreover, combining equation (C.6.27) and point $(viii)$ in Lemma C.3.1 yields, for all $g \in \tilde{\mathcal{G}}^c$,

$$\epsilon^c(\xi_g \circ \theta^c) = \epsilon^c(\theta^c \circ \rho_g^c). \qquad \text{(C.6.29)}$$

Define then $\tilde{\mathcal{G}} := \bigcap_{c \in C} \tilde{\mathcal{G}}^c$. As C is countable, we obtain a measurable set such that $\upsilon(\tilde{\mathcal{G}}) = 1$, and such that equation (C.6.27) holds for all $c \in C$ and all $g \in \tilde{\mathcal{G}}$. Combining this with equation (C.6.28), we obtain that for all $c \in C$ and all $g \in \tilde{\mathcal{G}}$, the map θ^c is a stationary Markov chain isomorphism from (ϵ^c, ρ_g^c) to (ν_0, ξ_g). Eventually, defining the map $\theta : \mathcal{X} \to \mathcal{P}$ by $\theta(x) := \theta^c(x)$ if $x \in C$, the measurability of θ can be proven similarly as we proved the measurability of ϕ in the proof of the direction $(i) \Rightarrow (ii)$ (this uses the countability of C).

Combining now θ with the projection on orbits κ, we indeed obtain a measured class-pose decomposition in the sense of Definition 3.5.16.

The second part of the statement is a direct consequence of how, in this proof, we constructed the measured class-pose decomposition and the isomorphic class-pose parametrisation from one-another.

This ends the proof of Theorem 3.5.17. □

C.7 Appendix for Section 3.6

C.7.1 Proof of Theorem 3.6.1

Theorem 3.6.1. *Let us recall that* $\mathsf{pr} : \mathcal{X} \to C$ *denotes the projection on ergodic components w.r.t. the Markov chain* τ*. The following holds:*

(i) For $\lambda = \Lambda$*, a channel* $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$ *is a solution to* (3.6.10) *if and only if* $\kappa = \iota \circ \mathsf{pr}$*, for a congruent*[8] *channel* $\iota \in \mathscr{K}_{\mathsf{cong}}(C, \mathcal{T})$*.*

(ii) For all $0 \leq \lambda \leq \Lambda$*, all solutions* $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$ *to* (3.6.10) *satisfy* $\kappa = \gamma \circ \mathsf{pr}$*, for some* $\gamma \in \mathscr{K}(\mathcal{X}, \mathcal{T})$*.*

We want to apply Theorem B.1.3 Chapter 2 to the space $\mathcal{A} := \mathcal{X}$, the probability $\mu := \mu_0$, the distortion $D : \mathscr{K}(\mathcal{X}, \mathcal{T}) \to [0, \bar{I}(X, X')]$ defined by $D(\kappa) := \bar{I}_\kappa(T, T')$, and the projection on ergodic components $\mathsf{pr} : \mathcal{X} \to C$. We thus need to verify that assumptions (a) and (b) in the latter theorem do hold here.

Let us define the channel $\epsilon \in \mathscr{K}(C, \mathcal{X})$ by $\epsilon := (x|c) := \epsilon^c(x)$. This channel coincides with the channel ϵ defined in Appendix B.1 in Chapter 2. More precisely:

Lemma C.7.1. *The following holds:*

(i) We have

$$\big[\bar{q}(X, X')(\mathsf{pr} \otimes \mathsf{pr})\big]^{\mathsf{T}} = \big[(\mathsf{pr} \otimes \mathsf{pr}) \cdot \bar{q}(X, X')\big](\epsilon \otimes \epsilon), \tag{C.7.1}$$

where we used the hook-up and transpose notations (see Definitions 3.2.3 and 3.2.4). In particular,

$$\bar{q}(X, X') = \big[(\epsilon \otimes \epsilon) \circ (\mathsf{pr} \otimes \mathsf{pr})\big] \cdot \bar{q}(X, X')$$

(ii) We have $\epsilon(x|c) = \frac{\mu(x)}{(\mathsf{pr} \cdot \mu)(c)} \delta_{\mathsf{pr}(x)=c}$ *for all* $x \in \mathcal{X}$*,* $c \in C$*.*

Proof. (i). To alleviate notations, here we denote $\bar{q}(X, X')$ simply by $\bar{q}$. From equation (3.6.3), we have

$$\begin{aligned}
\big[(\mathsf{pr} \otimes \mathsf{pr}) \cdot \bar{q}\big](c, c') &= \sum_{(x,x') \in \mathcal{X} \times \mathcal{X}} \bar{q}(x, x') \delta_{(\mathsf{pr} \otimes \mathsf{pr})(x,x') = (c,c')} \\
&= \sum_{(x,x') \in \mathcal{X} \times \mathcal{X}} \bar{q}(x, x') \delta_{(x,x') \in \mathcal{X}^c \times \mathcal{X}^{c'}} \\
&= \sum_{(x,x') \in \mathcal{X} \times \mathcal{X}} \left(\sum_{c'' \in C} \mu(\mathcal{X}^{c''}) \epsilon^{c''}(x) \epsilon^{c''}(x') \delta_{(x,x') \in \mathcal{X}^{c'} \times \mathcal{X}^{c'}} \right) \delta_{(x,x') \in \mathcal{X}^c \times \mathcal{X}^{c'}} \\
&= \mu(\mathcal{X}^c) \delta_{c=c'} \sum_{(x,x') \in \mathcal{X} \times \mathcal{X}} \epsilon^c(x) \epsilon^c(x') \, \delta_{(x,x') \in \mathcal{X}^c \times \mathcal{X}^c} \\
&= \mu(\mathcal{X}^c) \delta_{c=c'},
\end{aligned} \tag{C.7.2}$$

[8] See Definition 2.2.2.

where the third line uses equation (3.6.4), and the last line that ϵ^c is concentrated on $\mathcal{X}^c$ (see point (i) in Theorem 3.3.7). Therefore, for all $c, c' \in C$,

$$\begin{aligned}
\big[\bar{q}(\mathsf{pr}\otimes\mathsf{pr})\big](x,x',c,c') &= \bar{q}(x,x')\,\delta_{(x,x')\in\mathcal{X}^c\times\mathcal{X}^{c'}} \\
&= \left(\sum_{c''\in C}\mu(\mathcal{X}^{c''})\epsilon^{c''}(x)\epsilon^{c''}(x')\delta_{(x,x')\in\mathcal{X}^{c''}\times\mathcal{X}^{c''}}\right)\delta_{(x,x')\in\mathcal{X}^c\times\mathcal{X}^{c'}} \qquad \text{(C.7.3)}\\
&= \mu(\mathcal{X}^c)\,\delta_{c=c'}\,\epsilon^c(x)\epsilon^c(x')\,\delta_{(x,x')\in\mathcal{X}^c\times\mathcal{X}^{c'}} \\
&= \mu(\mathcal{X}^c)\,\delta_{c=c'}\,\epsilon^c(x)\epsilon^{c'}(x')\,\delta_{(x,x')\in\mathcal{X}^c\times\mathcal{X}^{c'}} \\
&= \big[(\mathsf{pr}\otimes\mathsf{pr})\cdot\bar{q}\big](c,c')\,\epsilon^c(x)\epsilon^{c'}(x')\,\delta_{(x,x')\in\mathcal{X}^c\times\mathcal{X}^{c'}} \qquad \text{(C.7.4)}\\
&= \big[(\mathsf{pr}\otimes\mathsf{pr})\cdot\bar{q}\big](c,c')\,\epsilon^c(x)\epsilon^{c'}(x') \qquad \text{(C.7.5)}\\
&= \big[\big((\mathsf{pr}\otimes\mathsf{pr})\cdot\bar{q}\big)(\epsilon\otimes\epsilon)\big]\,(c,c',x,x'),
\end{aligned}$$

where line (C.7.3) uses equation (3.6.4); line (C.7.4) uses equation (C.7.2) above; and line (C.7.5) that ϵ^c is concentrated on $\mathcal{X}^c$ (see point (i) in Theorem 3.3.7). This proves equation (C.7.1) in point (i). By marginalising each side of the latter equation on the coordinate $\mathcal{X}\times\mathcal{X}$, we obtain

$$\begin{aligned}
\bar{q}(X,X') &= (\epsilon\otimes\epsilon)\cdot\big((\mathsf{pr}\otimes\mathsf{pr})\cdot\bar{q}(X,X')\big) \\
&= \big((\epsilon\otimes\epsilon)\circ(\mathsf{pr}\otimes\mathsf{pr})\big)\cdot\bar{q}(X,X'),
\end{aligned}$$

which proves point (i).

(ii). By marginalising the first equality in point (i) on (X,T), we get

$$(\mu\mathsf{pr})^{\mathsf{T}} = (\mathsf{pr}\cdot\mu)\epsilon,$$

i.e., for $x\in\mathcal{X}$, $c\in C$,

$$\mu(x)\delta_{\mathsf{pr}(x)=c} = (\mathsf{pr}\cdot\mu)(c)\epsilon(x|c).$$

Note that, here, as we assumed μ fulll-support and as pr is surjective, we have $(\mathsf{pr}\cdot\mu)(c)>0$ for all $c\in C$. □

As in Appendix B.1 in Chapter 2, let us thus define $\bar{\kappa} := \kappa\circ\epsilon\circ\mathsf{pr}$. The following shows that assumption (a) in Theorem B.1.3 indeed holds:

Lemma C.7.2. *$\bar{q}_\kappa(T,T') = \bar{q}_{\bar{\kappa}}(T,T')$, and, in particular, $\bar{I}_\kappa(T;T') = \bar{I}_{\bar{\kappa}}(T;T')$.*

Proof. We have

$$\begin{aligned}
\bar{q}_\kappa(T,T') &= (\kappa\otimes\kappa)\cdot\bar{q}(X,X') \\
&= (\kappa\otimes\kappa)\cdot\Big[\big((\epsilon\otimes\epsilon)\circ(\mathsf{pr}\otimes\mathsf{pr})\big)\cdot\bar{q}(X,X')\Big] \\
&= \Big[(\kappa\otimes\kappa)\circ(\epsilon\otimes\epsilon)\circ(\mathsf{pr}\otimes\mathsf{pr})\Big]\cdot\bar{q}(X,X') \\
&= \Big[(\kappa\circ\epsilon\circ\mathsf{pr})\otimes(\kappa\circ\epsilon\circ\mathsf{pr})\Big]\cdot\bar{q}(X,X') \\
&= (\bar{\kappa}\otimes\bar{\kappa})\cdot\bar{q}(X,X') \\
&= \bar{q}_{\bar{\kappa}}(T,T'),
\end{aligned}$$

where the second line uses point (i) in Lemma C.7.1. □

Next, let us show that assumption (*b*) in Theorem B.1.3 holds as well, which will end the proof of Theorem 3.6.1:

Lemma C.7.3. *The following are equivalent:*

(i) The exists a function $h : \mathcal{T} \to \mathcal{C}$ *such that* $\mathsf{pr} = h \circ \kappa$.

(ii) $\bar{I}_\kappa(T;T') = \bar{I}(X;X')$

Proof. Using the chain rule for mutual information and the Markov chain $T - X - X' - T'$, one can easily verify that the equality $\bar{I}_\kappa(T;T') = \bar{I}(X;X')$ is equivalent to the Markov chain $X - T - T' - X'$.

If the latter Markov chain holds, then in particullar, we have $X - T - X'$, which means that κ is a sufficient statistic of X w.r.t. X' (see Definition 3.3.14), which is equivalent to the existence of a channel $\gamma \in \mathscr{K}(\mathcal{T}, \mathcal{X})$ such that $\bar{q}(X,X)' = \mu(\gamma \circ \kappa)$ (see Proposition 3.3.15). Therefore, from Proposition 3.3.17, there exists a function $h : \mathcal{X} \to \mathcal{C}$ such that $\mu\mathsf{pr} = \mu(h \circ \kappa)$. But as μ is here full-support, this is equivalent to $\mathsf{pr} = h \circ \kappa$.

Conversely, assume that there exists $h : \mathcal{T} \to \mathcal{X}$ such that $\mathsf{pr} = h \circ \kappa$. As pr is a minimal sufficient statistic both of X w.r.t. X' and of X w.r.t. X',

Let us define the distribution $\bar{q}(X, X', T, T', C)$ by extending $\bar{q}(X, X', T, T')$ through $C = \mathsf{pr}(X)$: i.e., for all $x, x' \in \mathcal{X}$, $t, t' \in \mathcal{T}$, $c \in \mathcal{C}$,

$$\bar{q}(x, x', t, t', c) = \bar{q}_\kappa(x, x', t, t')\delta_{\mathsf{pr}(x)=c}.$$

Then

$$\begin{aligned} \bar{q}_\kappa(X, X', T, T', C) &= \bar{q}(X, X', C)\, \bar{q}_\kappa(T, T'|X, X', C) \\ &= \bar{q}(C)\, \bar{q}(X|C)\, \bar{q}(X'|C)\, \bar{q}_\kappa(T, T'|X, X') \\ &= \bar{q}(C)\, \bar{q}(X|C)\, \bar{q}(X'|C)\, \bar{q}_\kappa(T|X)\, \bar{q}_\kappa(T'|X'),, \end{aligned} \tag{C.7.6}$$

where the conditional distributions are well-defined as the conditioning distributions are full-support; the second line uses the Markov chain $X - C - X'$ (which holds because pr defines a sufficient statistic of X w.r.t. X', see Theorem 3.3.18) and the fact that C is a deterministic function of X; and the third line uses the Markov chain $T - X - X' - T'$, which holds by construction of $\bar{q}_\kappa(X, X', T, T')$. Thus we have

$$T - X - C - X' - T',$$

and in particular,

$$(T, X) - C - (X', T'). \tag{C.7.7}$$

On the other hand, the equality $\mathsf{pr} = h \circ \kappa$ implies the Markov chains

$$X - T - C \tag{C.7.8}$$

and

$$C - T' - X'. \tag{C.7.9}$$

Combining (C.7.7), (C.7.8) and (C.7.9) yields, with a computation similar as in (C.7.6), the Markov chain

$$X - T - C - T' - X',$$

and, a fortiori,

$$X - T - T' - X'.$$

□

This ends the proof of Theorem 3.6.1.

C.7.2 Proof of Corollary 3.6.2

Corollary 3.6.2. *Let $\rho \in \mathscr{K}(\mathcal{X} \times \mathcal{G}, \mathcal{X})$ be an action of a finite group $\mathcal{G}$ on a finite state-space $\mathcal{X}$, and $\bar{\rho}$ defined as in (3.6.12). Let $\mu \in \Delta_{\mathcal{X}}$ be full-support and $\bar{\rho}$-stationary, let $q(X, X') := \mu\bar{\rho}$ and for all $\kappa \in \mathscr{K}(\mathcal{X}, \mathcal{T})$ with $\mathcal{T}$ countable,*

$$q_\kappa(X, X', T, T') := q(X, X')(\kappa \otimes \kappa) \in \Delta_{\mathcal{X}\times\mathcal{X}\times\mathcal{T}\times\mathcal{T}},$$

Then a channel κ is a solution of

$$\underset{\substack{\kappa\in\mathscr{K}(\mathcal{X},\mathcal{T})\\ I_\kappa(T;T')=I(X;X')}}{\arg\min} \quad I_\kappa(X;T) \tag{3.6.13}$$

if and only if $\kappa = \iota\circ\mathsf{pr}$, where $\iota \in \mathscr{K}_{\mathsf{cong}}(C, \mathcal{T})$ is a congruent channel, and $\mathsf{pr} \in \mathscr{K}(\mathcal{X}, C)$ is the projection on orbits w.r.t. the action ρ.

Proof. As mentioned before the statement of Corollary 3.6.2, the channel $\bar{\rho}$ defined in (3.6.12) is the update channel of the MDP (π_υ, ρ), where υ is the uniform distribution on $\mathcal{G}$ — which is the unique group-stationary probability — and π_υ the corresponding independent policy (see Definition 3.4.4). We are thus under the assumptions of Theorem 3.4.6 which shows that, here, the projection on ergodic components coincides with the projection on orbits.

Moreover, for all $n \geq 1$ and $x_0 \in \mathcal{X}$:

$$\begin{aligned}
\delta_{x_0}\bar{\rho}^n &= (\mathtt{Id} \otimes \bar{\rho}^n) \cdot (\delta_{(x_0,x_0)}) \\
&= (\mathtt{Id} \otimes \bar{\rho})^n \cdot (\delta_{(x_0,x_0)}) \\
&= (\mathtt{Id} \otimes \bar{\rho}) \cdot (\delta_{(x_0,x_0)}) \\
&= \delta_{x_0}\bar{\rho},
\end{aligned}$$

where the third line applies point (i) in Theorem 3.4.6 to the action $\mathtt{Id}_{\mathcal{X}} \otimes \bar{\rho}$ of $\mathcal{G}$ on the standard Borel space $\mathcal{X} \times \mathcal{X}$. Therefore, by linearity of the hook-up operation,

$$\mu\bar{\rho}^n = \mu\bar{\rho}.$$

I.e., if we apply the discussion above Theorem 3.6.1 to $\tau := \bar{\rho}$, we have $q(X_0, X_n) = q(X_0, X_1)$ for all $n \geq 1$, and thus, from equation (3.6.5) ,

$$\bar{q}(X, X', T, T') = q(X_0, X_1, T_0, T_1) = (\mu\bar{\rho})(\kappa \otimes \kappa)$$

The result is then a consequence of point (i) in Theorem 3.6.1. This ends the proof of Corollary 3.6.2. □

C.7.3 Proof of Theorem 3.6.9

Theorem 3.6.9. *Let $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$ be a finite family of finite-alphabet MDPs. Then:*

(i) For (ν_0, η, ξ), (ν_0', η', ξ') two joinings of $(\mu_0^c, \pi^c, \rho^c)_{c\in C}$, we have, using the j-factor notation (see Definition 3.5.10),

$$(\nu_0', \eta', \xi') \lesssim (\nu_0, \eta, \xi) \quad \Rightarrow \quad H(\nu_0'\eta'\xi') \leq H(\nu_0\eta\xi). \tag{3.6.16}$$

(ii) A joining has minimum entropy if and only if it is a minimal joining.

(iii) All minimal joinings are isomorphic as stationary MDPs.

Proof. Consider joinings (ν_0, η, ξ) and (ν_0', η', ξ'), with resp. state-space $\mathcal{P}$, $\mathcal{P}'$ and action space $\mathcal{K}$, $\mathcal{K}'$, such that

$$(\nu_0', \eta', \xi') \lesssim (\nu_0, \eta, \xi),$$

and denote by (Φ, Ψ) the factor maps that make the joining (ν_0', η', ξ') a j-factor of (ν_0, η, ξ). Let us also write $\mathcal{A}' := \mathcal{P}' \times \mathcal{K}' \times \mathcal{P}'$, $\mathcal{A} := \mathcal{P} \times \mathcal{K} \times \mathcal{P}$, $q' := \nu_0'\eta'\xi'$, $q := \nu_0\eta\xi$ and $f := \Phi \otimes \Psi \otimes \Phi$. Then, from Proposition 3.5.2, we have

$$f \cdot q = q'. \tag{C.7.10}$$

As f is deterministic, this implies

$$H(q') \leq H(q), \tag{C.7.11}$$

which proves point (i) in Theorem 3.6.9. Moreover,

$$\begin{cases} f \cdot q = q' \\ H(q') = H(q) \end{cases} \quad \Leftrightarrow \quad \begin{cases} f \cdot q = q' \\ f \text{ induces a bijection from } \mathrm{supp}(q') \text{ to } \mathrm{supp}(q) \end{cases}$$
$$\Leftrightarrow \quad f \text{ is a measured isomorphism from } (\mathcal{A}', q') \text{ to } (\mathcal{A}, q), \tag{C.7.12}$$

where the first line uses the strict concavity of entropy combined with the fact that $f \cdot q = q'$ implies that for all $p' \in \mathcal{P}'$, the probability $q'(p')$ is a convex combination of elements of $\{q(p)\}_{p\in\mathcal{P}}$; and the second line uses Lemma C.6.2.

Now, from Theorem 3.5.13, there always exists a minimal joining. Let us denote it by (ν_0', η', ξ'): as by definition, (ν_0', η', ξ') is a j-factor of (ν_0, η, ξ), the reasonings above still apply with the same notations.

On the one hand, the inequality (C.7.11), for fixed $q' = \nu_0'\eta'\xi'$ and all $q = \nu_0\eta\xi$, proves that the minimal joining (ν_0', η', ξ') is a minimum entropy joining.

Conversely, assume that (ν_0, η, ξ) is a minimum entropy joining. As it is a joining, we know from (C.7.10) that $f \cdot q = q'$. As (ν_0', η', ξ') is a joining and (ν_0, η, ξ) a minimum entropy joining, the equality is achieved in (C.7.11). Thus, from (C.7.12), the map $\Phi \otimes \Psi \otimes \Phi$ is a measured isomorphism from $(\mathcal{P}\times\mathcal{K}\times\mathcal{P}, \nu_0\eta\xi)$ to $(\mathcal{P}'\times\mathcal{K}'\times\mathcal{P}', \nu_0'\eta'\xi')$. Then, Lemma C.3.2 yields that Φ, resp. Ψ, is a measured isomoprhism from $(\mathcal{P}, \nu_0)$ to $(\mathcal{P}', \nu_0')$, resp. from $(\mathcal{K}, \eta\cdot\nu_0)$ to $(\mathcal{K}', \eta' \cdot \nu_0')$, and that, denoting by Φ^{-1} and Ψ^{-1} mod 0 inverses of resp. Φ and Ψ, we have

$$\Phi^{-1} \otimes \Psi^{-1} \otimes \Phi^{-1} \cdot \nu_0'\eta'\xi' = \nu_0\eta\xi. \tag{C.7.13}$$

From Proposition 3.5.2, equation (C.7.13) implies that (ν_0, η, ξ) is an MDP factor of (ν_0', η', ξ').

By marginalisation of equation (C.7.13), we also get

$$\Phi^{-1} \otimes \Psi^{-1} \cdot \nu_0'\eta' = \nu_0\eta,$$

which means that $\Phi^{-1} \otimes \Psi^{-1}$ is a measured isomorphism from $(\mathcal{P}' \times \mathcal{K}', \nu_0' \eta')$ to $(\mathcal{P} \times \mathcal{K}, \nu_0 \eta)$. But as we already showed that (ν_0, η, ξ) is an MDP factor of (ν_0', η', ξ'), this proves, from Lemma C.6.1, that (ν_0, η, ξ) is an j-factor of (ν_0', η', ξ').

Eventually, as (ν_0', η', ξ') is a minimal joining and by transitivity of the j-factor relation (see Proposition 3.5.11), this implies that (ν_0, η, ξ) is a j-factor of any other joining: i.e., that it is a minimal joining. Thus point (ii) in Theorem 3.6.9 is proven.

Moreover, we proved along the way that any minimum entropy joining is MDP isomorphic to any minimal joining. Combined with the fact that minimum entropy joinings coincide with minimal joinings which we just proved, we obtain that two minimal joinings are always MDP isomorphic. This proves point (iii).

This ends the proof of Theorem 3.6.9. □

Appendix D

Appendix for Chapter 4

D.1 Section 4.2 Details

D.1.1 Proof of Proposition 4.2.2

$(i) \Rightarrow (ii)$: Suppose that there are variables T_1 and $T_i := (T_{i-1}, S_i)$ for $2 \leq i \leq n$ such that each T_i is a bottleneck with parameter λ_i. Unrolling the iterative definitions of the T_i, we obtain

$$T_i = (T_1, S_2, \dots, S_i),$$

which implies that, if $j < i$, then T_j is a deterministic function of T_i; in other words, given T_i, the variable T_j is independent of any other variable. So, first, we have $X - T_n - T_{n-1}$. Now, assume that for a given i, we have

$$X - T_n - \cdots - T_i. \tag{D.1.1}$$

Given T_i, the variable T_{i-1} is independent of any other variable, so, in particular,

$$(X, T_n, \dots, T_{i+1}) - T_i - T_{i-1}. \tag{D.1.2}$$

The Markov chains (D.1.1) and (D.1.2) together imply that

$$X - T_n - \cdots - T_{i-1}.$$

Thus, a recurrence from $i = n$ to $i = 1$ proves that we do have $X - T_n - \cdots - T_1$, where, by assumption, each T_i is indeed a bottleneck of parameter λ_i.

$(iii) \Rightarrow (i)$: For all i, the Markov chain (4.2.2) implies that

$$I(X; T_i) = I(X; T_i'),$$
$$I(Y; T_i) = I(Y; T_i'),$$

where $T_i' := (T_i, \dots, T_1)$. The Markov chain (4.2.2) also implies that these T_i' satisfy $Y - X - T_i'$. Thus, the T_i' are also bottlenecks with respective trade-off parameters $\lambda_1, \dots, \lambda_n$. But, by construction, they satisfy $T_i' = (T_{i-1}', S_i)$, where, here, $S_i := T_i$.

$(ii) \Rightarrow (iii)$. We merely define $q(X, T_1, \dots, T_n, Y)$ through the density

$$q(x, t_1, \dots, t_n, y) := q(x, t_1, \dots, t_n) q(y|x).$$

From this construction and the fact that each individual bottleneck must by definition satisfy $Y - X - T_i$, it is clear that $q(X, T_1, \dots, T_n, Y)$ is indeed an extension of the individual

bottleneck probabilities $q(X, Y, T_i)$. Moreover, by construction, we have

$$Y - X - (T_n, \dots, T_1).$$

This latter Markov chain, combined with the assumed Markov chain (4.2.1), together imply that the Markov chain (4.2.2) holds.

This ends the proof of Proposition 4.2.2.

D.1.2 Operational Interpretation of Successive Refinement

This section describes the operational interpretation—for the case of discrete variables X,Y—of successive refinement, which was already proposed in (Tian et al., 2008) and (Tuncel, 2009), as well as, in a slightly more general fashion, in (Mahvari et al., 2020). We will here rely on the content from the latter work (even though our notations will be different). We will denote, for a variable Z, by Z^l, the concatenation of l i.i.d. variables with the same law as Z.

Definition D.1.1. For $l \in \mathbb{N}$, an *n-stage* $(l, M_1, \dots, M_n)$*-code* consists of n encoder functions

$$\phi_i^l \ : \ \mathcal{X}^l \to \{1, \dots, M_i\}$$

and n decoder functions

$$\psi_i^l \ : \ \{1, \dots, M_1\} \times \cdots \times \{1, \dots, M_i\} \to \mathcal{Y}^l.$$

For a given source X, the i-th output of the $(l, M_1, \dots, M_n)$-code will be written

$$\hat{Y}_i^l \ := \psi_i^l(\phi_1^l(X^l), \dots, \phi_i^l(X^l)).$$

Intuitively, each new encoder extracts additional information from the same source, and, crucially, each new decoder is allowed to rely on *all* the information encoded until the i-th stage. Note that the output space of the decoder is modelled on that of the relevancy variable because this is the one about which one wants to extract information.

Definition D.1.2. The *relevance-complexity region* is the set of tuples $(R_1, \dots, R_n, \mu_1, \dots, \mu_n)$ such that there exists a sequence of n-stage $(l, M_1, \dots, M_n)$-codes for all $1 \leq i \leq n$,

$$\forall l \in \mathbb{N}, \quad \frac{1}{l} \log M_i \ \leq \ R_i$$

and

$$\forall l \in \mathbb{N}, \quad \frac{1}{l} I(Y^l; \hat{Y}_i^l) \ \geq \ \mu_i.$$

Intuitively, for a tuple to be in the relevance-complexity region, there must be an n-stage code such that the i-th encoder adds information at a rate no larger than R_i, and the i-th decoder yields information about the target variable Y no lower than μ_i. In other words, the relevance-complexity region is made of all the tuples that are achievable by n-stage codes.

Now, let us give the operational definition of successive refinement. We will denote, for a parameter λ, by $I_Y(\lambda)$, the maximum value of $I(Y;T)$ in the primal IB problem (4.1.2).

Definition D.1.3. Let $0 \leq \lambda_1 < \cdots < \lambda_n$. An IB problem defined by $p(X, Y)$ is said to be *operationally successively refinable*, or *O-SR*, for rates $(\lambda_1, \dots, \lambda_n)$, if the tuple

$$(\lambda_1 \ , \ \lambda_2 - \lambda_1 \ , \ \dots \ , \ \lambda_n - \lambda_{n-1}, \ I_Y(\lambda_1), \ \dots, \ I_Y(\lambda_n))$$

is in the relevance-complexity region.

Intuitively, in the case $n = 2$, assume one is given a total rate λ_2 to "spend" on encoding a source X. One can choose to encode the source in a single processing stage, yielding at best, after decoding, asymptotic relevant information $I_Y(\lambda_2)$ (see (Gilad-Bachrach et al., 2003)). Alternatively, one can choose to break up the total rate λ_2 into two rates $R_1 := \lambda_1 < \lambda_2$ and $R_2 := \lambda_2 - \lambda_1$, and successively encode potentially different aspects of the source at these rates. Operational SR means that even though this second alternative "spends" the total rate λ_2 along two distinct stages, it can still, after decoding, also yield asymptotic relevant information of $I_Y(\lambda_2)$. Naturally, in this case, the relevant information decodable from only the first stage must also be the optimal one, i.e., $I_Y(\lambda_1)$—otherwise, the "waste" in spending the rate λ_1 would prevent the second-stage decoder, which partially relies on the information encoded at the first stage, from ever achieving the optimal relevant information $I_Y(\lambda_2)$.

We then have the following single-letter characterisation:

Proposition D.1.4. *The IB problem defined by $p(X,Y)$ is O-SR for rates $(\lambda_1, \dots, \lambda_n)$ if and only if there exist variables $T_1, \dots, T_n$ such that*

(i) We have the Markov chain $Y - X - T_n - \cdots - T_1$;

(ii) The variables $T_1, \dots, T_n$ are each bottlenecks with respective parameters $\lambda_1, \dots, \lambda_n$.

Proof. This single-letter characterisation is a consequence of Remark 1 in (Mahvari et al., 2020), which states the following: a tuple $(R_1, \dots, R_n, \mu_1, \dots, \mu_n)$ is in the relevance-complexity region if and only if there exist variables $T_1, \dots, T_n$ such that the Markov chain $Y - X - T_n - \cdots - T_1$ holds, and such that, for all $i = 1, \dots, n$,

$$\sum_{j=1}^{i} I(X; T_j | T_1, \dots, T_{j-1}) \leq \sum_{j=1}^{i} R_j, \tag{D.1.3}$$

$$I(Y; T_i) \geq \mu_i. \tag{D.1.4}$$

By simplifying the left-hand side in (D.1.3) through the chain rule for mutual information, defining $\lambda_i := \sum_{j=1}^{i} R_j$, and applying the statement with $\mu_i := I_Y(\lambda_i)$, we obtain that the IB problem is O-SR for rates $(\lambda_1, \dots, \lambda_n)$ if and only if there exist variables $T_1, \dots, T_n$ such that

(i) We have the Markov chain $Y - X - T_n - \cdots - T_1$; and

(ii) We have, for all $i = 1, \dots, n$,

$$I(X; T_i) \leq \lambda_i, \tag{D.1.5}$$

$$I(Y; T_i) \geq I_Y(\lambda_i). \tag{D.1.6}$$

However, if point 1 above holds, then, particularly for all $i = 1, \dots, n$, we have the Markov chain $Y - X - T_i$. As a consequence, by definition of the primal IB problem (4.1.2), the inequality in (D.1.6) can be replaced by an equality, and thus point (ii) as a whole can be replaced by the condition that T_i is a bottleneck of parameter λ_i for the IB problem defined by $p(X,Y)$. Hence, we are left with points (i) and (ii) of Theorem D.1.4's statement. □

It is clear that the conditions of Theorem D.1.4 are exactly those of Proposition 4.2.1-(iii), so the operational Definition D.1.3 and the single-letter Definition 4.2.1 are equivalent; in other words, the notion studied in our work does have an operational interpretation. Crucially, the operational construction of Definitions D.1.1–D.1.3 also goes clearly along the interpretation in terms of the successive incorporation of information.

For completeness, let us also mention that Proposition D.1.4 above is also essentially Theorem 7 in (Tian et al., 2008), which proves the same single-letter characterisation for the same operational problem—up to the difference that the result is limited to $n = 2$, and that the latter work does not consider any decoder functions ψ_i^I. The case $n = 2$ of Proposition D.1.4 is also a consequence of Lemma 4 in (Tuncel, 2009).

D.1.3 Proof of Proposition 4.2.4

First of all, note that even though in the Definition 4.2.1 of successive refinement, the term "bottleneck" refers to a solution to the primal problem (4.1.2), the definition makes as much sense if now by "bottleneck" we mean a solution to the Lagrangian problem (4.1.4). This is, therefore, what we will be speaking about in this section. With this Lagrangian version, the Markov chain characterisation given by Proposition 4.2.2 still holds. More precisely:

Proposition D.1.5. *Let (X, Y) be jointly Gaussian, and $1 \leq \beta_1 < \cdots < \beta_n$. The following are equivalent:*

(*i*) *There is successive refinement for Lagrangian parameters $(\beta_1, \ldots, \beta_n)$.*

(*ii*) *There exist Lagrangian bottlenecks $T_1, \ldots, T_n$, of common source X and relevancy Y, with respective parameters $\beta_1, \ldots, \beta_n$, and an extension $q(Y, X, T_1, \ldots, T_n)$ of the $q_i := q_i(Y, X, T_i)$, such that, under q, we have the Markov chain*

$$Y - X - T_n - \cdots - T_1. \tag{D.1.7}$$

Proof. One can directly verify that the proof given for Proposition 4.2.2 (see Appendix D.1.1) does not involve the explicit form of the IB problem, so the very same proof can be used for the Lagrangian formulation. □

The statement of Proposition 4.2.4 is now fully explicit.

Proof of Proposition 4.2.4. For the case of the Lagrangian IB problem with jointly Gaussian source X and relevancy Y, an analytic solution was given in (Chechik et al., 2005), which proves among other things that the functions $(\beta \mapsto I_\beta(X;T))$ and $(\beta \mapsto I_\beta(Y;T))$ are continuous and increasing, where $I_\beta(X;T)$ and $I_\beta(Y;T)$ are defined by bottlenecks T of Lagrangian trade-off parameter β. Let us define

$$\beta_{IB}(X, Y) := \sup\{\beta \in \mathbb{R} \ : \ I_\beta(X;T) = 0\},$$

where we must have $\beta_{IB}(X, Y) \geq 1$ (see Section 4.1.4). Moreover, from the continuity of the function $(\beta \mapsto I_\beta(X;T))$, this supremum is a maximum, and from the monotonicity of the latter function, $I_\beta(X;T) = 0$ for all $\beta \leq \beta_{IB}(X, Y)$, whereas, by definition of $\beta_{IB}(X, Y)$, we have $I_\beta(X;T) > 0$ for all $\beta > \beta_{IB}(X, Y)$. Thus, $\beta_{IB}(X, Y)$ delimits trivial from non-trivial solutions, and we can, without loss of generality, choose $\beta \geq \beta_{IB}(X, Y)$.

Let us now turn to the *semigroup structure* of the Gaussian IB problem, which was both defined and proved in (Kline et al., 2022). In short, this structure means that one can *compose* two Gaussian bottlenecks, while still obtaining a Gaussian bottleneck for the original problem. More precisely, let $\beta_2 > \beta_{IB}(X, Y)$, and define T_2 as the analytical solution to the Lagrangian IB from (Chechik et al., 2005). This provides one with a joint distribution $q_2(Y, X, T_2)$, which, importantly for us here, happens to define a Gaussian vector as well. Then, we consider a new IB problem with still the same relevancy variable Y, but now with T_2 as the source, i.e.,

$$\underset{q(T_1|T_2) \,:\, T_1 - T_2 - Y,}{\arg\min} \quad I(T_2;T_1) \ - \ \beta_1' \, I(Y;T_1), \tag{D.1.8}$$

where $\beta_1' \geq \beta_{IB}(T_2, Y)$. As T_2 and Y are jointly Gaussian, the problem above is again a Gaussian IB problem, so we can again analytically define a solution T_1 with the formulas from (Chechik et al., 2005), yielding a distribution $q_1(Y, T_2, T_1)$. The semigroup structure proven in (Kline et al., 2022) refers to the following feature:

Proposition D.1.6. *Assume that T_1 and T_2 are built as above, and define the extension $q(Y, X, T_1, T_2)$ of $q_1(Y, X, T_1)$ and $q_2(Y, X, T_2)$ through*

$$q(y, x, t_1, t_2) := q_2(y, x, t_2) q_1(t_1|t_2). \tag{D.1.9}$$

Then, the marginal $q(Y, X, T_1)$ defines a Lagrangian bottleneck of source X and relevancy Y for some parameter β_1 uniquely defined, with $\beta_{IB}(X, Y) \leq \beta_1 < \beta_2$.

Thus, we can define a binary operator "$\circ$", which, for every $\beta_2 > \beta_{IB}(X, Y) \geq 1$ and $\beta_1' \geq \beta_{IB}(T_1, Y)$, provides the parameter $\beta_1 := \beta_2 \circ \beta_1'$ defined by Proposition D.1.6. Ref. (Kline et al., 2022) gives an explicit formula for this binary operator :

$$\beta_1' \circ \beta_2 = \frac{\beta_1' \beta_2}{\beta_1' + \beta_2 - 1}, \tag{D.1.10}$$

which is well-defined for $\beta_2 > \beta_{IB}(X, Y)$ and $\beta_1' \geq \beta_{IB}(T_2, Y)$, because $\beta_{IB}(X, Y) \geq 1$ and $\beta_{IB}(T_2, Y) \geq 1 \geq 0$ imply that $\beta_1' + \beta_2 - 1 > 0$. This formula implies the following:

Proposition D.1.7. *Let $\beta_2 > \beta_{IB}(X, Y)$. For any β_1 such that $\beta_{IB}(X, Y) \leq \beta_1 < \beta_2$, there exists a β_1' such that $\beta_1 = \beta_1' \circ \beta_2$.*

Proof. Let f denote the function $\beta_1' \mapsto \beta_1' \circ \beta_2$, which is well-defined and continuous on the interval $[\beta_{IB}(T_1, Y), +\infty[$. It is clear from formula (D.1.10) that

$$\lim_{\beta_1' \to \infty} f(\beta_1') = \beta_2. \tag{D.1.11}$$

On the other hand, note first that as $\beta_{IB}(T_2, Y)$ delimits trivial from non-trivial solutions, we have $I_{\beta_{IB}(T_2,Y)}(T_2; T_1) = 0$. But, by construction, under q given by Equation (D.1.9), we have the Markov chain $Y - X - T_2 - T_1$. Thus, $I_{\beta_{IB}(T_2,Y) \circ \beta_2}(X; T_1) \leq I_{\beta_{IB}(T_2,Y)}(T_2, T_1)$, i.e., $I_{\beta_{IB}(T_2,Y) \circ \beta_2}(X; T_1) = 0$. So, by definition of $\beta_{IB}(X, Y)$, we have

$$\beta_{IB}(T_2, Y) \circ \beta_2 \leq \beta_{IB}(X, Y), \tag{D.1.12}$$

i.e.,

$$f(\beta_{IB}(T_2, Y)) \leq \beta_{IB}(X, Y). \tag{D.1.13}$$

Now, Equations (D.1.11) and (D.1.13), combined with the continuity of f, imply that

$$[\beta_{IB}(X, Y), \beta_2[\subseteq f\big([\beta_{IB}(T_2, Y), \infty[\big),$$

which yields the result. □

Now let us consider a family of parameters $\beta_{IB}(X, Y) \leq \beta_1 < \cdots < \beta_n$. By iterating Propositions D.1.6 and D.1.7 used together, we obtain that there exist bottlenecks $T_1, \ldots, T_n$ of common source X and relevancy Y, with respective parameters $\beta_1, \ldots, \beta_n$, and an extension $q(Y, X, T_1, \ldots, T_n)$ of these bottlenecks defined by

$$q(y, x, t_1, \ldots, t_n) := q(y, x, t_n) q(t_{n-1}|t_n) \ldots q(t_1|t_2).$$

By construction, under q, the Markov chain $Y - X - T_n - \cdots - T_1$ holds. In other words, condition (ii) from Proposition D.1.5 is satisfied, which proves the successive refinability of jointly Gaussian vectors for the Lagrangian IB problem.

This ends the proof of Proposition 4.2.4. □

D.1.4 Proof of Proposition 4.2.5

Here, for $\alpha \in [0,1]$, we denote by T_α the variable defined by

$$\begin{aligned} q(T_\alpha = Y|X) &= \alpha \\ q(T_\alpha = e|X) &= 1-\alpha, \end{aligned} \tag{D.1.14}$$

where e denotes a dummy symbol not pertaining to either $\mathcal{X}$ or $\mathcal{Y}$. It was proven in (Kolchinsky et al., 2017) that, for every primal parameter $\lambda \in [0, I(X;Y)]$, there exists an α such that T_α is a bottleneck of parameter λ. Note that we must have

$$\lambda = I(X;T_\alpha) = \alpha I(X;Y), \tag{D.1.15}$$

where the first equality comes the general fact that a bottleneck must saturate the information constraint in (4.1.2) (see Section 4.1.4), and the second equality is a direct computation from (D.1.14). Thus, α is a bijective and increasing function of λ, and it is sufficient, for proving successive refinement, to prove that, for $0 \leq \alpha_1 < \cdots < \alpha_n \leq 1$, we can design a joint distribution $q(X, T_{\alpha_1}, \ldots, T_{\alpha_n})$ such that we have the Markov chain

$$X - T_{\alpha_n} - \cdots - T_{\alpha_1}.$$

Let us first focus on the case $n = 2$. We define a bottleneck $T_2 := T_{\alpha_2}$, i.e, we set $q(X, T_2) := q(X, T_{\alpha_2})$ and then a distribution $q(T_1, T_2)$ through

$$\begin{aligned} q(T_1 = Y|T_2 = Y) &:= \frac{\alpha_1}{\alpha_2} \\ q(T_1 = e|T_2 = Y) &:= \frac{\alpha_2 - \alpha_1}{\alpha_2} \\ q(T_1 = Y|T_2 = e) &:= 0 \\ q(T_1 = e|T_2 = e) &:= 1. \end{aligned}$$

We then define an extension $q(X, T_1, T_2)$ of $q(X, T_2)$ and $q(T_1, T_2)$ through

$$q(x, t_1, t_2) := q(x, t_2) q(t_1|t_2),$$

which implies by construction the Markov chain $X - T_2 - T_1$. But it also implies that

$$\begin{aligned} q(T_1 = Y|x) &= q(T_1 = Y|T_2 = Y) q(T_2 = Y|x) + q(T_1 = Y|T_2 = e) q(T_2 = e|x) \\ &= \frac{\alpha_1}{\alpha_2} \alpha_2 + 0 \times (1 - \alpha_2) \\ &= \alpha_1, \end{aligned}$$

and thus, necessarily, $q(T_1 = e|X) = 1 - \alpha_1$. So, $q(X, T_1) = q(X, T_{\alpha_1})$. Thus, we built a joint law $q(X, T_{\alpha_1}, T_{\alpha_2})$ such that $X - T_{\alpha_2} - T_{\alpha_1}$, which proves successive refinement for the case $n = 2$. The case of arbitrary n follows by direct iteration of the previous reasoning, where one starts from defining $q(X, T_n)$ through $T_n := T_{\alpha_n}$, and then iteratively defines $q(X, T_i, T_{i+1}, \ldots, T_n)$ through a well-chosen $q(T_i|T_{i+1})$ and the Markov chain condition $X - T_n - \cdots - T_{i+1} - T_i$.

This ends the proof of Proposition 4.2.5.

D.1.5 Proof of Proposition 4.2.6

The result is a consequence of the following general fact, where we will eventually set $U := T_1$, $V := T_2$, and $W := X$.

Proposition D.1.8. *Let $q(U,W)$ and $q(V,W)$ be full-support consistent distributions, defined on discrete alphabets $\mathcal{U} \times \mathcal{W}$ and $\mathcal{V} \times \mathcal{W}$, respectively. Consider the following properties:*

(*i*) *There exists an extension $\tilde{q}(U,V,W)$ of $q(U,W)$ and $q(V,W)$ under which the Markov chain $U - V - W$ holds.*

(*ii*) *For each $u \in \mathcal{U}$, there exists a family of convex combination coefficients $\{\alpha_{v,u}\,,\ v \in \mathcal{V}\}$ such that*

$$q(W|u) = \sum_v \alpha_{v,u}\, q(W|v).$$

Then, we always have (i) $\Rightarrow$ (ii) and, if, moreover, the channel $q(W|V)$ is injective, then we also have (ii) $\Rightarrow$ (i), and the extension $\tilde{q}$ is uniquely defined.

Note the abuse of notations in the statement of Proposition D.1.8: we write q for both $q(U,V)$ and $q(V,W)$, which are distinct distributions on partially distinct alphabets, even though they are consistent; in addition, along the proof, context, if not explicit statements, will make clear which distribution we are referring to.

Proof. Along the proof, we will be using the fact that a probability distribution is equivalent to a family of convex combination coefficients several times; indeed, both notions define a family of non-negative numbers such that their sum equals one.

(*i*) $\Rightarrow$ (*ii*). For all u, w, assumption (*i*) provides a $\tilde{q}(U,V,W)$ such that

$$\begin{aligned} q(w|u) &= \tilde{q}(w|u) \\ &= \sum_v \tilde{q}(w,v|u) \\ &= \sum_v \tilde{q}(v|u)\, \tilde{q}(w|v) \\ &= \sum_v \tilde{q}(v|u)\, q(w|v), \end{aligned}$$

where the first and fourth equalities use the fact that $\tilde{q}(U,V,W)$ is an extension of $q(U,W)$ and $q(V,W)$, and the third equality uses the fact that, under $\tilde{q}(U,V,W)$, the Markov chain $U - V - W$ holds. Let us define $\alpha_{v,u} := q(v|u)$. For each $u \in \mathcal{U}$, the family $\{\alpha_{v,u},\ v \in \mathcal{V}\}$ is a probability distribution, and thus a family of convex combination coefficients.

(*ii*) $\Rightarrow$ (*i*). We want to design a distribution $\tilde{q}$ that is both consistent with $q(U,W)$ and $q(V,W)$, and satisfies $U - V - W$. Thus, such a distribution is wholly defined by $\tilde{q}(V|U)$, because it must satisfy

$$\begin{aligned} \tilde{q}(u,v,w) &= \tilde{q}(u)\tilde{q}(v|u)\tilde{q}(w|v) \\ &= q(u)\tilde{q}(v|u)q(w|v), \end{aligned} \tag{D.1.16}$$

where $q(U)$ is obtained by marginalising $q(U,W)$, whereas $q(W|V)$ is obtained from $q(V,W)$. Assumption (ii) provides a candidate: let us define $\tilde{q}(v|u) := \alpha_{v,u}$, which makes sense because, for each u, the family $(\alpha_{v,u})_v$ is made of convex combination coefficients. From assumption (ii), for all u, w,

$$q(w|u) = \sum_v \tilde{q}(v|u)\, q(w|v), \tag{D.1.17}$$

and the corresponding $\tilde{q}(U,V,W)$ defined through Equation (D.1.16) satisfies the Markov chain $U - V - W$.

To prove that $\tilde{q}$ is an extension of $q(U,W)$ and $q(V,W)$, let us prove first that $\tilde{q}$ is consistent with $q(U,W)$. We have

$$\begin{aligned}
\tilde{q}(u,w) &= \sum_v \tilde{q}(u,v,w) \\
&= \sum_v q(u)\tilde{q}(v|u)q(w|v) \\
&= q(u) \sum_v \tilde{q}(v|u)q(w|v) \\
&= q(u)q(w|u) \\
&= q(u,w),
\end{aligned}$$

where the first equality is the definition of the marginal $\tilde{q}(u,w)$; the second equality uses Equation (D.1.16); and the fourth equality uses (D.1.17). Thus, $\tilde{q}(U,V,W)$ is consistent with $q(U,W)$.

Now, let us prove that $\tilde{q}(V,W) = q(V,W)$. This is equivalent to the channel $\tilde{q}(V|U)$ sending the marginal $q(U)$ on the marginal $q(V)$:

Lemma D.1.9. *We have $\tilde{q}(V,W) = q(V,W)$ if and only if*

$$\tilde{Q}_{vu} q_u = q_v, \tag{D.1.18}$$

where q_u and q_v are the column vectors defined by $q(U)$ and $q(V)$, respectively, and $\tilde{Q}_{vu}$ is the column transition matrix defined by $\tilde{q}(V|U)$.

Proof. For all v, w,

$$\begin{aligned}
\tilde{q}(v,w) &= \sum_u \tilde{q}(u,v,w) \\
&= \Big(\sum_u q(u)\tilde{q}(v|u) \Big) q(w|v),
\end{aligned}$$

where the first equality is the definition of the marginal $\tilde{q}(v,w)$, and the second one uses Equation (D.1.16). Thus, for all v, w,

$$\begin{aligned}
\tilde{q}(v,w) = q(v,w) \quad &\Leftrightarrow \quad q(v,w) = \Big(\sum_u q(u)\tilde{q}(v|u) \Big) \tilde{q}(w|v) \\
&\Leftrightarrow \quad q(v)q(w|v) = \Big(\sum_u q(u)\tilde{q}(v|u) \Big) q(w|v),
\end{aligned}$$

and, eventually, for all v, w,

$$\tilde{q}(v,w) = q(v,w) \quad \Leftrightarrow \quad q(w|v) = 0 \quad \text{or} \quad q(v) = \sum_u q(u)\tilde{q}(v|u). \tag{D.1.19}$$

Let us momentarily fix $v \in \mathcal{V}$. Since $q(W|v)$ is a probability, there must be some w_0 such that $q(w_0|v) > 0$. Choosing that w_0, we find that, for the given v, the vector equality $\tilde{q}(v, W) = q(v, W)$ implies, through Equation (D.1.19), that the scalar equality $q(v) = \sum_u q(u)\tilde{q}(v|u)$. By now applying this reasoning to each $v \in \mathcal{V}$, we obtain that $\tilde{q}(V, W) = q(V, W)$ implies that

$$\forall v \in \mathcal{V}, \quad \sum_u q(u)\tilde{q}(v|u) = q(v), \tag{D.1.20}$$

whose matrix formulation is precisely (D.1.18). Conversely, if (D.1.20) holds, then Equation (D.1.19) shows that $\tilde{q}(V, W) = q(V, W)$. □

We now prove that Equation (D.1.18) indeed holds. Let us also write Q_{wv} and Q_{wu} for the column transition matrices defined by $q(W|V)$ and $q(W|U)$, respectively. Then, Equation (D.1.17), which, here, is our assumption, can be rewritten as

$$Q_{wu} = Q_{wv}\tilde{Q}_{vu}. \tag{D.1.21}$$

Thus,

$$Q_{wv}\tilde{Q}_{vu}q_u = Q_{wu}q_u = q_w = Q_{wv}q_v$$

where q_w is the column vector defined by $q(W)$, and the second and third equalities are the matrix versions of the decompositions $q(W) = \sum_u q(u)q(W|u)$ and $q(W) = \sum_v q(v)q(W|v)$, respectively. In other words,

$$Q_{wv}(\tilde{Q}_{vu}q_u - q_v) = 0. \tag{D.1.22}$$

The injectivity of Q_{wv} implies that (D.1.18) indeed holds, so, from Lemma D.1.9, we have $\tilde{q}(V, W) = q(V, W)$. We have thus proven that $\tilde{q}$ extends both $q(U, W)$ and $q(V, W)$, so point (ii) holds.

Eventually, let us prove the uniqueness. Let $\tilde{q}' := \tilde{q}'(U, V, W)$ be another extension of $q(U, W)$ and $q(V, W)$ such that, under $\tilde{q}'$, the Markov chain $U - V - W$ holds. For the same reasons as above, $\tilde{q}'$ must satisfy Equation (D.1.16) with $\tilde{q}$ replaced by $\tilde{q}'$, so $\tilde{q}'$ is wholly specified by $\tilde{q}'(V|U)$, and is enough to prove that $\tilde{q}'(V|U) = \tilde{q}(V|U)$. Now, using the assumptions of consistency and the Markov chain for $\tilde{q}'$, we obtain

$$\begin{aligned} q(w|u) &= \tilde{q}'(w|u) \\ &= \sum_v \tilde{q}'(v, w|u) \\ &= \sum_v \tilde{q}'(v|u)\tilde{q}'(w|v) \\ &= \sum_v \tilde{q}'(v|u)q(w|v), \end{aligned} \tag{D.1.23}$$

i.e., in matrix terms, if $\tilde{Q}'_{uv}$ is the column transition matrix representing $\tilde{q}'(V|U)$,

$$Q_{wu} = Q_{wv}\tilde{Q}'_{vu}.$$

Combining this with Equation (D.1.21), we have $Q_{wv}(\tilde{Q}'_{vu} - \tilde{Q}_{vu}) = 0$. In other words, if c_i is the i-th column of $\tilde{Q}'_{vu} - \tilde{Q}_{vu}$, then $Q_{wv}c_i = 0$, which, by injectivity of Q_{wv}, means that $c_i = 0$. Thus, $\tilde{Q}'_{vu} - \tilde{Q}_{vu} = 0$, i.e., $\tilde{q}(U|V) = \tilde{q}'(U|V)$.

This ends the proof of Proposition D.1.8. □

Now, first of all, note that if we set $U := T_1$, $V := T_2$ and $W := X$, then point (ii) in Proposition D.1.8 is equivalent to the convex hull condition (4.2.4).

If there is successive refinement for parameters (λ_1, λ_2), then, from Proposition 4.2.2, there are bottlenecks T_1, T_2 of parameters λ_1, λ_2, respectively, such that $X - T_2 - T_1$; and the direction $(i) \Rightarrow (ii)$ of Proposition D.1.8 implies that the convex hull condition (4.2.4) is satisfied.

Conversely, assume that the convex hull condition is satisfied for some bottlenecks T_1, T_2 of parameters λ_1, λ_2, respectively, such that $q_2(X|T_2)$ is injective. Then, the sense $(ii) \Rightarrow (i)$ of Proposition D.1.8 shows that there exists a unique extension $\tilde{q}(X, T_1, T_2)$ of $q_1(X, T_1)$ and $q_2(X, T_2)$ such that we have $X - T_2 - T_1$. We then conclude with the Markov chain characterisation of successive refinement (Proposition 4.2.2).

This ends the proof of Proposition 4.2.6.

D.1.6 Linear Program Used to Compute the Convex Hull Condition (4.2.4)

Consider, for points $u, v_1, \dots, v_k \in \mathbb{R}^m$, the condition

$$u \in \mathsf{Hull}\{v_i,\ i = 1, \dots, k\}. \tag{D.1.24}$$

A linear program can be used to check whether this condition holds or not; in short, it consists of the first step of the simplex method (see, e.g., (Matousek et al., 2007), Section 5.6), which asserts the existence or not of an initial feasible basis, and computes this basis if it exists. More precisely, let us first note V the $m \times k$ matrix whose columns are the points v_i, and define

$$M := \begin{pmatrix} & V & \\ 1 & \dots & 1 \end{pmatrix}, \qquad \tilde{u} := \begin{pmatrix} u \\ 1 \end{pmatrix}.$$

Then, condition (D.1.24) can be reformulated as

$$\exists \alpha := (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^k \ : \quad \begin{cases} M\alpha = \tilde{u}, \\ \alpha_i \geq 0 \ \text{ for } \ i = 1, \dots, k. \end{cases} \tag{D.1.25}$$

We now consider the linear program defined for the augmented variable

$$\tilde{\alpha} := (\alpha_1, \dots, \alpha_k, \alpha_{k+1}, \dots, \alpha_{k+m+1}) \in \mathbb{R}^{k+m+1}$$

as

$$\min_{\substack{\tilde{M}\tilde{\alpha} = \tilde{u} \\ \forall i = 1, \dots, k+m+1,\ \ \alpha_i \geq 0}} \quad \alpha_{k+1} + \dots + \alpha_{k+m+1}, \tag{D.1.26}$$

where $\tilde{M} := (M | I_{m+1})$ is obtained by appending the $(m+1) \times (m+1)$ identity matrix to M to the right. It can be directly verified that (D.1.25), and thus, equivalently, (D.1.24), holds if and only if the minimum is 0 in the linear program (D.1.26), and that if this is the case, then the first k coordinates $\alpha_1, \dots, \alpha_k$ of any of the program's solutions provide coefficients for obtaining u as a convex combination of the v_i.

Now, consider two bottleneck distributions $q_1 := q_1(X, T_1)$ and $q_2 := q_2(X, T_2)$ such that $q(X|T_2)$ is injective. We want to check the convex hull condition (4.2.4), which holds if and only if for every $t_1 \in \mathcal{T}_1$, we have

$$q(X|t_1) \in \mathsf{Hull}\{q(X|t_2),\ t_2 \in \mathcal{T}_2\}. \tag{D.1.27}$$

This condition can be checked, for every fixed t_1, with the linear program described above, where if the condition holds, the algorithm also outputs a family of coefficients $(\alpha_{t_2,t_1})_{t_2}$ such that

$$q(X|t_1) = \sum_{t_2} \alpha_{t_2,t_1} q(X|t_2). \tag{D.1.28}$$

Let us define $q(t_2|t_1) := \alpha_{t_2,t_1}$ and a joint distribution $q(X, T_1, T_2)$ through

$$q(x, t_1, t_2) := q_1(t_1)q(t_2|t_1)q_2(x|t_2). \tag{D.1.29}$$

By construction, under q, we have the Markov chain $X - T_2 - T_1$. Moreover thanks to Equation (D.1.28) and the injectivity of $q(X|T_2)$, Proposition D.1.8 shows that q is indeed an extension of $q_1(X, T_1)$ and $q_2(X, T_2)$. Thus, the linear program above allows one both to check whether or not the convex hull condition holds and, when it does, to obtain Theorem 4.2.6's unique extension $q(X, T_1, T_2)$ such that $X - T_2 - T_1$.

See the published version of this work (Charvin et al., 2023a) for details on the algorithm's complexity.

Note also that as the convex hull condition holds if and only if the linear program's output is 0 for all $t_1 \in \mathcal{T}_1$, in numerical computations, the threshold for rounding the program's output impacts the answer. In our numerical experiments, we chose the threshold 10^{-6}.

D.1.7 Proof of Proposition 4.2.7

We will first present the framework developed in (Asoodeh et al., 2020; Witsenhausen et al., 1975) and then the original content of this proof, which starts with Lemma D.1.13 below. A full plan of this proof is presented in the main text.

We already noticed (in Section 4.2.2) that the primal IB problem (4.1.2) can be reformulated as an optimisation over the pairs $(q(T), q(X|T))$, i.e., Equation (4.2.3). Using the identity $I(U;V) = H(U) - H(U|V)$, and recalling that a bottleneck T must satisfy $I(X;T) = \lambda$ (Asoodeh et al., 2020), we can further reformulate the problem (4.2.3) as

$$\underset{\substack{(q(T),\, q(X|T)) \\ \sum_t q(t)q(X|t)=p(X) \\ H(X|T)=\nu}}{\arg\min} \quad H(Y|T), \tag{D.1.30}$$

where $\nu := H(X) - \lambda$. In particular, we can assume, without loss of generality, that $0 \leq \nu \leq H(X)$ (see Section 4.1.4), where $\nu = H(X)$ corresponds to $I(X;T) = 0$. Similarly as we denoted before by $I_Y(\lambda)$ the maximum in the classic IB problem (4.1.2), here, we denote by $H_Y(\nu)$ the minimum in (D.1.30). Rather than considering the information curve, i.e., the graph of I_Y, and following (Witsenhausen et al., 1975) upon which we rely, here, we consider the graph of H_Y, which we will refer to as the conditional entropy (CE) curve. This curve is convex (Witsenhausen et al., 1975), and it is just an affine translation of the information curve. Let us now define, for $\beta \geq 1$, the function

$$\begin{aligned} F_\beta \;:\; \Delta_{\mathcal{X}} &\to \mathbb{R} \\ p &\mapsto H(\kappa p) - \beta^{-1} H(p), \end{aligned}$$

where κ is the column transition matrix defined by the conditional probability $p(Y|X)$. Note that, for $p = p(X)$, we have $\kappa p = p(Y)$. (In this section, we choose notations close to those from (Asoodeh et al., 2020), as long as they do not clash with the ones we already established; most notably, what we denote here by β would correspond to β^{-1} in (Asoodeh et al., 2020).)

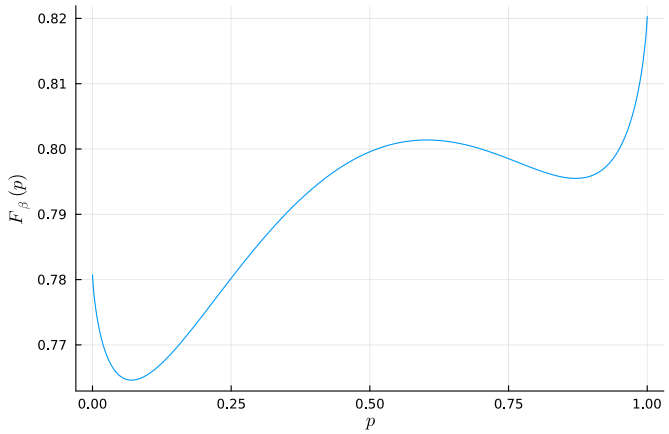


FIGURE D.1: The function F_β for example values of β and $p(X,Y)$, with binary X and Y. Here, p parameterises the source distribution $[p, 1-p]$.

The function F_β is plotted in Figure D.1 for example values of β and $p(X,Y)$, where the source and relevancy are binary. As a difference in concave functions, the function is a priori neither concave nor convex, but we can define its *lower convex envelope*, i.e., the largest convex function, which is still inferior or equal to F_β everywhere: we will denote it by $\mathcal{K}_\cup\big(F_\beta\big)$. In (Witsenhausen et al., 1975), through convex duality arguments, the following relationship between bottlenecks and F_β was proven:

Proposition D.1.10 ((Witsenhausen et al., 1975), Section IV)**.** *If a pair* $(q(T), q(X|T))$ *solves the problem* (D.1.30)*, then*

$$\sum_t q(t) F_\beta(q(X|t)) = \mathcal{K}_\cup\big(F_\beta\big)(p(X)), \tag{D.1.31}$$

for some $\beta \geq 1$ *such that* β^{-1} *is the slope of a tangent to the CE curve at the point* $(v, H_Y(v))$.

Let us also define the set of points where F_β differs from its lower convex envelope:

$$\mathcal{P}(\beta) := \big\{p \in \Delta_{\mathcal{X}} \ : \ F_\beta(p) \neq \mathcal{K}_\cup\big(F_\beta\big)(p)\big\}, \tag{D.1.32}$$

which will happen to be crucial for our considerations on successive refinement. Previous work showed that this set grows when β increases:

Lemma D.1.11 ((Asoodeh et al., 2020), Section II.B)**.** *If* $\beta_1 \leq \beta_2$*, then* $\mathcal{P}(\beta_1) \subseteq \mathcal{P}(\beta_2)$.

Proof. For the sake of self-containedness, we reproduce the computation from (Asoodeh et al., 2020). Let $p \notin \mathcal{P}(\beta_2)$, which means that $\mathcal{K}_\cup\Big(F_{\beta_2}\Big)(p) = F_{\beta_2}(p)$. For all $\beta_1 \leq \beta_2$,

$$\begin{aligned} F_{\beta_1}(p) &= H(\kappa p) - \beta_1^{-1} H(p) \\ &= F_{\beta_2}(p) - (\beta_1^{-1} - \beta_2^{-1}) H(p), \end{aligned}$$

so

$$\begin{aligned} \mathcal{K}_\cup\Big(F_{\beta_1}\Big)(p) &= \mathcal{K}_\cup\Big(F_{\beta_2} - (\beta_1^{-1} - \beta_2^{-1})H\Big)(p) \\ &\geq \mathcal{K}_\cup\Big(F_{\beta_2}\Big)(p) + \mathcal{K}_\cup\Big(-(\beta_1^{-1} - \beta_2^{-1})H\Big)(p) \\ &= \mathcal{K}_\cup\Big(F_{\beta_2}\Big)(p) - (\beta_1^{-1} - \beta_2^{-1})H(p), \end{aligned}$$

where the last equality comes from the convexity of the function $p \mapsto -(\beta_1^{-1} - \beta_2^{-1})H(p)$. Thus,

$$\begin{aligned}\mathcal{K}_\cup\Big(F_{\beta_1}\Big)(p) &\geq \mathcal{K}_\cup\Big(F_{\beta_1}\Big)(p) - (\beta_1^{-1} - \beta_2^{-1})H(p)\\ &= F_{\beta_2}(p) - (\beta_1^{-1} - \beta_2^{-1})H(p)\\ &= F_{\beta_1}(p).\end{aligned}$$

But, by definition, we have $\mathcal{K}_\cup\Big(F_{\beta_1}\Big)(p) \leq F_{\beta_1}(p)$, so $\mathcal{K}_\cup\Big(F_{\beta_1}\Big)(p) = F_{\beta_1}(p)$; in other words, $p \notin \mathcal{P}(\beta_1)$. Thus, we have proved that $\mathcal{P}(\beta_2)^c \subseteq \mathcal{P}(\beta_1)^c$, which is equivalent to $\mathcal{P}(\beta_1) \subseteq \mathcal{P}(\beta_2)$. □

Let us now assume that $|\mathcal{X}| = |\mathcal{Y}| = 2$. As we already proved successive refinability for deterministic $p(Y|X)$ in Proposition 4.2.5, we can assume that $p(Y|X)$ is not deterministic. But, the case of $|\mathcal{X}| = |\mathcal{Y}| = 2$ and non-deterministic $p(Y|X)$ is exhaustively studied in (Witsenhausen et al., 1975) (Section IV.A, IV.B and IV.D). The latter work implies that, in this case:

Lemma D.1.12 ((Witsenhausen et al., 1975))**.** *Let $0 \leq \nu < H(X)$, let $(q(T), q(X|T))$ be a a solution to (D.1.30) with parameter ν, and let β be given by Proposition D.1.10. Then, the set $\mathcal{P}(\beta)$ is a non-empty open interval and, for a pair $(q(T), q(X|T))$ to satisfy (D.1.31), the set of points*

$$\{q(X|t),\, t \in \mathcal{T}\}$$

must coincide with the extreme points of the interval $\mathcal{P}(\beta)$.

Equipped with these previously established facts, we can leverage them to prove successive refinement when $|\mathcal{X}| = |\mathcal{Y}| = 2$ and $p(Y|X)$ is not deterministic.

Lemma D.1.13. *Let $0 \leq \nu < H(X)$. Then, we can assume, without loss of generality, that $|\mathcal{T}| = 2$. Moreover, in this case, a solution $(q(T), q(X|T))$ to the reformulated IB problem (D.1.30) is such that $q(X|T)$, seen as a probability transition matrix, is injective.*

Proof. Let $(q(T), q(X|T))$ be a solution to (D.1.30) for parameter ν, and let β be given by Proposition D.1.10. From Lemma D.1.12, each $q(X|t)$ must correspond to one of the two extreme points of the interval $\mathcal{P}(\beta)$. Moreover, from Appendix B.4.3 in Chapter 2, for any primal bottleneck (or equivalently, any solution to (D.1.30)), we still obtain a bottleneck for the same parameter if we merge symbols t with identical $q(X|t)$. Thus, we can assume, without loss of generality, that $|\mathcal{T}| = 2$, and, in this case, the decoder $q(X|T)$ is, up to permutation of bottleneck symbols, uniquely defined by β.

Moreover, as $\mathcal{P}(\beta)$ is open and non-empty, these extreme points are distinct; in other words, the column transition matrix Q defined by $q(X|T)$ has its columns made of two distinct points on the simplex $\Delta_{\mathcal{X}}$. These points must thus be linearly independent as vectors in $\mathbb{R}^2$, so the rank of Q is 2. By the null rank theorem and as $|\mathcal{T}| = 2$, this implies that Q is injective. □

Let us now first consider SR for the case of $n = 2$ processing stages. Let $0 < \lambda_1 < \lambda_2 \leq H(X)$, and let T_1, T_2 be solutions to the primal IB problem (4.1.2) of respective parameters λ_1, λ_2. Equivalently, T_1 and T_2 are solutions to the reformulated IB problem (D.1.30) with resp. parameters ν_1, ν_2, where $0 \leq \nu_2 < \nu_1 < H(X)$. From Lemma D.1.13, we can assume that $q(X|T_2)$ is injective. Moreover, from Proposition D.1.10, the bottleneck pairs $(q(T_1), q(X|T_1))$ and $(q(T_2), q(X|T_2))$ are solutions to (D.1.31) for parameters β_1, β_2, respectively, which correspond to inverse slopes of the CE curve at $(\nu_1, H_Y(\nu_1))$ and $(\nu_2, H_Y(\nu_2))$,

respectively. By convexity of the CE curve (Witsenhausen et al., 1975), we have $\beta_1 \leq \beta_2$. Thus, from Lemma D.1.11,

$$\mathcal{P}(\beta_1) \subseteq \mathcal{P}(\beta_2).$$

This is equivalent to

$$\mathsf{Hull}\big(\overline{\mathcal{P}(\beta_1)}\big) = \overline{\mathcal{P}(\beta_1)} \subseteq \overline{\mathcal{P}(\beta_2)} = \mathsf{Hull}\big(\overline{\mathcal{P}(\beta_2)}\big),$$

where $\overline{E}$ denotes the closure of a set E, so, here, $\mathcal{P}(\beta_i)$ and $\overline{\mathcal{P}(\beta_i)}$ only differ by taking or not taking the segment's extreme points, and the equalities come from the convexity of this segment. From Lemma D.1.12, this can be rewritten as

$$\mathsf{Hull}\big\{q(X|t_1),\, t_1 \in \mathcal{T}_1\big\} \subseteq \mathsf{Hull}\big\{q(X|t_2),\, t_2 \in \mathcal{T}_2\big\}.$$

But this is exactly the convex hull condition (4.2.4). As we chose an injective $q(X|T_2)$, we can use the convex hull characterisation (Theorem 4.2.6) to conclude that T_1 and T_2 achieve successive refinement. This ends the proof of Proposition 4.2.7.

D.1.8 Computation of bifurcations values

In this work, we compute the bottlenecks' bifurcation parameters as the values where the effective cardinality changes (Zaslavsky et al., 2019): i.e., a bifurcation is a trade-off parameter value λ for which the number of distinct $q_\lambda(X|t)$ changes in a neighborhood of λ (see Section 4.1.4). With this naive method, the threshold chosen to numerically equate points $q(X|t)$ impacts the computed critical values, which could be avoided by using more sophisticated methods for computing these bifurcation values (Gedeon et al., 2012; Wu et al., 2020; Zaslavsky et al., 2019). However, the bifurcation values computed by our naive method did correspond, on our minimal examples, to parameters where the smoothness of the functions $I_X(\beta) := I_\beta(X;T)$ and $I_Y(\beta) := I_\beta(Y;T)$ breaks. Thus, our method seemingly identifies discontinuities of the first-order derivative of I_X and I_Y, which are those of second-order derivatives of the Lagrangian in (4.1.4) (see Corollary 1 in (Zaslavsky et al., 2019)). In this sense, our naive method still identifies the IB bifurcations, if defined as second-order bifurcations of the IB Lagragian as in, e.g., (Wu et al., 2020; Zaslavsky et al., 2019).

D.2 Section 4.3 Details

D.2.1 Proof of Proposition 4.3.3

We recall that Δ_{q_1,q_2} is the space of extensions $q(X,T_1,T_2)$ of $q_1(X,T_1)$ and $q_2(X,T_2)$, and that $\Delta_{SR,2}$ is the space of all distributions $r(X,T_1,T_2)$ (not necessarily consistent with q_1 and q_2) under which the Markov chain $X - T_2 - T_1$ holds. We write the proof for discrete variables for ease of presentation, but the very same proof works for continuous variables if we replace

sums by integrals. For $q(X,T_1,T_2) \in \Delta_{q_1,q_2}$ and $r(X,T_1,T_2) \in \Delta_{SR,2}$,

$$\begin{aligned} D_{KL}(q||r) &= \sum q(x,t_1,t_2)\log\left(\frac{q(x,t_1,t_2)}{r(x,t_1,t_2)}\right) \\ &= \sum q(x,t_1,t_2)\log\left(\frac{q(x,t_2)q(t_1|x,t_2)}{r(x,t_2)r(t_1|t_2)}\right) \\ &= \sum q(x,t_1,t_2)\log\left(\frac{q(t_1|x,t_2)}{r(t_1|t_2)}\right) + D_{KL}(q(X,T_2)||r(X,T_2)) \\ &\geq \sum q(x,t_1,t_2)\log\left(\frac{q(t_1|x,t_2)}{r(t_1|t_2)}\right) \\ &= \sum q(x,t_1,t_2)\log\left(\frac{q(t_1|x,t_2)}{q(t_1|t_2)}\right) + \sum q(t_2)D_{KL}\left(q(T_1|t_2)||r(T_1|t_2)\right) \\ &\geq \sum q(x,t_1,t_2)\log\left(\frac{q(t_1|x,t_2)}{q(t_1|t_2)}\right) \end{aligned} \tag{D.2.1}$$

The last term is $D_{KL}(q||r_0)$, with

$$r_0(X,T_1,T_2) := q(X)q(T_2|X)q(T_1|T_2) \in \Delta_{SR,2},$$

because, under r_0, the Markov chain $X - T_2 - T_1$ holds. So, from the last inequality in (D.2.1),

$$\inf_{r\in\Delta_{SR,2}} D_{KL}(q||r) = D_{KL}(q||r_0).$$

But, the last term of (D.2.1) is also $I_q(X;T_1|T_2)$. Thus,

$$\begin{aligned} D_{KL}(\Delta_{q_1,q_2}||\Delta_{SR}) &= \inf_{q\in\Delta_{q_1,q_2}} \inf_{r\in\Delta_{SR}} D_{KL}(q,r) \\ &= \inf_{q\in\Delta_{q_1,q_2}} D_{KL}(q||r_0) \\ &= \inf_{q\in\Delta_{q_1,q_2}} I_q(X;T_1|T_2) \\ &= UI(X : T_1 \setminus T_2). \end{aligned}$$

This ends the proof of Proposition 4.3.3.

D.3 Sample $p(Y|X)$ used in Sections 4.2.3 and 4.3.2

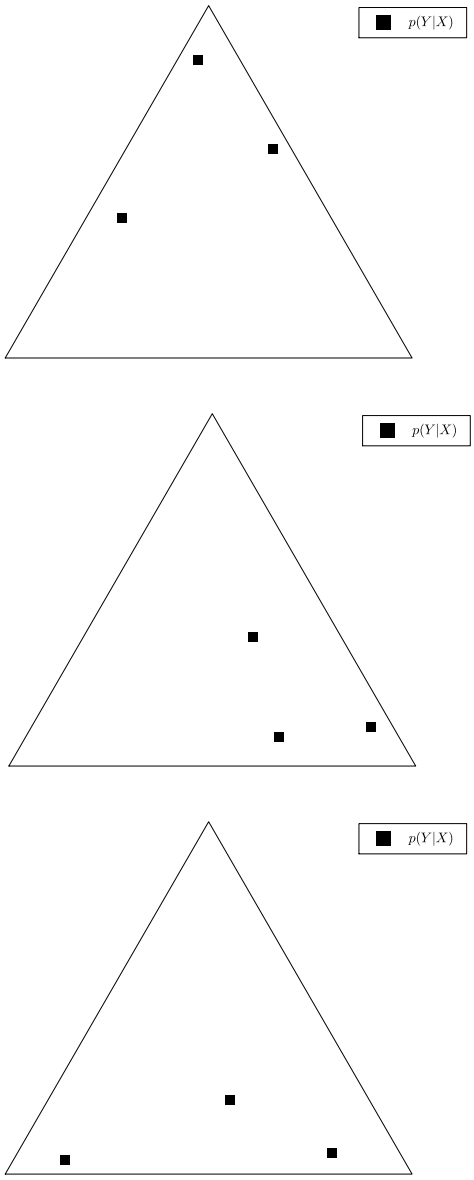


FIGURE D.2: Plot of the sample distributions $p(Y|X)$ used in, respectively, from top to bottom: (i) Figures 4.4a and 4.5a; (ii) Figures 4.4b and 4.5b; (iii) Figures 4.4c and 4.5c. The simplex depicted here is $\Delta_{\mathcal{Y}}$, where $|\mathcal{Y}| = 3$, and each black square corresponds to a symbol-wise conditional probability $p(Y|x) \in \Delta_{\mathcal{Y}}$. Note that the corresponding $p(X) \in \Delta_{\mathcal{X}}$ is shown in the left parts of Figures 4.4a–4.5c, which depict the simplex $\Delta_{\mathcal{X}}$, where, here, we also have $|\mathcal{X}| = 3$. The explicit values of the corresponding $p(X,Y)$ can be found at: `https://gitlab.com/uh-adapsys/successive-refinement-ib/` (accessed on 14 October 2025).

Bibliography

Achille, Alessandro et al. (Feb. 2018a). “Emergence of Invariance and Disentanglement in Deep Representations”. In: *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9. DOI: 10.1109/ITA.2018.8503149. (Visited on 04/28/2025).

— (Dec. 2018b). “Information Dropout: Learning Optimal Representations Through Noisy Computation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.12, pp. 2897–2905. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2017.2784440. (Visited on 10/24/2025).

Agmon, Shlomi et al. (July 2021). “Critical Slowing Down Near Topological Transitions in Rate-Distortion Problems”. In: *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 2625–2630. DOI: 10.1109/ISIT45174.2021.9517956. (Visited on 04/28/2025).

Aguerri, Inaki Estella et al. (Oct. 2017). *Distributed Information Bottleneck Method for Discrete and Gaussian Sources*. arXiv: 1709.09082 [cs, math]. (Visited on 07/11/2023).

Aguilera, Miguel et al. (Aug. 2013). “The Situated HKB Model: How Sensorimotor Spatial Coupling Can Alter Oscillatory Brain Dynamics”. In: *Frontiers in Computational Neuroscience* 7. ISSN: 1662-5188. DOI: 10.3389/fncom.2013.00117. (Visited on 10/18/2025).

Ahissar, Ehud et al. (May 2016). “Perception as a Closed-Loop Convergence Process”. In: *eLife* 5. Ed. by David Kleinfeld, e12830. ISSN: 2050-084X. DOI: 10.7554/eLife.12830.

Ahissar, Ehud et al. (Oct. 2025). “Closed-Loop Perception: Gaps between Artificial Intelligence and Biology”. In: *Current Opinion in Behavioral Sciences* 65, p. 101572. ISSN: 2352-1546. DOI: 10.1016/j.cobeha.2025.101572. (Visited on 11/02/2025).

Alemi, Alexander A. et al. (Feb. 2017). “Deep Variational Information Bottleneck”. In: *International Conference on Learning Representations*. (Visited on 04/29/2025).

Aliprantis, Charalambos D. et al. (2006). *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. Berlin/Heidelberg: Springer-Verlag. ISBN: 978-3-540-29586-0. DOI: 10.1007/3-540-29587-9. (Visited on 09/11/2025).

Amari, S.-I. (July 2001). “Information Geometry on Hierarchy of Probability Distributions”. In: *IEEE Transactions on Information Theory* 47.5, pp. 1701–1711. ISSN: 00189448. DOI: 10.1109/18.930911. (Visited on 07/12/2023).

Amir, Nadav et al. (Dec. 2015). “Past-Future Information Bottleneck for Linear Feedback Systems”. In: *2015 54th IEEE Conference on Decision and Control (CDC)*. Osaka: IEEE, pp. 5737–5742. ISBN: 978-1-4799-7886-1. DOI: 10.1109/CDC.2015.7403120. (Visited on 07/11/2023).

Appelle, S. (Oct. 1972). “Perception and Discrimination as a Function of Stimulus Orientation: The "Oblique Effect" in Man and Animals”. In: *Psychological Bulletin* 78.4, pp. 266–278. ISSN: 0033-2909. DOI: 10.1037/h0033117.

Apraez, Daniel Ordoñez et al. (Sept. 2025). “Morphological Symmetries in Robotics”. In: *The International Journal of Robotics Research* 44.10-11, pp. 1743–1766. ISSN: 0278-3649. DOI: 10.1177/02783649241282422. (Visited on 03/07/2026).

Arató, József et al. (May 2024). “Eye Movements Reflect Active Statistical Learning”. In: *Journal of Vision* 24.5, p. 17. ISSN: 1534-7362. DOI: 10.1167/jov.24.5.17. (Visited on 03/19/2026).

Archer, Karen et al. (Dec. 2022). "A Space of Goals: The Cognitive Geometry of Informationally Bounded Agents". In: *Royal Society Open Science* 9.12, p. 211800. DOI: 10.1098/rsos.211800. (Visited on 10/13/2025).

Ashman, Matthew et al. (Dec. 2024). "Approximately Equivariant Neural Processes". In: *Advances in Neural Information Processing Systems* 37, pp. 97088–97123. (Visited on 03/20/2025).

Asoodeh, Shahab et al. (Nov. 2020). "Bottleneck Problems: An Information and Estimation-Theoretic View". In: *Entropy* 22, p. 1325. DOI: 10.3390/e22111325.

Atay, Fatihcan M. et al. (Mar. 2017). "Lumpability of Linear Evolution Equations in Banach Spaces". In: *Evolution Equations and Control Theory* 6.1, pp. 15–34. DOI: 10.3934/eect.2017002. (Visited on 04/29/2026).

Aubret, Arthur et al. (Feb. 2023). "An Information-Theoretic Perspective on Intrinsic Motivation in Reinforcement Learning: A Survey". In: *Entropy* 25.2, p. 327. ISSN: 1099-4300. DOI: 10.3390/e25020327. (Visited on 12/01/2025).

Ay, Nihat (Apr. 2015). "Information Geometry on Complexity and Stochastic Interaction". In: *Entropy* 17.4, pp. 2432–2458. ISSN: 1099-4300. DOI: 10.3390/e17042432. (Visited on 07/12/2023).

Ay, Nihat et al. (Dec. 2003). "Dynamical Properties of Strongly Interacting Markov Chains". In: *Neural Networks* 16.10, pp. 1483–1497. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(03)00190-4. (Visited on 05/18/2024).

Ay, Nihat et al. (Sept. 2011). "A Geometric Approach to Complexity". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 21.3, p. 037103. ISSN: 1054-1500. DOI: 10.1063/1.3638446. (Visited on 05/18/2024).

Ay, Nihat et al. (Sept. 2012). "Information-Driven Self-Organization: The Dynamical System Approach to Autonomous Robot Behavior". In: *Theory in Biosciences* 131.3, pp. 161–179. ISSN: 1611-7530. DOI: 10.1007/s12064-011-0137-9.

Ay, Nihat et al. (2014). "On the Causal Structure of the Sensorimotor Loop". In: *Guided Self-Organization: Inception*. Ed. by Mikhail Prokopenko. Vol. 9. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 261–294. ISBN: 978-3-642-53733-2 978-3-642-53734-9. DOI: 10.1007/978-3-642-53734-9_9. (Visited on 07/11/2023).

Ay, Nihat et al. (Dec. 2015). "The Umwelt of an Embodied Agent-a Measure-Theoretic Definition". In: *Theory in Biosciences* 134. DOI: 10.1007/s12064-015-0217-3.

Ay, Nihat et al. (2017). *Information Geometry*. Vol. 64. Ergebnisse Der Mathematik Und Ihrer Grenzgebiete 34. Cham: Springer International Publishing. ISBN: 978-3-319-56477-7 978-3-319-56478-4. DOI: 10.1007/978-3-319-56478-4. (Visited on 11/12/2023).

Ay, Nihat et al. (2022). "Information and Complexity, Or: Where Is the Information?" In: *Complexity and Emergence*. Ed. by Sergio Albeverio et al. Cham: Springer International Publishing, pp. 87–105. ISBN: 978-3-030-95703-2. DOI: 10.1007/978-3-030-95703-2_4.

Ball, Robin C. et al. (June 2010). "Quantifying Emergence in Terms of Persistent Mutual Information". In: *Advances in Complex Systems* 13.03, pp. 327–338. ISSN: 0219-5259. DOI: 10.1142/S021952591000258X. (Visited on 11/06/2025).

Banerjee, Pradeep et al. (June 2018). "Computing the Unique Information". In: pp. 141–145. DOI: 10.1109/ISIT.2018.8437757.

Barandiaran, Xabier E. (Sept. 2017). "Autonomy and Enactivism: Towards a Theory of Sensorimotor Autonomous Agency". In: *Topoi* 36.3, pp. 409–430. ISSN: 1572-8749. DOI: 10.1007/s11245-016-9365-4. (Visited on 10/26/2025).

Barandiaran, Xabier E. et al. (Jan. 2014). "Norm-Establishing and Norm-Following in Autonomous Agency". In: *Artificial Life* 20.1, pp. 5–28. ISSN: 1064-5462. DOI: 10.1162/ARTL_a_00094. (Visited on 12/01/2025).

Barnett, Lionel et al. (Aug. 2021). *Dynamical Independence: Discovering Emergent Macroscopic Processes in Complex Dynamical Systems*. arXiv: 2106.06511 [nlin]. (Visited on 07/12/2023).

Barnett, Nix et al. (Oct. 2015). "Computational Mechanics of Input–Output Processes: Structured Transformations and the Epsilon-Transducer". In: *Journal of Statistical Physics* 161.2, pp. 404–451. ISSN: 1572-9613. DOI: 10.1007/s10955-015-1327-5. (Visited on 09/25/2025).

Barrett, Nathaniel F. et al. (Nov. 2025). "The Challenge of Normativity: An Examination of Three Minimal Models". In: *Adaptive Behavior*, p. 10597123251396761. ISSN: 1059-7123. DOI: 10.1177/10597123251396761. (Visited on 12/01/2025).

Baspinar, Emre et al. (Mar. 2021). "A Cortical-Inspired Sub-Riemannian Model for Poggendorff-Type Visual Illusions". In: *Journal of Imaging* 7.3, p. 41. ISSN: 2313-433X. DOI: 10.3390/jimaging7030041. (Visited on 03/11/2026).

Beer, Randall D. et al. (Jan. 2015). "Information Processing and Dynamics in Minimally Cognitive Agents". In: *Cognitive Science* 39.1, pp. 1–38. ISSN: 03640213. DOI: 10.1111/cogs.12142. (Visited on 09/27/2023).

Belfiore, Jean-Claude et al. (June 2022). *Topos and Stacks of Deep Neural Networks*. DOI: 10.48550/arXiv.2106.14587. arXiv: 2106.14587 [math]. (Visited on 11/25/2025).

Benedetto, Alessandro et al. (Apr. 2023). "Active Vision: How You Look Reflects What You Are Looking For". In: *Current Biology* 33.8, R303–R305. ISSN: 0960-9822. DOI: 10.1016/j.cub.2023.03.012. (Visited on 10/20/2025).

Benger, Etam et al. (May 2023). *The Cardinality Bound on the Information Bottleneck Representations Is Tight*. arXiv: 2305.07000 [cs, math]. (Visited on 07/11/2023).

Benigno, Gabriel B. et al. (June 2023). "Waves Traveling over a Map of Visual Space Can Ignite Short-Term Predictions of Sensory Input". In: *Nature Communications* 14.1, p. 3409. ISSN: 2041-1723. DOI: 10.1038/s41467-023-39076-2. (Visited on 03/12/2026).

Bennequin, Daniel et al. (July 2009). "Movement Timing and Invariance Arise from Several Geometries". In: *PLoS Computational Biology* 5.7, e1000426. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1000426. (Visited on 11/25/2025).

Bennequin, Daniel et al. (2017). "Several Geometries for Movements Generations". In: *Geometric and Numerical Foundations of Movements*. Ed. by Jean-Paul Laumond et al. Cham: Springer International Publishing, pp. 13–42. ISBN: 978-3-319-51547-2. DOI: 10.1007/978-3-319-51547-2_2. (Visited on 11/25/2025).

— (Sept. 2025). "Brain's Geometries for Movements and Beauty Judgments. A Contribution of Topos Geometries". In: *Frontiers in Psychology* 16. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2025.1583185. (Visited on 11/25/2025).

Berthoz, Alain et al. (2000). *The Brain's Sense of Movement*. Harvard University Press. ISBN: 978-0-674-80109-7. JSTOR: j.ctt1cc2mz2. (Visited on 03/03/2025).

Bertoni, Federico et al. (Nov. 2021). "Emergence of Lie Symmetries in Functional Architectures Learned by CNNs". In: *Frontiers in Computational Neuroscience* 15. ISSN: 1662-5188. DOI: 10.3389/fncom.2021.694505. (Visited on 05/21/2024).

Bertschinger, Nils et al. (Nov. 2013). "Quantifying Unique Information". In: *Entropy* 16. DOI: 10.3390/e16042161.

Bertschinger, Nils et al. (June 2014). "The Blackwell Relation Defines No Lattice". In: *2014 IEEE International Symposium on Information Theory*, pp. 2479–2483. DOI: 10.1109/ISIT.2014.6875280. arXiv: 1401.3146 [cs, math, stat]. (Visited on 07/12/2023).

Bialek, William et al. (July 2006). "Efficient Representation as a Design Principle for Neural Coding and Computation". In: *2006 IEEE International Symposium on Information Theory*. Seattle, WA: IEEE, pp. 659–663. ISBN: 978-1-4244-0505-3. DOI: 10.1109/ISIT.2006.261867. (Visited on 07/11/2023).

Biehl, Martin et al. (Sept. 2013). "Some Ways to See Two in One". In: *Proceedings of the 12th European Conference on the Synthesis and Simulation of Living Systems : Advances in Artificial Life, ECAL 2013*. ITA: MIT Press, pp. 1099–1106. ISBN: 978-0-262-31709-2. DOI: 10.7551/978-0-262-31709-2-ch165. (Visited on 10/20/2025).

Billingsley, Patrick (1965). *Ergodic Theory and Information*. John Wiley & Sons, Inc.

Blackwell, David (1953). "Equivalent Comparisons of Experiments". In: *Annals of Mathematical Statistics* 24, pp. 265–272.

Boi, Marco et al. (May 2017). "Consequences of the Oculomotor Cycle for the Dynamics of Perception". In: *Current Biology* 27.9, pp. 1268–1277. ISSN: 0960-9822. DOI: 10.1016/j.cub.2017.03.034. (Visited on 06/03/2024).

Bounoua, Mustapha et al. (Mar. 2025). *Learning to Match Unpaired Data with Minimum Entropy Coupling*. DOI: 10.48550/arXiv.2503.08501. arXiv: 2503.08501 [cs]. (Visited on 07/10/2025).

Bressloff, Paul C. et al. (Mar. 2001). "Geometric Visual Hallucinations, Euclidean Symmetry and the Functional Architecture of Striate Cortex". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356.1407, pp. 299–330. DOI: 10.1098/rstb.2000.0769. (Visited on 05/21/2024).

Brette, Romain (Jan. 2019). "Is Coding a Relevant Metaphor for the Brain?" In: *Behavioral and Brain Sciences* 42, e215. ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X19000049. (Visited on 10/25/2025).

Brezis, Haim (2011). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. New York, NY: Springer. ISBN: 978-0-387-70913-0 978-0-387-70914-7. DOI: 10.1007/978-0-387-70914-7. (Visited on 09/11/2025).

Buddha, S. Kartik et al. (2013). "Function Identification in Neuron Populations via Information Bottleneck". In: *Entropy* 15.5, pp. 1587–1608. ISSN: 1099-4300. DOI: 10.3390/e15051587.

Buesing, Lars et al. (Aug. 2010). "A Spiking Neuron as Information Bottleneck". In: *Neural Computation* 22.8, pp. 1961–1992. ISSN: 0899-7667. DOI: 10.1162/neco.2010.08-09-1084.

Buhrmann, Thomas et al. (2013). "A Dynamical Systems Account of Sensorimotor Contingencies". In: *Frontiers in Psychology* 4. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00285.

Buhrmann, Thomas et al. (2014). "Non-Representational Sensorimotor Knowledge". In: *From Animals to Animats 13*. Ed. by Angel P. del Pobil et al. Cham: Springer International Publishing, pp. 21–31. ISBN: 978-3-319-08864-8. DOI: 10.1007/978-3-319-08864-8_3.

Buzsáki, György et al. (May 2019). *The Brain from Inside Out*. Oxford, New York: Oxford University Press. ISBN: 978-0-19-090538-5.

Capdepuy, Philippe et al. (2007). "Constructing the Basic Umwelt of Artificial Agents: An Information-Theoretic Approach". In: *Advances in Artificial Life*. Ed. by Fernando Almeida E Costa et al. Vol. 4648. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 375–383. ISBN: 978-3-540-74912-7. DOI: 10.1007/978-3-540-74913-4_38. (Visited on 07/11/2023).

Caselles-Dupré, Hugo et al. (2019). "Symmetry-Based Disentangled Representation Learning Requires Interaction with Environments". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. (Visited on 10/29/2025).

Caselles-Dupré, Hugo et al. (Jan. 2021a). *On the Sensory Commutativity of Action Sequences for Embodied Agents*. DOI: 10.48550/arXiv.2002.05630. arXiv: 2002.05630 [cs]. (Visited on 12/12/2024).

— (July 2021b). *SCOD: Active Object Detection for Embodied Agents Using Sensory Commutativity of Action Sequences*. DOI: 10.48550/arXiv.2107.02069. arXiv: 2107.02069 [cs]. (Visited on 12/12/2024).

Casile, Antonino et al. (Jan. 2019). "Contrast Sensitivity Reveals an Oculomotor Strategy for Temporally Encoding Space". In: *eLife* 8. Ed. by Fred Rieke et al., e40924. ISSN: 2050-084X. DOI: 10.7554/eLife.40924. (Visited on 10/26/2025).

Catenacci Volpi, Nicola et al. (Oct. 2020). "Space Emerges from What We Know-Spatial Categorisations Induced by Information Constraints". In: *Entropy* 20, p. 1179. DOI: 10.3390/e22101179.

Chalk, Matthew et al. (2018). "Toward a Unified Theory of Efficient, Predictive, and Sparse Coding". In: *Proceedings of the National Academy of Sciences* 115.1, pp. 186–191. DOI: 10.1073/pnas.1711114115.

Charvin, Hippolyte et al. (Nov. 2022). "Successive Refinement and Coarsening of the Information Bottleneck". In: *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*. (Visited on 06/06/2024).

Charvin, Hippolyte et al. (Sept. 2023a). "Exact and Soft Successive Refinement of the Information Bottleneck". In: *Entropy* 25.9, p. 1355. ISSN: 1099-4300. DOI: 10.3390/e25091355. (Visited on 06/06/2024).

Charvin, Hippolyte et al. (Nov. 2023b). "Towards Information Theory-Based Discovery of Equivariances". In: *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*. (Visited on 04/25/2025).

— (Feb. 2025). "An Information Parsimony Perspective on Probabilistic Symmetries". In: *NeurIPS 2024 Workshop on Symmetry and Geometry in Neural Representations*. (Visited on 11/08/2025).

Chechik, Gal et al. (Jan. 2005). "Information Bottleneck for Gaussian Variables". In: *The Journal of Machine Learning Research* 6, pp. 165–188.

Chirikjian, Gregory S. (2012). *Stochastic Models, Information Theory, and Lie Groups, Volume 2: Analytic Methods and Modern Applications*. Applied and Numerical Harmonic Analysis. Boston: Birkhäuser. ISBN: 978-0-8176-4943-2 978-0-8176-4944-9. DOI: 10.1007/978-0-8176-4944-9. (Visited on 10/27/2025).

Chklovskii, Dmitri B. et al. (Apr. 2002). "Wiring Optimization in Cortical Circuits". In: *Neuron* 34.3, pp. 341–347. ISSN: 0896-6273. DOI: 10.1016/S0896-6273(02)00679-7. (Visited on 03/13/2026).

Chong, Isis et al. (Jan. 2020). "On the Evolution of a Radical Concept: Affordances According to Gibson and Their Subsequent Use and Development". In: *Perspectives on Psychological Science* 15.1, pp. 117–132. ISSN: 1745-6916. DOI: 10.1177/1745691619868207. (Visited on 04/10/2026).

Cicalese, Ferdinando et al. (June 2019). "Minimum-Entropy Couplings and Their Applications". In: *IEEE Transactions on Information Theory* 65.6, pp. 3436–3451. ISSN: 1557-9654. DOI: 10.1109/TIT.2019.2894519. (Visited on 07/10/2025).

Cisek, Paul et al. (June 2024). "Toward a Neuroscience of Natural Behavior". In: *Current Opinion in Neurobiology* 86, p. 102859. ISSN: 0959-4388. DOI: 10.1016/j.conb.2024.102859. (Visited on 10/30/2025).

Clark, Andy (Jan. 2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press. ISBN: 978-0-19-021701-3. DOI: 10.1093/acprof:oso/9780190217013.001.0001.

Clark, Ashley M. et al. (Dec. 2022). "Eye Drift during Fixation Predicts Visual Acuity". In: *Proceedings of the National Academy of Sciences* 119.49, e2200256119. DOI: 10.1073/pnas.2200256119. (Visited on 03/19/2026).

Clark, David et al. (2019). "Unsupervised Discovery of Temporal Structure in Noisy Data with Dynamical Components Analysis". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. (Visited on 11/25/2025).

Cloney, Richard (Nov. 1982). "Ascidian Larvae and the Events of Metamorphosis1". In: *American Zoologist* 22.4, pp. 817–826. ISSN: 0003-1569. DOI: 10.1093/icb/22.4.817. (Visited on 11/30/2025).

Cohn, Donald L. (2013). *Measure Theory: Second Edition*. Birkhäuser Advanced Texts Basler Lehrbücher. New York, NY: Springer. ISBN: 978-1-4614-6955-1 978-1-4614-6956-8. DOI: 10.1007/978-1-4614-6956-8. (Visited on 07/30/2025).

Coombes, Stephen et al., eds. (2014). *Neural Fields: Theory and Applications*. Berlin, Heidelberg: Springer. ISBN: 978-3-642-54592-4 978-3-642-54593-1. DOI: 10.1007/978-3-642-54593-1. (Visited on 03/11/2026).

Coudène, Yves (2016). *Ergodic Theory and Dynamical Systems*. Universitext. London: Springer. ISBN: 978-1-4471-7285-7 978-1-4471-7287-1. DOI: 10.1007/978-1-4471-7287-1. (Visited on 06/17/2025).

Cover, Thomas et al. (2009). *Elements of Information Theory*. 2nd edition. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience.

Creutzig, Felix et al. (May 2009). "Past-Future Information Bottleneck in Dynamical Systems". In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 79, p. 041925. DOI: 10.1103/PhysRevE.79.041925.

Crutchfield, James P. (Oct. 2017). *The Origins of Computational Mechanics: A Brief Intellectual History and Several Clarifications*. DOI: 10.48550/arXiv.1710.06832. arXiv: 1710.06832 [cond-mat]. (Visited on 10/24/2025).

Csiszár, Imre et al. (2011). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. 2nd ed. Cambridge: Cambridge University Press. ISBN: 978-0-521-19681-9. DOI: 10.1017/CBO9780511921889. (Visited on 09/27/2023).

de la Rue, Thierry (2006). "An Introduction to Joinings in Ergodic Theory". In: *Discrete and Continuous Dynamical Systems* 15.1, pp. 121–142. ISSN: 1078-0947. DOI: 10.3934/dcds.2006.15.121. (Visited on 07/14/2025).

— (2023). "Joinings in Ergodic Theory". In: *Ergodic Theory*. Springer, New York, NY, pp. 149–168. ISBN: 978-1-0716-2388-6. DOI: 10.1007/978-1-0716-2388-6_300. (Visited on 11/28/2025).

De Llanza Varona, Miguel et al. (2024). "Exploring Action-Centric Representations Through the Lens of Rate-Distortion Theory". In: *Active Inference*. Ed. by Christopher L. Buckley et al. Cham: Springer Nature Switzerland, pp. 189–203. ISBN: 978-3-031-47958-8. DOI: 10.1007/978-3-031-47958-8_12.

Dean, Alexander et al. (July 2025). *Algebras of Actions in an Agent's Representations of the World*. DOI: 10.48550/arXiv.2310.01536. arXiv: 2310.01536 [cs]. (Visited on 10/29/2025).

Degenaar, Jan et al. (Sept. 2017). "Sensorimotor Theory and Enactivism". In: *Topoi* 36.3, pp. 393–407. ISSN: 1572-8749. DOI: 10.1007/s11245-015-9338-z. (Visited on 03/18/2026).

Deng, Weijian et al. (Dec. 2022). "On the Strong Correlation Between Model Invariance and Generalization". In: *Advances in Neural Information Processing Systems* 35, pp. 28052–28067. (Visited on 10/25/2025).

Di Paolo, Ezequiel et al. (May 2017). *Sensorimotor Life: An Enactive Proposal*. Oxford University Press. ISBN: 978-0-19-878684-9. DOI: 10.1093/acprof:oso/9780198786849.001.0001.

Di Paolo, Ezequiel Alejandro et al. (July 2014). "Learning to Perceive in the Sensorimotor Approach: Piaget's Theory of Equilibration Interpreted Dynamically". In: *Frontiers in Human Neuroscience* 8. ISSN: 1662-5161. DOI: 10.3389/fnhum.2014.00551. (Visited on 10/26/2025).

Díaz Ledezma, Fernando et al. (Dec. 2023). "Machine Learning–Driven Self-Discovery of the Robot Body Morphology". In: *Science Robotics* 8.85, eadh0972. DOI: 10.1126/scirobotics.adh0972. (Visited on 10/28/2025).

Dikshtein, Michael et al. (Oct. 2021). *A Class of Nonbinary Symmetric Information Bottleneck Problems*. arXiv: 2110.00985 [cs, math]. (Visited on 07/11/2023).

Dorrell, Will et al. (Sept. 2022). "Actionable Neural Representations: Grid Cells from Minimal Constraints". In: *The Eleventh International Conference on Learning Representations*. (Visited on 10/27/2025).

Dragoi, Valentin et al. (Dec. 2001). "Stability of Cortical Responses and the Statistics of Natural Scenes". In: *Neuron* 32.6, pp. 1181–1192. ISSN: 0896-6273. DOI: 10.1016/S0896-6273(01)00540-2. (Visited on 03/10/2026).

Dudley, R. (1966). "Convergence of Baire Measures". In: *Studia Mathematica* 27.3, pp. 251–268. ISSN: 0039-3223. (Visited on 09/11/2025).

Eagleman, David M. et al. (Jan. 2023). "The Future of Sensory Substitution, Addition, and Expansion via Haptic Devices". In: *Frontiers in Human Neuroscience* 16. ISSN: 1662-5161. DOI: 10.3389/fnhum.2022.1055546. (Visited on 10/26/2025).

Egbert, Matthew D. et al. (Aug. 2014). "Modeling Habits as Self-Sustaining Patterns of Sensorimotor Behavior". In: *Frontiers in Human Neuroscience* 8. ISSN: 1662-5161. DOI: 10.3389/fnhum.2014.00590. (Visited on 10/29/2025).

— (Dec. 2022). "Using Enactive Robotics to Think Outside of the Problem-Solving Box: How Sensorimotor Contingencies Constrain the Forms of Emergent Autononomous Habits". In: *Frontiers in Neurorobotics* 16. ISSN: 1662-5218. DOI: 10.3389/fnbot.2022.847054. (Visited on 10/23/2025).

Einsiedler, Manfred et al. (2011). *Ergodic Theory: With a View towards Number Theory*. London: Springer. ISBN: 978-0-85729-020-5 978-0-85729-021-2. DOI: 10.1007/978-0-85729-021-2. (Visited on 11/13/2025).

Elad, Adar et al. (Oct. 2019). "Direct Validation of the Information Bottleneck Principle for Deep Nets". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Seoul, Korea (South): IEEE, pp. 758–762. ISBN: 978-1-7281-5023-9. DOI: 10.1109/ICCVW.2019.00099. (Visited on 09/27/2023).

Emmert-Streib, Frank et al., eds. (2009). *Information Theory and Statistical Learning*. Boston, MA: Springer US. ISBN: 978-0-387-84815-0 978-0-387-84816-7. DOI: 10.1007/978-0-387-84816-7. (Visited on 04/25/2026).

Equitz, W.H.R. et al. (1991). "Successive Refinement of Information". In: *IEEE Transactions on Information Theory* 37.2, pp. 269–275. DOI: 10.1109/18.75242.

Favela, Luis (Jan. 2024). *The Ecological Brain: Unifying the Sciences of Brain, Body, and Environment*. ISBN: 978-1-003-00995-5. DOI: 10.4324/9781003009955.

Favela, Luis H. et al. (June 2023). "Investigating the Concept of Representation in the Neural and Psychological Sciences". In: *Frontiers in Psychology* 14. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2023.1165622. (Visited on 03/17/2026).

Field, D. J. (Dec. 1987). "Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells". In: *Journal of the Optical Society of America. A, Optics and Image Science* 4.12, pp. 2379–2394. ISSN: 0740-3232. DOI: 10.1364/josaa.4.002379.

Forestier, Sébastien et al. (2022). "Intrinsically Motivated Goal Exploration Processes with Automatic Curriculum Learning". In: *Journal of Machine Learning Research* 23.152, pp. 1–41. ISSN: 1533-7928. (Visited on 12/01/2025).

Friston, Karl J. et al. (Nov. 2024). "Supervised Structure Learning". In: *Biological Psychology* 193, p. 108891. ISSN: 0301-0511. DOI: 10.1016/j.biopsycho.2024.108891. (Visited on 10/11/2025).

Fritz, Tobias (Aug. 2020). "A Synthetic Approach to Markov Kernels, Conditional Independence and Theorems on Sufficient Statistics". In: *Advances in Mathematics* 370, p. 107239. ISSN: 0001-8708. DOI: 10.1016/j.aim.2020.107239. (Visited on 09/26/2025).

Fritz, Tobias et al. (Apr. 2025). *Categories of Abstract and Noncommutative Measurable Spaces*. DOI: 10.48550/arXiv.2504.13708. arXiv: 2504.13708 [math]. (Visited on 09/24/2025).

Gedeon, Tomáš et al. (2012). "The Mathematical Structure of Information Bottleneck Methods". In: *Entropy* 14.3, pp. 456–479. ISSN: 1099-4300. DOI: 10.3390/e14030456.

Gerken, Jan E. et al. (Dec. 2023). "Geometric Deep Learning and Equivariant Neural Networks". In: *Artificial Intelligence Review* 56.12, pp. 14605–14662. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10502-7. (Visited on 05/21/2024).

Gibson, J.J. (2014). *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press. DOI: 10.4324/9781315740218.

Gilad-Bachrach, Ran et al. (2003). "An Information Theoretic Tradeoff between Complexity and Accuracy". In: *Learning Theory and Kernel Machines*. Ed. by Gerhard Goos et al. Vol. 2777. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 595–609. ISBN: 978-3-540-40720-1 978-3-540-45167-9. DOI: 10.1007/978-3-540-45167-9_43. (Visited on 07/11/2023).

Girshick, Ahna R. et al. (July 2011). "Cardinal Rules: Visual Orientation Perception Reflects Knowledge of Environmental Statistics". In: *Nature Neuroscience* 14.7, pp. 926–932. ISSN: 1546-1726. DOI: 10.1038/nn.2831. (Visited on 03/10/2026).

Glasner, Eli (Feb. 2003). *Ergodic Theory via Joinings*. https://www.ams.org/surv/101. DOI: 10.1090/surv/101. (Visited on 07/14/2025).

Globerson, Amir et al. (2003). "Sufficient Dimensionality Reduction". In: *Journal of Machine Learning Research* 3.Mar, pp. 1307–1331. ISSN: ISSN 1533-7928. (Visited on 11/19/2025).

Goasguen, Loïc et al. (Dec. 2023). "From State Transitions to Sensory Regularity: Structuring Uninterpreted Sensory Signals From Naive Sensorimotor Experiences". In: *IEEE Transactions on Cognitive and Developmental Systems* 15.4, pp. 1864–1878. ISSN: 2379-8939. DOI: 10.1109/TCDS.2022.3226531. (Visited on 10/28/2025).

Godon, Jean-Merwan et al. (2020). "A Formal Account of Structuring Motor Actions With Sensory Prediction for a Naive Agent". In: *Frontiers in Robotics and AI* 7. ISSN: 2296-9144. DOI: 10.3389/frobt.2020.561660.

Goyal, Anirudh et al. (2019). "Transfer and Exploration via the Information Bottleneck". In: *International Conference on Learning Representations*.

Grassberger, Peter (Sept. 1986). "Toward a Quantitative Theory of Self-Generated Complexity". In: *International Journal of Theoretical Physics* 25.9, pp. 907–938. ISSN: 1572-9575. DOI: 10.1007/BF00668821. (Visited on 11/06/2025).

Gray, Robert M. (2009). *Probability, Random Processes, and Ergodic Properties*. Boston, MA: Springer US. ISBN: 978-1-4419-1089-9 978-1-4419-1090-5. DOI: 10.1007/978-1-4419-1090-5. (Visited on 06/17/2025).

— (2011). *Entropy and Information Theory*. Boston, MA: Springer US. ISBN: 978-1-4419-7969-8 978-1-4419-7970-4. DOI: 10.1007/978-1-4419-7970-4. (Visited on 06/17/2025).

Greschonig, Gernot et al. (2000). "Ergodic Decomposition of Quasi-Invariant Probability Measures". In: *Colloquium Mathematicae* 84/85.2, pp. 495–514. ISSN: 0010-1354. (Visited on 11/20/2025).

Griffith, Virgil et al. (2014). "Quantifying Synergistic Mutual Information". In: *Guided Self-Organization: Inception*. Ed. by Mikhail Prokopenko. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 159–190. ISBN: 978-3-642-53734-9. DOI: 10.1007/978-3-642-53734-9_6.

Guerreiro, Marinês (Dec. 2016). "Group Algebras and Coding Theory". In: *São Paulo Journal of Mathematical Sciences* 10.2, pp. 346–371. ISSN: 2316-9028. DOI: 10.1007/s40863-016-0040-x. (Visited on 10/27/2025).

Hafed, Ziad M. et al. (Apr. 2021). "Active Vision at the Foveal Scale in the Primate Superior Colliculus". In: *Journal of Neurophysiology* 125.4, pp. 1121–1138. ISSN: 0022-3077. DOI: 10.1152/jn.00724.2020. (Visited on 03/19/2026).

Harder, Malte et al. (Jan. 2013). "Bivariate Measure of Redundant Information". In: *Phys. Rev. E* 87.1, p. 012130. DOI: 10.1103/PhysRevE.87.012130.

Higgins, Irina et al. (Dec. 2018). *Towards a Definition of Disentangled Representations*. arXiv: 1812.02230 [cs, stat]. (Visited on 07/12/2023).

Higgins, Irina et al. (2022). "Symmetry-Based Representations for Artificial and Biological General Intelligence". In: *Frontiers in Computational Neuroscience* 16. ISSN: 1662-5188. (Visited on 10/06/2023).

Hoffmann, Matej et al. (Sept. 2017). "Development of Reaching to the Body in Early Infancy: From Experiments to Robotic Models". In: *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 112–119. DOI: 10.1109/DEVLRN.2017.8329795. (Visited on 10/25/2025).

Hsu, Hsiang et al. (Nov. 2018). *Generalizing Bottleneck Problems*. arXiv: 1802.05861 [cs, math]. (Visited on 07/11/2023).

Hu, Shizhe et al. (Aug. 2024). "A Survey on Information Bottleneck". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.8, pp. 5325–5344. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2024.3366349. (Visited on 11/09/2025).

Husemoller, Dale (1994). *Fibre Bundles*. Vol. 20. Graduate Texts in Mathematics. New York, NY: Springer. ISBN: 978-1-4757-2263-5 978-1-4757-2261-1. DOI: 10.1007/978-1-4757-2261-1. (Visited on 11/23/2025).

Intoy, Janis et al. (Feb. 2020). "Finely Tuned Eye Movements Enhance Visual Acuity". In: *Nature Communications* 11.1, p. 795. ISSN: 2041-1723. DOI: 10.1038/s41467-020-14616-2. (Visited on 10/26/2025).

Intoy, Janis et al. (July 2024). "Consequences of Eye Movements for Spatial Selectivity". In: *Current Biology* 34.14, 3265–3272.e4. ISSN: 0960-9822. DOI: 10.1016/j.cub.2024.06.016. (Visited on 10/26/2025).

Jacobs, Mozes et al. (June 2025). *Traveling Waves Integrate Spatial Information Through Time*. DOI: 10.48550/arXiv.2502.06034. arXiv: 2502.06034 [cs]. (Visited on 03/10/2026).

Jacquey, Lisa et al. (Dec. 2019). "Sensorimotor Contingencies as a Key Drive of Development: From Babies to Robots". In: *Frontiers in Neurorobotics* 13. ISSN: 1662-5218. DOI: 10.3389/fnbot.2019.00098. (Visited on 10/26/2025).

Jamneshan, Asgar et al. (2023). "Foundational aspects of uncountable measure theory: Gelfand duality, Riesz representation, canonical models, and canonical disintegration". In: *Fundamenta Mathematicae* 261, pp. 1–98. ISSN: 0016-2736, 1730-6329. DOI: 10.4064/fm226-7-2022. (Visited on 09/24/2025).

Jamone, Lorenzo et al. (Mar. 2018). "Affordances in Psychology, Neuroscience, and Robotics: A Survey". In: *IEEE Transactions on Cognitive and Developmental Systems* 10.1, pp. 4–25. ISSN: 2379-8920, 2379-8939. DOI: 10.1109/TCDS.2016.2594134. (Visited on 07/12/2023).

Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4, pp. 620–630. DOI: 10.1103/PhysRev.106.620.

Jirsa, Viktor et al. (Feb. 2022). "Entropy, Free Energy, Symmetry and Dynamics in the Brain". In: *Journal of Physics: Complexity* 3.1, p. 015007. ISSN: 2632-072X. DOI: 10.1088/2632-072X/ac4bec. (Visited on 06/06/2024).

Jost, Jürgen (2016). "Sensorimotor Contingencies and the Dynamical Creation of Structural Relations Underlying Percepts". In: *The Pragmatic Turn : Toward Action-Oriented Views in Cognitive Science*. MIT Press, pp. 121–138.

Kawaguchi, Kenji et al. (May 2023). *How Does Information Bottleneck Help Deep Learning?* arXiv: 2305.18887 [cs, math]. (Visited on 07/11/2023).

Kechris, Alexander S. (1995). *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. New York, NY: Springer. ISBN: 978-1-4612-8692-9 978-1-4612-4190-4. DOI: 10.1007/978-1-4612-4190-4. (Visited on 07/19/2025).

Keller, T. Anderson (Dec. 2025). *Flow Equivariant Recurrent Neural Networks*. DOI: 10.48550/arXiv.2507.14793. arXiv: 2507.14793 [cs]. (Visited on 03/12/2026).

Keller, T. Anderson et al. (July 2023a). "Neural Wave Machines: Learning Spatiotemporally Structured Representations with Locally Coupled Oscillatory Recurrent Neural Networks". In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, pp. 16168–16189. (Visited on 06/04/2024).

Keller, T. Anderson et al. (Oct. 2023b). "Traveling Waves Encode The Recent Past and Enhance Sequence Learning". In: *The Twelfth International Conference on Learning Representations*. (Visited on 03/10/2026).

Keller, T. Anderson et al. (Feb. 2026). *A Spatiotemporal Perspective on Dynamical Computation in Neural Information Processing Systems*. DOI: 10.48550/arXiv.2409.13669. arXiv: 2409.13669 [q-bio]. (Visited on 03/05/2026).

Kerr, David et al. (2016). *Ergodic Theory*. Springer Monographs in Mathematics. Cham: Springer International Publishing. ISBN: 978-3-319-49845-4 978-3-319-49847-8. DOI: 10.1007/978-3-319-49847-8. (Visited on 10/17/2025).

Keurti, Hamza et al. (July 2023). "Homomorphism AutoEncoder – Learning Group Structured Representations from Observed Transitions". In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, pp. 16190–16215. (Visited on 10/19/2025).

Keurti, Hamza et al. (June 2024). "Stitching Manifolds: Leveraging Interaction to Compose Object Representations into Scenes." In: *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*. (Visited on 12/12/2024).

Klampfl, Stefan et al. (Apr. 2009). "Spiking Neurons Can Learn to Solve Information Bottleneck Problems and Extract Independent Components". In: *Neural Computation* 21.4, pp. 911–959. ISSN: 0899-7667. DOI: 10.1162/neco.2008.01-07-432.

Kleinman, Michael et al. (July 2023). *A Cortical Information Bottleneck during Decision-Making*. Preprint. Neuroscience. DOI: 10.1101/2023.07.12.548742. (Visited on 09/27/2023).

Kline, Adam (Mar. 2025). "Principles for Coarse-Graining in Biological Systems". PhD thesis. University of Chicago. DOI: 10.6082/uchicago.14378. (Visited on 11/29/2025).

Kline, Adam G. et al. (Mar. 2022). "Gaussian Information Bottleneck and the Non-Perturbative Renormalization Group". In: *New Journal of Physics* 24.3, p. 033007. DOI: 10.1088/1367-2630/ac395d.

Klyubin, A.S. et al. (2004). "Organization of the Information Flow in the Perception-Action Loop of Evolved Agents". In: *Proceedings. 2004 NASA/DoD Conference on Evolvable Hardware, 2004.* Pp. 177–180. DOI: 10.1109/EH.2004.1310828.

Kolchinsky, Artemy et al. (May 2017). "Nonlinear Information Bottleneck". In: *Entropy* 21. DOI: 10.3390/e21121181.

Kolchinsky, Artemy et al. (Dec. 2018). "Semantic Information, Autonomous Agency and Non-Equilibrium Statistical Physics". In: *Interface Focus* 8.6, p. 20180041. ISSN: 2042-8898, 2042-8901. DOI: 10.1098/rsfs.2018.0041. (Visited on 07/12/2023).

Kolchinsky, Artemy et al. (Feb. 2019). *Caveats for Information Bottleneck in Deterministic Scenarios*. arXiv: 1808.07593 [cs, stat]. (Visited on 07/11/2023).

König, Heinz (2012). "Stochastic Processes in Terms of Inner Premeasures". In: *Measure and Integration: Publications 1997-2011*. Ed. by Heinz König. Basel: Springer, pp. 313–342. ISBN: 978-3-0348-0382-3. DOI: 10.1007/978-3-0348-0382-3_14. (Visited on 09/23/2025).

Koshelev, V.N. (1980). "Hierarchical Coding of Discrete Sources". In: *Probl. Peredachi Inf.* 16.3, pp. 31–49.

Kostina, Victoria et al. (Nov. 2018). *Successive Refinement of Abstract Sources*. arXiv: 1707.09567 [cs, math]. (Visited on 07/12/2023).

Krakauer, David et al. (June 2020). "The Information Theory of Individuality". In: *Theory in Biosciences* 139. DOI: 10.1007/s12064-020-00313-7.

Laflaquière, Alban (Oct. 2020). *Emergence of Spatial Coordinates via Exploration*. arXiv: 2010.15469 [cs]. (Visited on 07/12/2023).

Laflaquière, Alban et al. (Aug. 2015a). "Grounding Object Perception in a Naive Agent's Sensorimotor Experience". In: *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 276–282. DOI: 10.1109/DEVLRN.2015.7346156. (Visited on 03/17/2026).

Laflaquière, Alban et al. (Sept. 2015b). "Learning Agent's Spatial Configuration from Sensorimotor Invariants". In: *Robotics and Autonomous Systems* 71, pp. 49–59. ISSN: 09218890. DOI: 10.1016/j.robot.2015.01.003. arXiv: 1810.01872 [cs, stat]. (Visited on 07/12/2023).

Laflaquière, Alban et al. (Oct. 2018). *Grounding Perception: A Developmental Approach to Sensorimotor Contingencies*. arXiv: 1810.01870 [cs, stat]. (Visited on 07/12/2023).

Laflaquière, Alban et al. (2019). "Unsupervised Emergence of Egocentric Spatial Structure from Sensorimotor Prediction". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. (Visited on 03/17/2026).

Lamb, Alex et al. (Dec. 2022). *Guaranteed Discovery of Control-Endogenous Latent States with Multi-Step Inverse Models*. arXiv: 2207.08229 [cs, stat]. (Visited on 07/12/2023).

Langer, Carlotta et al. (2021). "How Morphological Computation Shapes Integrated Information in Embodied Agents". In: *Frontiers in Psychology* 12. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.716433.

— (Dec. 2024). "An Information-Theoretic Perspective on Acting Agents". In: *IOP Conference Series: Materials Science and Engineering* 1321.1, p. 012008. ISSN: 1757-899X. DOI: 10.1088/1757-899X/1321/1/012008. (Visited on 11/09/2025).

Langlois, Eya Torkhani et al. (Apr. 2024). "Role of the Cerebellum in the Construction of Functional and Geometrical Spaces". In: *The Cerebellum*. ISSN: 1473-4230. DOI: 10.1007/s12311-024-01693-y. (Visited on 06/19/2024).

Lastras, L. et al. (2001). "All Sources Are Nearly Successively Refinable". In: *IEEE Transactions on Information Theory* 47.3, pp. 918–926. DOI: 10.1109/18.915645.

Ledezma, Fernando Diaz et al. (June 2025). *Unsupervised Discovery of Behavioral Primitives from Sensorimotor Dynamic Functional Connectivity*. DOI: 10.48550/arXiv.2506.22473. arXiv: 2506.22473 [cs]. (Visited on 10/23/2025).

Lemaréchal, Claude (2001). "Lagrangian Relaxation". In: *Computational Combinatorial Optimization: Optimal or Provably Near-Optimal Solutions*. Ed. by Michael Jünger et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 112–156. ISBN: 978-3-540-45586-8. DOI: 10.1007/3-540-45586-8_4.

Li, Cheuk Ting (Aug. 2021). "Efficient Approximate Minimum Entropy Coupling of Multiple Probability Distributions". In: *IEEE Transactions on Information Theory* 67.8, pp. 5259–5268. ISSN: 1557-9654. DOI: 10.1109/TIT.2021.3076986. (Visited on 07/10/2025).

Liboni, Luisa H. B. et al. (Jan. 2025). "Image Segmentation with Traveling Waves in an Exactly Solvable Recurrent Neural Network". In: *Proceedings of the National Academy of Sciences* 122.1, e2321319121. DOI: 10.1073/pnas.2321319121. (Visited on 03/10/2026).

Lillemark, Hansen et al. (Oct. 2025). "Flow Equivariant World Modeling for Partially Observed Dynamic Environments". In: (visited on 03/10/2026).

Lin, Yen-Chu et al. (Apr. 2023). "Cognitive Influences on Fixational Eye Movements". In: *Current Biology* 33.8, 1606–1612.e4. ISSN: 0960-9822. DOI: 10.1016/j.cub.2023.03.026. (Visited on 10/26/2025).

Lindgren, Kristian (2024). *Information Theory for Complex Systems: An Information Perspective on Complexity in Dynamical Systems and Statistical Mechanics*. Understanding Complex Systems. Berlin, Heidelberg: Springer. ISBN: 978-3-662-68212-8 978-3-662-68214-2. DOI: 10.1007/978-3-662-68214-2. (Visited on 03/16/2026).

Liu, Ziming et al. (July 2023). *Discovering New Interpretable Conservation Laws as Sparse Invariants*. arXiv: 2305.19525 [nlin, physics:physics]. (Visited on 07/12/2023).

Lorenzen, Stephan Sloth et al. (Feb. 2022). *Information Bottleneck: Exact Analysis of (Quantized) Neural Networks*. DOI: 10.48550/arXiv.2106.12912. arXiv: 2106.12912 [cs]. (Visited on 10/14/2025).

Lyle, Clare et al. (May 2020). *On the Benefits of Invariance in Neural Networks*. DOI: 10.48550/arXiv.2005.00178. arXiv: 2005.00178 [cs]. (Visited on 10/25/2025).

Ma, Ya-Jing et al. (May 2025). *An Explicit Description of Extreme Points of the Set of Couplings with Given Marginals: With Application to Minimum-Entropy Coupling Problems*. DOI: 10.48550/arXiv.2505.12227. arXiv: 2505.12227 [math]. (Visited on 07/10/2025).

Mac Lane, Saunders (1978). *Categories for the Working Mathematician*. Vol. 5. Graduate Texts in Mathematics. New York, NY: Springer. ISBN: 978-1-4419-3123-8 978-1-4757-4721-8. DOI: 10.1007/978-1-4757-4721-8. (Visited on 12/01/2025).

Mahvari, Mohammad Mahdi et al. (Nov. 2020). *On the Relevance-Complexity Region of Scalable Information Bottleneck*. DOI: 10.48550/arXiv.2011.01352. arXiv: 2011.01352 [cs]. (Visited on 10/14/2025).

— (Feb. 2021). *Scalable Vector Gaussian Information Bottleneck*. arXiv: 2102.07525 [cs, math]. (Visited on 07/12/2023).

Marcel, Valentin et al. (June 2017). "Building a Sensorimotor Representation of a Naive Agent's Tactile Space". In: *IEEE Transactions on Cognitive and Developmental Systems* 9.2, pp. 141–152. ISSN: 2379-8939. DOI: 10.1109/TCDS.2016.2617922. (Visited on 10/28/2025).

Marcel, Valentin et al. (Sept. 2022). "Learning to Reach to Own Body from Spontaneous Self-Touch Using a Generative Model". In: *2022 IEEE International Conference on Development and Learning (ICDL)*. London, United Kingdom: IEEE, pp. 328–335. ISBN: 978-1-6654-1311-4. DOI: 10.1109/ICDL53763.2022.9962186. (Visited on 07/12/2023).

Marchetti, Giovanni Luca et al. (Apr. 2023). "Equivariant Representation Learning via Class-Pose Decomposition". In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 4745–4756. (Visited on 05/21/2024).

Martini, K. Michael et al. (June 2024). "Data Efficiency, Dimensionality Reduction, and the Generalized Symmetric Information Bottleneck". In: *Neural Computation* 36.7, pp. 1353–1379. ISSN: 0899-7667. DOI: 10.1162/neco_a_01667. (Visited on 11/19/2025).

Marzen, Sarah (Dec. 2025). "Resource-Rational Reinforcement Learning and Sensorimotor Causal States, and Resource-Rational Maximiners". In: *Interface Focus* 15.5, p. 20240062. ISSN: 2042-8898. DOI: 10.1098/rsfs.2024.0062. (Visited on 03/12/2026).

Matousek, Jiri et al. (2007). *Understanding and Using Linear Programming*. 1st ed. Universitext. Berlin, Heidelberg: Springer Berlin Heidelberg : Imprint: Springer. ISBN: 978-3-540-30697-9.

Mazzetti, Caterina et al. (Feb. 2026). *A Sub-Riemannian Model of Neural States in the Primary Motor Cortex*. DOI: 10.48550/arXiv.2501.03247. arXiv: 2501.03247 [q-bio]. (Visited on 03/11/2026).

Mediano, Pedro A. M. et al. (May 2022a). "Greater than the Parts: A Review of the Information Decomposition Approach to Causal Emergence". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 380.2227, p. 20210246. ISSN: 1364-503X. DOI: 10.1098/rsta.2021.0246. (Visited on 04/28/2026).

Mediano, Pedro A. M. et al. (Jan. 2022b). "Integrated Information as a Common Signature of Dynamical and Information-Processing Complexity". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 32.1, p. 013115. ISSN: 1054-1500, 1089-7682. DOI: 10.1063/5.0063384. (Visited on 07/12/2023).

Möller, Marco et al. (2023). "Emergence of Common Concepts, Symmetries and Conformity in Agent Groups—an Information-Theoretic Model". In: *Interface Focus* 13.3, p. 20230006. DOI: 10.1098/rsfs.2023.0006.

Montúfar, Guido et al. (Sept. 2015). "A Theory of Cheap Control in Embodied Systems". In: *PLOS Computational Biology* 11.9, pp. 1–22. DOI: 10.1371/journal.pcbi.1004427.

Moss, Sean et al. (Dec. 2023). "A Category-Theoretic Proof of the Ergodic Decomposition Theorem". In: *Ergodic Theory and Dynamical Systems* 43.12, pp. 4166–4192. ISSN: 0143-3857, 1469-4417. DOI: 10.1017/etds.2023.6. (Visited on 09/24/2025).

Muller, Lyle et al. (Sept. 2012). "Propagating Waves in Thalamus, Cortex and the Thalamocortical System: Experiments and Models". In: *Journal of Physiology-Paris*. New Trends in Neurogeometrical Approaches to the Brain and Mind Problem 106.5, pp. 222–238. ISSN: 0928-4257. DOI: 10.1016/j.jphysparis.2012.06.005. (Visited on 03/10/2026).

Muller, Lyle et al. (May 2018). "Cortical Travelling Waves: Mechanisms and Computational Principles". In: *Nature Reviews. Neuroscience* 19.5, pp. 255–268. ISSN: 1471-0048. DOI: 10.1038/nrn.2018.20.

Murphy, Kieran A. (2024). "Machine-Learning Optimized Measurements of Chaotic Dynamical Systems via the Information Bottleneck". In: *Physical Review Letters* 132.19. DOI: 10.1103/PhysRevLett.132.197201.

Murphy, Kieran A. et al. (Oct. 2022a). *Characterizing Information Loss in a Chaotic Double Pendulum with the Information Bottleneck*. arXiv: 2210.14220 [nlin]. (Visited on 07/11/2023).

— (Apr. 2022b). *The Distributed Information Bottleneck Reveals the Explanatory Structure of Complex Systems*. arXiv: 2204.07576 [cond-mat]. (Visited on 07/11/2023).

Musall, Simon et al. (Oct. 2019). "Single-Trial Neural Dynamics Are Dominated by Richly Varied Movements". In: *Nature Neuroscience* 22.10, pp. 1677–1686. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-019-0502-4. (Visited on 07/12/2023).

Nehaniv, Chrystopher L. et al. (2002). "Meaningful Information, Sensor Evolution, and the Temporal Horizon of Embodied Organisms". In: *Proceedings of the Eighth International Conference on Artificial Life*. ICAL 2003. Cambridge, MA, USA: MIT Press, pp. 345–349. ISBN: 0-262-69281-3.

Nelinger, Guy et al. (July 2025). *Object Detection through Dynamic Motor-Sensory Convergence*. DOI: 10.1101/2025.07.16.665117. (Visited on 11/02/2025).

Ngampruetikorn, Vudtiwat et al. (Oct. 2021). *Perturbation Theory for the Information Bottleneck*. arXiv: 2105.13977 [cond-mat, physics:physics]. (Visited on 07/11/2023).

Nguyen, Hai Huu et al. (Dec. 2023). "Equivariant Reinforcement Learning under Partial Observability". In: *Proceedings of The 7th Conference on Robot Learning*. PMLR, pp. 3309–3320. (Visited on 03/07/2026).

No, Albert (2019). "Universality of Logarithmic Loss in Successive Refinement". In: *Entropy* 21.2. ISSN: 1099-4300. DOI: 10.3390/e21020158.

Oizumi, Masafumi et al. (July 2025). *Principal Bundle Geometry of Qualia: Understanding the Quality of Consciousness from Symmetry*. DOI: 10.31234/osf.io/agupq_v2. (Visited on 09/24/2025).

Olsson, Lars et al. (June 2006). "From Unknown Sensors and Actuators to Actions Grounded in Sensorimotor Perceptions". In: *Connect. Sci.* 18, pp. 121–144. DOI: 10.1080/09540090600768542.

O'Regan, J. et al. (Nov. 2001). "A Sensorimotor Account of Vision and Visual Consciousness". In: *The Behavioral and brain sciences* 24, 939–73; discussion 973. DOI: 10.1017/S0140525X01000115.

O'Regan, J. Kevin (June 2011). *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*. Oxford University Press. ISBN: 978-0-19-977522-4. DOI: 10.1093/acprof:oso/9780199775224.001.0001. (Visited on 11/12/2025).

Ortega, Pedro A. et al. (2013). "Thermodynamics as a Theory of Decision-Making with Information-Processing Costs". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 469.2153, p. 20120683. DOI: 10.1098/rspa.2012.0683.

Ouderaa, Tycho F. A. van der et al. (Aug. 2022). "Learning Invariant Weights in Neural Networks". In: *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*. PMLR, pp. 1992–2001. (Visited on 10/27/2025).

Pacelli, Vincent et al. (May 2019). "Task-Driven Estimation and Control via Information Bottlenecks". In: *2019 International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, pp. 2061–2067. ISBN: 978-1-5386-6027-0. DOI: 10.1109/ICRA.2019.8794213. (Visited on 07/11/2023).

Pak, Denizhan (Oct. 2025). *Sensorimotor Contingencies and The Sensorimotor Approach to Cognition*. DOI: 10.48550/arXiv.2510.14227. arXiv: 2510.14227 [q-bio]. (Visited on 11/11/2025).

Palmer, Stephanie E. et al. (2015). "Predictive Information in a Sensory Population". In: *Proceedings of the National Academy of Sciences* 112.22, pp. 6908–6913. DOI: 10.1073/pnas.1506855112.

Parker, Albert E. et al. (Apr. 2022). *The Lack of Convexity of the Relevance-Compression Function*. arXiv: 2204.10957 [cs, math]. (Visited on 07/11/2023).

Pérez Rey, Luis Armando et al. (2023). "Equivariant Representation Learning in the Presence of Stabilizers". In: *Machine Learning and Knowledge Discovery in Databases: Research Track*. Ed. by Danai Koutra et al. Cham: Springer Nature Switzerland, pp. 693–708. ISBN: 978-3-031-43421-1. DOI: 10.1007/978-3-031-43421-1_41.

Perrone, Paolo (Mar. 2024). "Markov Categories and Entropy". In: *IEEE Transactions on Information Theory* 70.3, pp. 1671–1692. ISSN: 1557-9654. DOI: 10.1109/TIT.2023.3328825. (Visited on 09/24/2025).

Petitot, Jean (2017). *Elements of Neurogeometry: Functional Architectures of Vision*. Lecture Notes in Morphogenesis. Cham: Springer International Publishing. ISBN: 978-3-319-65589-5 978-3-319-65591-8. DOI: 10.1007/978-3-319-65591-8. (Visited on 03/11/2026).

Pezzulo, Giovanni et al. (Feb. 2024). "Active Inference as a Theory of Sentient Behavior". In: *Biological Psychology* 186, p. 108741. ISSN: 0301-0511. DOI: 10.1016/j.biopsycho.2023.108741. (Visited on 10/25/2025).

Pfante, Oliver et al. (Mar. 2014). "Comparison between Different Methods of Level Identification". In: *Advances in Complex Systems* 17.02, p. 1450007. ISSN: 0219-5259. DOI: 10.1142/S0219525914500076. (Visited on 10/27/2025).

Pfante, Oliver et al. (2015). "Operator-Theoretic Identification of Closed Sub-Systems of Dynamical Systems". In: *Discontinuity, Nonlinearity, and Complexity* 4.1, pp. 91–109.

Philipona, D. et al. (Sept. 2003). "Is There Something Out There? Inferring Space from Sensorimotor Dependencies". In: *Neural Computation* 15.9, pp. 2029–2049. ISSN: 0899-7667. DOI: 10.1162/089976603322297278. (Visited on 10/28/2025).

Piaget, Jean (1964). "Part I: Cognitive Development in Children: Piaget Development and Learning". In: *Journal of Research in Science Teaching* 2.3, pp. 176–186. ISSN: 1098-2736. DOI: 10.1002/tea.3660020306. (Visited on 10/29/2025).

Plato (1943). *The Republic*. Art-type edition. New York : Books, Inc., 1943.

— (1952). *Phaedrus*. Cambridge : University Press, 1952.

Poincaré, Henri (1952). *Science and Hypothesis*. Dover Publcations, Inc.

Polani, Daniel et al. (2001). "An Information-Theoretic Approach for the Quantification of Relevance". In: *Advances in Artificial Life*. Ed. by Jozef Kelemen et al. Berlin, Heidelberg: Springer, pp. 704–713. ISBN: 978-3-540-44811-2. DOI: 10.1007/3-540-44811-X_82.

Polani, Daniel et al. (2009). "Models of Information Processing in the Sensorimotor Loop". In: *Information Theory and Statistical Learning*. Ed. by Frank Emmert-Streib et al. Boston, MA: Springer US, pp. 289–308. ISBN: 978-0-387-84816-7. DOI: 10.1007/978-0-387-84816-7_12. (Visited on 03/02/2025).

Poletti, Martina et al. (Dec. 2015). "Head-Eye Coordination at a Microscopic Scale". In: *Current Biology* 25.24, pp. 3253–3259. ISSN: 0960-9822. DOI: 10.1016/j.cub.2015.11.004. (Visited on 10/26/2025).

Quessard, Robin et al. (2020). "Learning Disentangled Representations and Group Structure of Dynamical Environments". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 19727–19737. (Visited on 10/19/2025).

Rauh, Johannes et al. (Oct. 2017). "Coarse-Graining and the Blackwell Order". In: *Entropy* 19.10, p. 527. ISSN: 1099-4300. DOI: 10.3390/e19100527. arXiv: 1701.07602 [cs, math]. (Visited on 07/12/2023).

Rauh, Johannes et al. (July 2019). "Unique Information and Secret Key Decompositions". In: *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE. DOI: 10.1109/isit.2019.8849550.

Ravindran, Balaraman et al. (2002). "Model Minimization in Hierarchical Reinforcement Learning". In: *Abstraction, Reformulation, and Approximation*. Ed. by Sven Koenig et al. Berlin, Heidelberg: Springer, pp. 196–211. ISBN: 978-3-540-45622-3. DOI: 10.1007/3-540-45622-8_15.

Rimoldi, B. (1994). "Successive Refinement of Information: Characterization of the Achievable Rates". In: *IEEE Transactions on Information Theory* 40.1, pp. 253–259. DOI: 10.1109/18.272493.

Rolfs, Martin (Oct. 2009). "Microsaccades: Small Steps on a Long Way". In: *Vision Research* 49.20, pp. 2415–2441. ISSN: 0042-6989. DOI: 10.1016/j.visres.2009.08.010. (Visited on 03/19/2026).

— (Aug. 2015). "Attention in Active Vision: A Perspective on Perceptual Continuity Across Saccades". In: *Perception* 44.8-9, pp. 900–919. ISSN: 0301-0066. DOI: 10.1177/0301006615594965. (Visited on 03/18/2026).

Rolfs, Martin et al. (Feb. 2022). "Coupling Perception to Action through Incidental Sensory Consequences of Motor Behaviour". In: *Nature Reviews Psychology* 1.2, pp. 112–123. ISSN: 2731-0574. DOI: 10.1038/s44159-021-00015-x. (Visited on 06/04/2024).

Rolfs, Martin et al. (May 2025). "Lawful Kinematics Link Eye Movements to the Limits of High-Speed Perception". In: *Nature Communications* 16.1, p. 3962. ISSN: 2041-1723. DOI: 10.1038/s41467-025-58659-9. (Visited on 10/26/2025).

Romero, David W. et al. (Dec. 2022). "Learning Partial Equivariances From Data". In: *Advances in Neural Information Processing Systems* 35, pp. 36466–36478. (Visited on 02/26/2025).

Rosas, Fernando et al. (May 2025). "AI in a Vat: Fundamental Limits of Efficient World Modelling for Agent Sandboxing and Interpretability". In: *Reinforcement Learning Conference*. (Visited on 11/11/2025).

Rosas, Fernando E. et al. (2020). "Causal Blankets: Theory and Algorithmic Framework". In: *Active Inference*. Ed. by Tim Verbelen et al. Cham: Springer International Publishing, pp. 187–198. ISBN: 978-3-030-64919-7. DOI: 10.1007/978-3-030-64919-7_19.

Rosas, Fernando E. et al. (June 2024). *Software in the Natural World: A Computational Approach to Hierarchical Emergence*. DOI: 10.48550/arXiv.2402.09090. arXiv: 2402.09090 [nlin]. (Visited on 10/17/2025).

Rose, K. (1998). "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems". In: *Proceedings of the IEEE* 86.11, pp. 2210–2239. DOI: 10.1109/5.726788.

Rucci, Michele et al. (Apr. 2015). "The Unsteady Eye: An Information-Processing Stage, Not a Bug". In: *Trends in Neurosciences* 38.4, pp. 195–206. ISSN: 0166-2236. DOI: 10.1016/j.tins.2015.01.005. (Visited on 06/04/2024).

Rucci, Michele et al. (Oct. 2018). "Temporal Coding of Visual Space". In: *Trends in Cognitive Sciences*. Special Issue: Time in the Brain 22.10, pp. 883–895. ISSN: 1364-6613. DOI: 10.1016/j.tics.2018.07.009. (Visited on 06/04/2024).

Rudin, Walter (Jan. 1987). *Real and Complex Analysis*. McGraw-Hill, Inc.

Rudolph, Daniel J. (Dec. 1979). "An Example of a Measure Preserving Map with Minimal Self-Joinings, and Applications". In: *Journal d'Analyse Mathématique* 35.1, pp. 97–122. ISSN: 1565-8538. DOI: 10.1007/BF02791063. (Visited on 07/31/2025).

Rupe, Adam et al. (Aug. 2022). "Algebraic Theory of Patterns as Generalized Symmetries". In: *Symmetry* 14.8, p. 1636. ISSN: 2073-8994. DOI: 10.3390/sym14081636. arXiv: 2206.15050 [cond-mat, physics:nlin]. (Visited on 07/11/2023).

— (June 2024). "On Principles of Emergent Organization". In: *Physics Reports*. On Principles of Emergent Organization 1071, pp. 1–47. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2024.04.001. (Visited on 04/29/2026).

Sachdeva, Vedant et al. (Mar. 2021). "Optimal Prediction with Resource Constraints Using the Information Bottleneck". In: *PLOS Computational Biology* 17.3, pp. 1–27. DOI: 10.1371/journal.pcbi.1008743.

Salge, Christoph et al. (2014). "Empowerment–An Introduction". In: *Guided Self-Organization: Inception*. Ed. by Mikhail Prokopenko. Berlin, Heidelberg: Springer, pp. 67–114. ISBN: 978-3-642-53734-9. DOI: 10.1007/978-3-642-53734-9_4. (Visited on 11/20/2025).

Santos, Bruno A. et al. (Feb. 2022). "Active Role of Self-Sustained Neural Activity on Sensory Input Processing: A Minimal Theoretical Model". In: *Neural Computation* 34.3, pp. 686–715. ISSN: 0899-7667. DOI: 10.1162/neco_a_01471. (Visited on 10/18/2025).

Sarti, Alessandro et al. (Jan. 2008). "The Symplectic Structure of the Primary Visual Cortex". In: *Biological Cybernetics* 98.1, pp. 33–48. ISSN: 1432-0770. DOI: 10.1007/s00422-007-0194-9. (Visited on 03/11/2026).

Sarti, Alessandro et al. (2022). "Differential Heterogenesis". In: *Differential Heterogenesis: Mutant Forms, Sensitive Bodies*. Ed. by Alessandro Sarti et al. Cham: Springer International Publishing, pp. 55–96. ISBN: 978-3-030-97797-9. DOI: 10.1007/978-3-030-97797-9_4. (Visited on 03/13/2026).

Sasakura, Yasunori et al. (2012). "Ascidians as Excellent Chordate Models for Studying the Development of the Nervous System during Embryogenesis and Metamorphosis". In: *Development, Growth & Differentiation* 54.3, pp. 420–437. ISSN: 1440-169X. DOI: 10.1111/j.1440-169X.2012.01343.x. (Visited on 03/16/2026).

Sasakura, Yasunori et al. (2018). "Formation of Adult Organs through Metamorphosis in Ascidians". In: *WIREs Developmental Biology* 7.2, e304. ISSN: 1759-7692. DOI: 10.1002/wdev.304. (Visited on 11/30/2025).

Sato, Tatsuo K. et al. (July 2012). "Traveling Waves in Visual Cortex". In: *Neuron* 75.2, pp. 218–229. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2012.06.029. (Visited on 03/10/2026).

Saxe, Andrew M. et al. (Dec. 2019). "On the Information Bottleneck Theory of Deep Learning". In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12, p. 124020. DOI: 10.1088/1742-5468/ab3985.

Sericola, Bruno (Aug. 2013). "Markov Chains. Theory and Applications". In: DOI: 10.1002/9781118731543.

Seth, Anil K. (Apr. 2014). "A Predictive Processing Theory of Sensorimotor Contingencies: Explaining the Puzzle of Perceptual Presence and Its Absence in Synesthesia". In: *Cognitive Neuroscience* 5.2, pp. 97–118. ISSN: 1758-8928. DOI: 10.1080/17588928.2013.877880. (Visited on 06/04/2024).

Shalizi, Cosma Rohilla (2001). "Causal Architecture, Complexity and Self -Organization in Time Series and Cellular Automata". PhD thesis. United States – Wisconsin: The University of Wisconsin - Madison. ISBN: 978-0-493-22589-0. (Visited on 10/24/2025).

Shalizi, Cosma Rohilla et al. (Aug. 2001). "Computational Mechanics: Pattern and Prediction, Structure and Simplicity". In: *Journal of Statistical Physics* 104.3, pp. 817–879. ISSN: 1572-9613. DOI: 10.1023/A:1010388907793. (Visited on 10/24/2025).

Shamir, Ohad et al. (2010). "Learning and Generalization with the Information Bottleneck". In: *Theoretical Computer Science* 411.29, pp. 2696–2711. ISSN: 0304-3975. DOI: 10.1016/j.tcs.2010.04.006.

Shannon, C.E. (July 1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27, pp. 379–423.

Shields, P.C. (Oct. 1998). "The Interactions between Ergodic Theory and Information Theory". In: *IEEE Transactions on Information Theory* 44.6, pp. 2079–2093. ISSN: 1557-9654. DOI: 10.1109/18.720532. (Visited on 04/25/2026).

Shwartz-Ziv, Ravid et al. (Apr. 2017). *Opening the Black Box of Deep Neural Networks via Information*. arXiv: 1703.00810 [cs]. (Visited on 07/11/2023).

Shwartz-Ziv, Ravid et al. (2019). *Representation Compression and Generalization in Deep Neural Networks*.

Shwartz Ziv, Ravid et al. (Mar. 2024). "To Compress or Not to Compress—Self-Supervised Learning and Information Theory: A Review". In: *Entropy* 26.3, p. 252. ISSN: 1099-4300. DOI: 10.3390/e26030252. (Visited on 10/27/2025).

Slonim, Noam et al. (Aug. 2006). "Multivariate Information Bottleneck". In: *Neural Computation* 18.8, pp. 1739–1789. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.2006.18.8.1739. (Visited on 07/11/2023).

Song, Yue et al. (Dec. 2023). "Flow Factorized Representation Learning". In: *Advances in Neural Information Processing Systems* 36, pp. 49761–49782. (Visited on 05/21/2024).

Still, S. (Jan. 2009). "Information-Theoretic Approach to Interactive Learning". In: *Europhysics Letters* 85.2, p. 28005. DOI: 10.1209/0295-5075/85/28005.

Stronks, H. Christiaan et al. (Oct. 2016). "Visual Task Performance in the Blind with the BrainPort V100 Vision Aid". In: *Expert Review of Medical Devices* 13.10, pp. 919–931. ISSN: 1743-4440. DOI: 10.1080/17434440.2016.1237287. (Visited on 10/26/2025).

Sun, Qingyun et al. (June 2022). "Graph Structure Learning with Variational Information Bottleneck". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.4, pp. 4165–4174. ISSN: 2374-3468, 2159-5399. DOI: 10.1609/aaai.v36i4.20335. (Visited on 07/12/2023).

Tao, Terence (2011). *An Introduction to Measure Theory*. Graduate Studies in Mathematics 126. American Mathematical Society. ISBN: 978-0-8218-6919-2.

Teichner, Ron et al. (Mar. 2023). "Identifying Regulation with Adversarial Surrogates". In: *Proceedings of the National Academy of Sciences* 120.12, e2216805120. DOI: 10.1073/pnas.2216805120. (Visited on 11/25/2025).

Terekhov, Alexander V. et al. (Mar. 2016). "Space as an Invention of Active Agents". In: *Frontiers in Robotics and AI* 3. ISSN: 2296-9144. DOI: 10.3389/frobt.2016.00004. (Visited on 10/19/2025).

Thura, David et al. (Dec. 2022). "Integrated Neural Dynamics of Sensorimotor Decisions and Actions". In: *PLOS Biology* 20.12, e3001861. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3001861. (Visited on 11/23/2025).

Tian, Chao et al. (2008). "Successive Refinement for Hypothesis Testing and Lossless One-Helper Problem". In: *IEEE Transactions on Information Theory* 54.10, pp. 4666–4681. DOI: 10.1109/TIT.2008.928951.

Tishby, Naftali et al. (Apr. 2000). "The Information Bottleneck Method". In: *arXiv preprint*. arXiv: physics/0004057. (Visited on 07/11/2023).

Tishby, Naftali et al. (Jan. 2011). "The Information Theory of Decision and Action". In: *Percept. Action Cycle Springer Ser. in Cognitive Neural Syst.* Vol. 19, pp. 601–636. ISBN: 978-1-4419-1451-4. DOI: 10.1007/978-1-4419-1452-1_19.

Tishby, Naftali et al. (Mar. 2015). *Deep Learning and the Information Bottleneck Principle*. arXiv: 1503.02406 [cs]. (Visited on 07/11/2023).

Tkačik, Gašper et al. (Mar. 2016). "Information Processing in Living Systems". In: *Annual Review of Condensed Matter Physics* 7.Volume 7, 2016, pp. 89–117. ISSN: 1947-5454, 1947-5462. DOI: 10.1146/annurev-conmatphys-031214-014803. (Visited on 12/12/2024).

Touchette, Hugo et al. (Jan. 2004). "Information-Theoretic Approach to the Study of Control Systems". In: *Physica A: Statistical Mechanics and its Applications* 331.1-2, pp. 140–172. ISSN: 03784371. DOI: 10.1016/j.physa.2003.09.007. arXiv: physics/0104007. (Visited on 07/12/2023).

Travers, Nicholas F. et al. (Jan. 2025). "Equivalence of History and Generator ϵ-Machines". In: *Symmetry* 17.1, p. 78. ISSN: 2073-8994. DOI: 10.3390/sym17010078. (Visited on 10/27/2025).

Tsao, Thomas et al. (Oct. 2022). "A Topological Solution to Object Segmentation and Tracking". In: *Proceedings of the National Academy of Sciences* 119.41, e2204248119. DOI: 10.1073/pnas.2204248119. (Visited on 10/23/2025).

Tschantz, Alexander et al. (Apr. 2020). "Learning Action-Oriented Models through Active Inference". In: *PLOS Computational Biology* 16.4, e1007805. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007805. (Visited on 11/01/2025).

Tucker, Mycal et al. (2022). "Trading off Utility, Informativeness, and Complexity in Emergent Communication". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al.

Tuncel, E. et al. (2003). "Computation and Analysis of the N-Layer Scalable Rate-Distortion Function". In: *IEEE Transactions on Information Theory* 49.5, pp. 1218–1230. DOI: 10.1109/TIT.2003.810627.

Tuncel, Ertem (May 2009). "Capacity/Storage Tradeoff in High-Dimensional Identification Systems". In: *IEEE Transactions on Information Theory* 55.5, pp. 2097–2106. ISSN: 1557-9654. DOI: 10.1109/TIT.2009.2016057. (Visited on 10/14/2025).

van der Ouderaa, Tycho F. et al. (Dec. 2024). "Noether's Razor: Learning Conserved Quantities". In: *Advances in Neural Information Processing Systems* 37, pp. 135943–135965. (Visited on 03/17/2025).

van der Pol, Elise et al. (2020). "MDP Homomorphic Networks: Group Symmetries in Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 4199–4210. (Visited on 09/15/2024).

van der Wilk, Mark et al. (2018). "Learning Invariances Using the Marginal Likelihood". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. (Visited on 10/27/2025).

van Dijk, Sander G. et al. (July 2012). *Informational Drives for Sensor Evolution*. Vol. ALIFE 2012: The Thirteenth International Conference on the Synthesis and Simulation of Living Systems. ALIFE 2022: The 2022 Conference on Artificial Life. DOI: 10.1162/978-0-262-31050-5-ch044.

Varela, Francisco J. et al. (Nov. 1992). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA, USA: MIT Press. ISBN: 978-0-262-72021-2.

Virgo, Nathaniel et al. (Jan. 2022). "Embracing Sensorimotor History: Time-synchronous and Time-Unrolled Markov Blankets in the Free-Energy Principle". In: *Behavioral and Brain Sciences* 45, e215. ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X22000334. (Visited on 11/24/2025).

Virgo, Nathaniel et al. (Aug. 2025). *A "Good Regulator Theorem" for Embodied Agents*. DOI: 10.48550/arXiv.2508.06326. arXiv: 2508.06326 [cs]. (Visited on 11/23/2025).

Volpi, Nicola Catenacci et al. (June 2023). "Goal-Directed Empowerment: Combining Intrinsic Motivation and Task-Oriented Behavior". In: *IEEE Transactions on Cognitive and Developmental Systems* 15.2, pp. 361–372. ISSN: 2379-8939. DOI: 10.1109/TCDS.2020.3042938. (Visited on 12/01/2025).

Wang, Dian et al. (Jan. 2022a). "Equivariant Q Learning in Spatial Action Spaces". In: *Proceedings of the 5th Conference on Robot Learning*. PMLR, pp. 1713–1723. (Visited on 03/07/2026).

Wang, Rui et al. (June 2022b). "Approximately Equivariant Networks for Imperfectly Symmetric Dynamics". In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR, pp. 23078–23091. (Visited on 03/05/2026).

Wang, Siwei et al. (May 2021). "Maximally Efficient Prediction in the Early Fly Visual System May Support Evasive Flight Maneuvers". In: *PLOS Computational Biology* 17.5, pp. 1–27. DOI: 10.1371/journal.pcbi.1008965.

Westphal, Charles et al. (Oct. 2025). *A Generalized Information Bottleneck Theory of Deep Learning*. DOI: 10.48550/arXiv.2509.26327. arXiv: 2509.26327 [cs]. (Visited on 10/24/2025).

Willard, Stephen (1970). *General Topology*. Addison-Wesley Series in Mathematics. Addison-Wesley Publishing Company. ISBN: 0-201-08707-3.

Willeke, Konstantin F. et al. (Aug. 2019). "Memory-Guided Microsaccades". In: *Nature Communications* 10.1, p. 3710. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11711-x. (Visited on 03/19/2026).

Williams, Paul L. et al. (Apr. 2010). *Nonnegative Decomposition of Multivariate Information*. arXiv: 1004.2515 [math-ph, physics:physics, q-bio]. (Visited on 09/27/2023).

Winter, Robin et al. (Dec. 2022). "Unsupervised Learning of Group Invariant and Equivariant Representations". In: *Advances in Neural Information Processing Systems* 35, pp. 31942–31956. (Visited on 10/12/2025).

Witsenhausen, H. et al. (1975). "A Conditional Entropy Bound for a Pair of Discrete Random Variables". In: *IEEE Transactions on Information Theory* 21.5, pp. 493–501. DOI: 10.1109/TIT.1975.1055437.

Worm, Daniël T. H. et al. (Apr. 2011). "Ergodic Decompositions Associated with Regular Markov Operators on Polish Spaces". In: *Ergodic Theory and Dynamical Systems* 31.2, pp. 571–597. ISSN: 1469-4417, 0143-3857. DOI: 10.1017/S0143385710000039. (Visited on 09/06/2025).

Wu, Eric G. et al. (Sept. 2024). "Fixational Eye Movements Enhance the Precision of Visual Information Transmitted by the Primate Retina". In: *Nature Communications* 15.1, p. 7964. ISSN: 2041-1723. DOI: 10.1038/s41467-024-52304-7. (Visited on 03/19/2026).

Wu, Tailin et al. (Jan. 2020). *Phase Transitions for the Information Bottleneck in Representation Learning*. arXiv: 2001.01878 [cs, math, stat]. (Visited on 07/11/2023).

Yang, Qianqian et al. (Nov. 2017). *The Multi-layer Information Bottleneck Problem*. arXiv: 1711.05102 [cs, math, stat]. (Visited on 07/12/2023).

Yarbus, Alfred L. (1967). *Eye Movements and Vision*. Boston, MA: Springer US. ISBN: 978-1-4899-5381-0 978-1-4899-5379-7. DOI: 10.1007/978-1-4899-5379-7. (Visited on 10/26/2025).

Yeung, Raymond W. (2008). *Information Theory and Network Coding*. Springer.

Yousfi, Yassine et al. (Nov. 2020). "Successive Information Bottleneck and Applications in Deep Learning". In: *2020 54th Asilomar Conference on Signals, Systems, and Computers*. Pacific Grove, CA, USA: IEEE, pp. 1210–1213. ISBN: 978-0-7381-3126-9. DOI: 10.1109/IEEECONF51394.2020.9443491. (Visited on 07/12/2023).

Zaidi, Abdellatif et al. (2020). "On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views". In: *Entropy* 22.2. ISSN: 1099-4300. DOI: 10.3390/e22020151.

Zaslavsky, Noga et al. (Aug. 2019). "Deterministic Annealing and the Evolution of Information Bottleneck Representations". In: *Preprint*.

Zaslavsky, Noga et al. (2022). "The Evolution of Color Naming Reflects Pressure for Efficiency: Evidence from the Recent Past". In: *bioRxiv*. DOI: 10.1101/2021.11.03.467047.

Zhao, Zhetuo et al. (Jan. 2023). "Inferring Visual Space from Ultra-Fine Extra-Retinal Knowledge of Gaze Position". In: *Nature Communications* 14.1, p. 269. ISSN: 2041-1723. DOI: 10.1038/s41467-023-35834-4. (Visited on 09/27/2023).