Successive Refinement and Coarsening of the Information Bottleneck

Hippolyte Charvin
Adaptive Systems Research Group
University of Hertfordshire
h.charvin@herts.ac.uk

Nicola Catenacci Volpi
Adaptive Systems Research Group
University of Hertfordshire
n.catenacci-volpi@herts.ac.uk

Daniel Polani
Adaptive Systems Research Group
University of Hertfordshire
d.polani@herts.ac.uk

Abstract

We study two central aspects of information processing in cognitive systems: one is the ability to incorporate fresh information to already learnt models; the other is the “trickling” of information through the many layers of a cognitive processing pipeline. We investigate the extent to which these specific structures of cognitive processing impact their informational optimal limits. To do so, we present mathematical characterisations and low-dimensional numerical examples, which explore formal properties of the Information Bottleneck method: namely, how it relates to successive refinement, and successive coarsening of information.

1 Introduction

One of the most crucial tensions characterising the activity of embodied agents, is that between pursuing one’s own interests, thus behaving in a purposeful and relevant manner, and doing so without violating the constraints on cognitive processing specific to each agent and situation. The framework of information-theoretic bounded rationality [14, 7, 18] operationalises this tension as the fundamental trade-off resulting from the optimization of the agent’s preference, while keeping its information-processing cost below a given bound. More precisely, on the one hand this information-processing cost is measured with the mutual information $I(X;T)$ between an environmental variable X and an agent’s representation T. On the other hand, behavioral relevancy is most often quantified with an explicit utility function, but it can also be measured implicitly through the mutual information $I(Y;T)$ that the representation T extracts about a “relevancy” variable Y. Assuming the information about Y can be only extracted from the environmental variable X, i.e., that the Markov chain $T - X - Y$ holds, the representation T then solves the so-called Information Bottleneck (IB) problem [20, 8]

$$\underset{\substack{q(T|X)\,:\\ T-X-Y,\; I(X;T)\leq\lambda}}{\arg\max} \qquad I(Y;T), \tag{1}$$

where $\lambda > 0$ controls the complexity-relevancy trade-off — so that the solutions to (1) for varying λ draw the optimal limits on the feasible trade-offs. Note that for the actual computation of solutions to (1), due to the non-linearity of the constraint, one usually rather maximises the Lagrangian $I(Y;T) - \beta I(X;T)$, where the trade-off is now parametrised by the multiplier β.

A richer description of real-world biological and artificial agents demands taking into the picture another key characteristic of their information-processing dynamics: namely, that information-

Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

processing operates along hierarchies [11, 10, 5, 9], producing sequences of representations, instead of single ones. Hierarchical and sequential generalisations of the IB problem (1), or related problems, have been proposed: in particular concerning multiple-stage decision-making [7, 13] and the analysis of deep networks' learning dynamics [19, 22, 21]. In this paper, we are interested in the following question: how are the information-processing limits drawn by solutions to (1) — which are not constrained to satisfy any specific structure — affected when one does impose hierarchical sequential structures? We will take a look at two cases of sequential processing. The first one investigates to which extent the optimal information bounds on (1) can be implemented through incremental incorporation of fresh information into already learned models [12]. In other words, in a developmental language, can the optimal bounds on the information trade-off (1) be achieved by a system which first passes through a given stage of development, and then builds on it to acquire further information about the environment [16, 17]? The second one considers the loss of optimality, in terms of discrepancy from the bound (1), induced by information being feed-forwardly processed through many layers of a cognitive processing pipeline. We will investigate these questions from a formal point of view, with mathematical characterisations and numerical experiments on low-dimensional examples.

2 Successive refinement

Consider an incremental learning process where each step incorporates new information so as to refine the previously acquired representations. In a similar spirit to the classic IB, one can describe the optimal informational bounds on this multiple-stage processing: this is precisely what the notion of *successive refinement* does. This notion has initially been formulated for classic rate-distortion problems [6, 15], but inspired by recent work [13, 7], we adapt it to the IB framework. Consider the following two-stage bottleneck problem[1]: for $\lambda_2 > \lambda_1 \geq 0$, we first design a bottleneck $T = T_1$ that solves (1) with $\lambda = \lambda_1$; and then design a solution S_2 to

$$\underset{\substack{q(S_2|X,T_1)\,:\\ (T_1,S_2)-X-Y,\; I(X;S_2|T_1)\leq\lambda_2-\lambda_1}}{\arg\max} \qquad I(Y;S_2|T_1) \tag{2}$$

The second, *conditional* bottleneck S_2 optimally "supplements" the information already optimally acquired in T_1, yielding a final "two-stage bottleneck" $T_2 := (T_1, S_2)$.

Definition 1. There is *successive refinement* from parameters λ_1 to $\lambda_2 > \lambda_1$, if there is a two-stage bottleneck $T_2 := (T_1, S_2)$ with parameters (λ_1, λ_2), such that T_2 is also a one-stage bottleneck with parameter λ_2, i.e., solves the IB problem (1) with $T = T_2$ and $\lambda = \lambda_2$.

Using the chain rule for mutual information [4] and the fact that the inequality constraint must be saturated for the IB problem [1], one can easily prove that there is successive refinement from λ_1 to λ_2 if and only if there exists one-stage bottlenecks T_1 and T_2, with respective parameters λ_1 and λ_2, such that the Markov chain $T_1 - T_2 - X$ holds. To determine whether or not these equivalent properties hold true, we propose a geometric approach. Denote by $Hull(E)$ the convex hull of a set E, and consider the condition

$$\forall t_1 \in \mathcal{T}_1, \qquad p(X|t_1) \in Hull\{p(X|t_2),\ t_2 \in \mathcal{T}_2\}, \tag{3}$$

where t_1, t_2, resp. $\mathcal{T}_1, \mathcal{T}_2$, denote the symbols, resp. alphabets, of distinct one-stage bottlenecks T_1 and T_2, with respective parameters λ_1, λ_2, where $\lambda_1 < \lambda_2$.

Proposition 2. *We use the notations defined above. Successive refinement from λ_1 to λ_2 implies* (3)*, and* (3) *implies successive refinement if the transition matrix defined by $q(X|T_2)$ is injective.*

Moreover, if the bottleneck problem (1) *has strictly concave information curve*[2]*, then for any bottleneck T, the transition matrix defined by $q(X|T)$ can always be chosen injective.*[3]

In short, Proposition 2 says that under conditions that are easily satisfied, there is successive refinement if the convex hull of the $q(X|t_2)$ still contains the $q(X|t_1)$. Equipped with this new characterisation,

[1] Our formulation is inspired by, and equivalent to, the asymptotic formulation in [13]. It is also related to so-called "parallel information-processing hierarchy" in [7].

[2] The information curve is the curve drawn by points $(I(X;T), I(Y;T))$, for bottlenecks T corresponding to every possible $\lambda \geq 0$. It is always concave, and generically strictly concave if $p(Y|X)$ is not deterministic.

[3] The proof of this proposition will be presented in a publication under preparation.

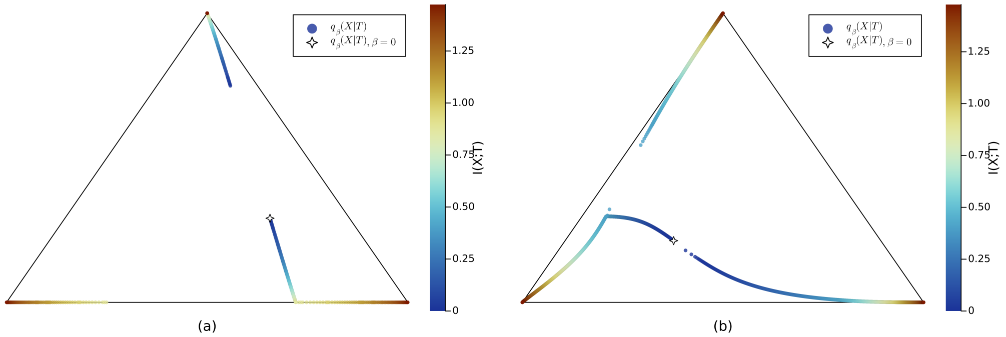

Figure 1: Example trajectories of the $q_\beta(X|T)$. Each point is a $q_\beta(X|t)$ for a fixed symbol t. Colours have been indexed by $I_\beta(X;T)$ instead of β because it allowed better visualisation.

we computed bottlenecks on synthetic examples, with example distributions $p(X,Y)$ sampled uniformly on the simplex. We used the Lagrangian formulation of the IB (see Section 1) — where β is the trade-off parameter — and optimised it with the modified Blahut-Arimoto algorithm [20] combined with reverse deterministic annealing [23], starting at $\beta = \infty$ from the generic optimum $T = X$. We chose $|\mathcal{X}| = |\mathcal{Y}| = 3$, which allows to plot the trajectories of $q_\beta(X|T)$ on the simplex $\Delta_\mathcal{X}$. We only focused on examples with strictly concave information curve and injective $q(X|T_2)$, so that condition (3) does characterise successive refinement.

Figure 1 shows two representative examples[4]. On both plots 1-(a) and 1-(b), we can visualise the bifurcations at which some $q_\beta(X|t)$ splits into two distinct distributions as β grows, yielding an increase of the effective bottleneck cardinality [23]. More importantly for us here, we observe that $\mathit{Hull}\{q_{\beta_2}(X|t), t \in \mathcal{T}\}$ seems to have a general tendency to contain the $q_{\beta_1}(X|t)$ for $\beta_1 < \beta_2$. For instance, this is the case after the split into 3 distinct symbols. But the condition can fail: for instance if the condition held, the first curve segments where we have only two symbols (bluer families of points in Figures 1-(a) and 1-(b)) should be linear segments. But this is not the case for Figure 1-(b): there the first trajectory segment is non-negligibly curved — a relatively typical behavior. Thus, even though there are many pairs of β_1 and $\beta_2 > \beta_1$ for which condition (3) holds, these numerical results suggest that the $q_{\beta_1}(X|t)$, for $\beta_1 < \beta_2$, are not always encompassed by $\mathit{Hull}\{q_{\beta_2}(X|t), t \in \mathcal{T}\}$.

We propose to quantify this "spreading out" from the convex hull with the measure of *unique information* proposed in [3], using the algorithm[5] designed in [2]:

$$UI(X : T_1 \setminus T_2) := \underset{r(X,T_1,T_2) \in \Delta_q}{\arg\min} \; I_r(X; T_1|T_2), \tag{4}$$

where Δ_q is the set of probabilities $r(X,T_1,T_2)$ consistent with the bottleneck probabilities[6], i.e., such that $r(X,T_1) = q(X,T_1)$ and $r(X,T_2) = q(X,T_2)$. This quantity vanishes if and only if there is successive refinement from T_1 to T_2. In the case of Figure 1-(a), where the convex hull condition (3) seems visually always satisfied, the maximum of $UI(X : T_1 \setminus T_2)$ over all T_2 finer than T_1 (i.e., with $\beta_2 > \beta_1$) is approx. $4.0 \cdot 10^{-5}$; whereas in Figure 1-(b), where (3) is clearly broken, it is approx. 0.008. Importantly, even if $UI(X : T_1 \setminus T_2)$ is nonzero in the second case, it is still noticeably small. Among all the examples we studied, it never exceeded 5.4% of $I(X\,;\,T_1, T_2)$.

Overall, these numerical results suggest that imposing two-stage processing, with incorporation of new information at the second stage, does not essentially degrade the optimal informational limits drawn by the classic IB — and the same statement can be made for n-stage processing, by direct iterations of the definitions, arguments and computations in this section. This is encouraging for the feasibility of the optimal IB trade-offs (1) in real-world cognitive processing systems, as the representations they possess about their environment are most often the result of a sequential – and in particular, developmental — process. However, as the modified BA algorithm is only guaranteed to find local optima [20], one should be wary that the observed patterns do not necessarily reflect the behavior of actual, global solutions to (1).

[4] See Appendix B for values of the sample $p(X,Y)$.

[5] Thanks to J. Rauh and P. Banerjee for insightful comments on the algorithm.

[6] The IB method provides us only with $q(X,T_1)$ and $q(X,T_2)$, which does not wholly constrain $q(X,T_1,T_2)$.

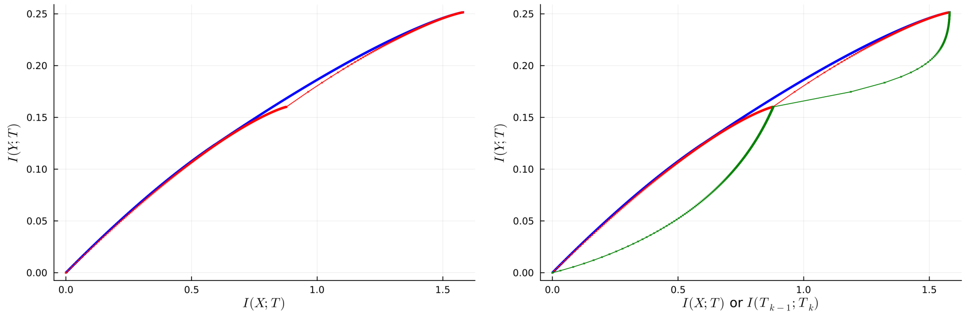

Figure 2: Left: IB curve (blue) vs. DIC curve (red). Right: same plot, superimposed with the curve made of the $(I(T_{k-1};T_k), I(Y;T_k))$ (green). Note that even though here the cusps of the red and green curve touch one another, it was not always the case in other explored examples (not shown).

3 Dense information coarsening

Here we consider the converse situation of that of Section 2, where along a feed-forward cognitive processing pipeline, the information is coarsened rather than refined, without additional external input between layers. In other words, we consider a hierarchy of representations $T_1, \ldots, T_n$, that, by assumption, satisfies the Markov chain $X - T_1 - \cdots - T_n$, where X is an environmental variable, and T_k has only direct access to the previous layer T_{k-1} — deep neural networks being an example of this setting [19]. Note that the indices are in reverse order as those of the previous section, because we stick to an indexing describing the order in which representations are produced. If each layer extracts information about a fixed relevancy variable Y, then the optimal information trade-off at each layer is drawn by bottlenecks T_k with relevancy Y, but where the source X in (1) is replaced by T_{k-1}. The question is then: given these hard constraints on the cognitive processing pipeline, how much information optimality do we lose as compared to a one-stage processing as described by (1)?

We focus on the limit of an infinite number of processing stages, which models cognitive systems where the number of information-processing stages stemming from the sensory interfaces is large. Moreover we consider a processing highly distributed among stages: in other words, each stage performs only a small amount of compression, resulting in a "trickling down" of information through dense information-processing layers. We thus consider a sequence of bottlenecks $(T_k)_{k\in\mathbb{N}^*}$, where T_k is a bottleneck with source T_{k-1} and relevancy Y. We choose the sequence $(\beta_k)_{k\in\mathbb{N}^*}$ with small increments, such that $I(X;T_1) \approx I(X;Y)$ and $I(X;T_k) \to 0$ when $k \to +\infty$. We will refer here to this process as the "dense information coarsening" (DIC for short).

We compared, for distributions $p(X,Y)$ sampled uniformly on the simplex, the information curve of the regular IB (1) with fixed source X, with the DIC information curve — i.e., the curve drawn by points $(I(X;T_k), I(Y;T_k))$. Figure 2 (left) shows a representative example[7] with $|\mathcal{X}| = |\mathcal{Y}| = 3$. Here, and in the other examples explored (not shown), the DIC curve tracks rather neatly the regular IB curve. However, the discrepancy between the curves is non-negligible and the DIC curve displays a distinctive structure: as opposed to the IB curve which is always concave, the DIC curve is only piece-wise concave, resulting in slight cusps at the junctions — in Figure 2 (left), there are two such pieces of concave curve and one cusp. These cusps correspond to symbol splitting points of the DIC trajectories on the simplex $\Delta_{\mathcal{X}}$ (see Appendix A for further details). In Figure 2 (right), we superimposed these graphs with the trajectories of points $(I(T_{k-1};T_k), I(Y;T_k))$ which suggests a hypothesis on the reason for the discrepancy between the DIC and IB curve. Indeed, the curves show that the increase in discrepancy between IB and DIC curves occur concomitantly with a brutal decrease of the information $I(T_{k-1};T_k)$ between consecutive layers, suggesting that this abrupt decrease in cross-layer information induces a divergence from the optimal informational trade-off.

Whether this hypothesis is satisfied in real-world cognitive systems or not, these numerical results suggest a phenomenon of divergence, at critical bifurcation points, from the optimal trade-off drawn by the regular IB (1), for information-processing pipelines with multiple layers resembling very deep neural networks [19]. This phenomenon will be better clarified in future work.

[7] See Appendix B for values of the sample $p(X,Y)$.

References

[1] S. Asoodeh and F. Calmon. Bottleneck problems: An information and estimation-theoretic view. *Entropy*, 22:1325, 11 2020. doi: 10.3390/e22111325.

[2] P. Banerjee, J. Rauh, and G. Montufar. Computing the unique information. pages 141–145, 06 2018. doi: 10.1109/ISIT.2018.8437757.

[3] N. Bertschinger, J. Rauh, E. Olbrich, and N. Ay. Quantifying unique information. *Entropy*, 16, 11 2013. doi: 10.3390/e16042161.

[4] T. Cover and J. Thomas. *Elements of Information Theory, 2nd edition*. Wiley-Interscience, 2006.

[5] A. Diaconescu, V. Litvak, C. Mathys, L. Kasper, K. Friston, and K. Stephan. A computational hierarchy in human cortex. 09 2017.

[6] W. Equitz and T. Cover. Successive refinement of information. *IEEE Transactions on Information Theory*, 37(2):269–275, 1991. doi: 10.1109/18.75242.

[7] T. Genewein, F. Leibfried, J. Grau-Moya, and D. Braun. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, 2, 11 2015. doi: 10.3389/frobt.2015.00027.

[8] R. Gilad-Bachrach, A. Navot, and N. Tishby. An information theoretic tradeoff between complexity and accuracy. *Lecture Notes in Computer Science*, 07 2003. doi: 10.1007/978-3-540-45167-9_43.

[9] S. Hochstein and M. Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002. ISSN 0896-6273. doi: https://doi.org/10.1016/S0896-6273(02)01091-7. URL `https://www.sciencedirect.com/science/article/pii/S0896627302010917`.

[10] E. Koechlin and C. Summerfield. An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6):229–235, 2007. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2007.04.005. URL `https://www.sciencedirect.com/science/article/pii/S1364661307001052`.

[11] E. Koechlin, C. Ody, and F. Kouneiher. The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648):1181–1185, 2003. doi: 10.1126/science.1088545. URL `https://www.science.org/doi/abs/10.1126/science.1088545`.

[12] Y. Luo, L. Yin, W. Bai, and K. Mao. An appraisal of incremental learning methods. *Entropy*, 22(11), 2020. ISSN 1099-4300. doi: 10.3390/e22111190. URL `https://www.mdpi.com/1099-4300/22/11/1190`.

[13] M. Mahvari, M. Kobayashi, and A. Zaidi. On the relevance-complexity region of scalable information bottleneck, 11 2020.

[14] P. A. Ortega and D. A. Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683, 2013. doi: 10.1098/rspa.2012.0683. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2012.0683`.

[15] B. Rimoldi. Successive refinement of information: characterization of the achievable rates. *IEEE Transactions on Information Theory*, 40(1):253–259, 1994. doi: 10.1109/18.272493.

[16] K. S. Scherf, J. A. Sweeney, and B. Luna. Brain basis of developmental change in visuospatial working memory. *Journal of Cognitive Neuroscience*, 18(7):1045–1058, 2006. doi: 10.1162/jocn.2006.18.7.1045.

[17] N. M. Scott, M. D. Sera, and A. P. Georgopoulos. An information theory analysis of spatial decisions in cognitive development. *Frontiers in Neuroscience*, 9, 2015. ISSN 1662-453X. doi: 10.3389/fnins.2015.00014. URL `https://www.frontiersin.org/articles/10.3389/fnins.2015.00014`.

[18] N. Tishby and D. Polani. *The Information Theory of Decision and Action*, volume 19, pages 601–636. 01 2011. ISBN 978-1-4419-1451-4. doi: 10.1007/978-1-4419-1452-1_19.

[19] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop, ITW 2015*, 03 2015. doi: 10.1109/ITW.2015.7133169.

[20] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, 49, 07 2001.

[21] Q. Yang, P. Piantanida, and D. Gündüz. The multi-layer information bottleneck problem. 11 2017.

[22] Y. Yousfi and E. Akyol. Successive information bottleneck and applications in deep learning. In *2020 54th Asilomar Conference on Signals, Systems, and Computers*, pages 1210–1213, 2020. doi: 10.1109/IEEECONF51394.2020.9443491.

[23] N. Zaslavsky and N. Tishby. Deterministic annealing and the evolution of information bottleneck representations. 08 2019.

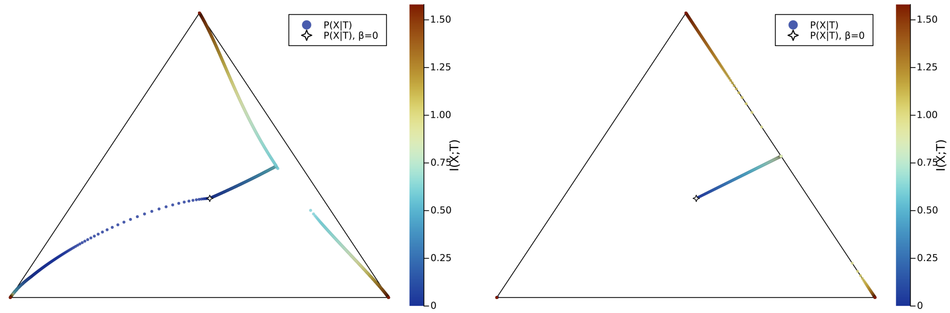

Figure 3: Left: IB trajectory of $q_\beta(X|T)$ distributions over β. Right: DIC trajectory of $q_(X|T_k)$ distributions over k.

Appendices

A IB and DIC trajectories on the simplex

One can compare the simplex trajectories of the IB and DIC methods. Even though a deeper understanding of the observed patterns (in particular how they relate to the main content in Section 3) is left to future work, we still describe them here. In Figure 3 (left), we plot the IB trajectory, on the simplex $\Delta_\mathcal{X}$, of the $q_\beta(X|T)$ distributions corresponding to the IB solutions drawing the optimal IB curve in Figure 2 (blue curve). This IB trajectory is compared, in Figure 3 (right), with the DIC trajectory, still on $\Delta_\mathcal{X}$, of the $q(X|T_k)$ distributions that draw the DIC curve in Figure 2 (red curve). See resp. Sections 2 and 3 for notations.

The two kinds of trajectories are to a large extent similar. Crucially, the DIC trajectory displays symbol merges as k increases, just like the IB trajectory displays symbol merges when β decreases. Moreover the symbol merge in Figure 3 (right) corresponds exactly to the cusp which can be observed on the DIC information curve in Figure 2.

However we notice two key differences: first, contrarily to the IB trajectories which have in general non-zero curvatures, the DIC trajectories are always composed of straight lines between two symbol merges. Second, the symbol merge, which can occur in the interior of the simplex for the IB case, happens on the border of the simplex for the DIC trajectories — a feature which was always satisfied for other explored examples of DIC trajectories.

These numerical results thus hint towards possible ways of understanding the discrepancy between the IB and DIC and information curve: it might be partly explained by the discrepancy between the distributions $q_\beta(X|T)$, resp. $q(X|T_k)$, at which symbol merge occurs in resp. the IB and DIC case; and it might also be related to a different nature of the trajectories on the simplex.

B Values of sample distributions

The distributions $p(X,Y)$ used for our numerical experiments were sampled from the uniform distribution on the probability simplex. In this paper, we presented only features which were qualitatively always satisfied among the examples we studied.

For Figure 1-(a), the sample $p(X,Y)$ was defined from[8]

$$p(X) = \begin{pmatrix} 0.198 \\ 0.511 \\ 0.291 \end{pmatrix}, \qquad p(Y|X) = \begin{pmatrix} 0.401 & 0.328 & 0.271 \\ 0.315 & 0.490 & 0.195 \\ 0.109 & 0.389 & 0.502 \end{pmatrix};$$

[8] The values are rounded to 10^{-3}, and $p(Y|X)$ is written as a column transition matrix.

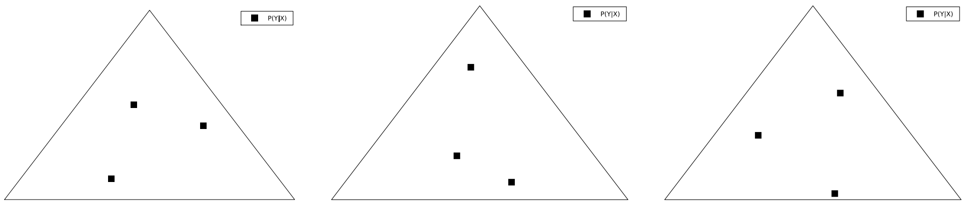

Figure 4: Distributions $p(Y|X)$ on the simplex $\Delta_{\mathcal{Y}}$. Each black square is a $p(Y|x)$ for a given symbol x. The resp. left, middle and right $p(Y|X)$ distributions is that used for resp. Figure 1-(a), Figure 1-(b), and Figure 2.

for Figure 1-(b), from

$$p(X) = \begin{pmatrix} 0.516 \\ 0.270 \\ 0.213 \end{pmatrix}, \qquad p(Y|X) = \begin{pmatrix} 0.519 & 0.310 & 0.226 \\ 0.301 & 0.128 & 0.683 \\ 0.180 & 0.562 & 0.091 \end{pmatrix};$$

and for Figure 2, from

$$p(X) = \begin{pmatrix} 0.298 \\ 0.353 \\ 0.349 \end{pmatrix}, \qquad p(Y|X) = \begin{pmatrix} 0.410 & 0.133 & 0.518 \\ 0.558 & 0.317 & 0.149 \\ 0.032 & 0.550 & 0.332 \end{pmatrix}.$$

The respective conditional distributions $p(Y|X)$ are represented on the $\mathcal{Y}$ simplex in Figure 4.