Formatting Instructions For InfoCog@NeurIPS 2022

September 30, 2022

Abstract

We discuss two models for two central issues appearing in conjunction with cognitive processing. One is the ability to incorporate fresh information to already learnt models; the other one is moving to a better understanding of what happens when information "trickles" through many layers of a cognitive processing pipeline. We do so by investigating formal properties of the Information Bottleneck method: namely how it relates to successive refinement, and successive coarsening of information.

1 Introduction

We discuss two models for two central issues appearing in conjunction with cognitive processing. One is the ability to incorporate fresh information to already learnt models; the other one is moving to a better understanding of what happens when information "trickles" through many layers of a cognitive processing pipeline such as a deep network, or any cognitive system with a sensory interface and a rich inner information processing. Both can be treated using two — in a way — "dual" perspectives of the successive refinement and coarsening of the Information Bottleneck (IB) method [5, 2]. This method describes fundamental limits on the optimal relevancy/complexity trade-offs that one can achieve while extracting (in a single stage of information processing) relevant features within a given system. Similarly, one can use an analogous framework to understand optimal limits on several stages information processing, and compare them to the single-stage case.

Formally, the IB method considers a given joint probability $p(X, Y)$, where X is the source of information (e.g., environmental variables), and Y the variable of interest (e.g., to the agent). The IB method produces a "bottleneck" T (e.g., the agent's representation), which maximises the information about the relevancy Y under constraints on the information extracted from X: i.e., T solves

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

where $\lambda \geq 0$ controls the trade-off between relevancy and complexity. In the next sections, we study first refinement, and then coarsening of information.

2 Successive refinement

The IB framework, like classical rate-distortion problems before it [1, 4], has recently been adapted to the case of several-stages information processing [3]. In learning-theoretic terms[1], a two-stage bottleneck with parameters (λ_1, λ_2), where $\lambda_2 > \lambda_1$, can be described as the successive computation of a solution T_1 to the classic IB

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

and then a solution S_2 to

$$\underset{\substack{q(S_2|X,T_1)\,:\\ (S_2,T_1)-X-Y,\ I(X;S_2|T_1)\leq\lambda_2-\lambda_1}}{\arg\max} \quad I(Y;S_2|T_1) \tag{3}$$

The second, *conditional* bottleneck S_2 optimally "supplements" the information already optimally acquired in T_1, yielding a final representation $T = (T_1, S_2)$. The notion of successive refinement [1, 4, 3] describes if and when this two-stage IB (2)-(3) differs from the one-stage IB (1):

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

We then have simple characterisations:[2]

Proposition 2. *There is successive refinement from λ_1 to $\lambda_2 > \lambda_1$ if and only if there exists distinct one-stage bottlenecks T_1 and T_2, corresponding respectively to λ_1 and λ_2, such that, equivalently*

(i) We have the Markov chain $T_1 - T_2 - X - Y$;

(ii) Or for all $t_1 \in \mathcal{T}_1$, we have $p(X|t_1) \in Hull\{\{p(X|t_2),\ t_2 \in \mathcal{T}_2\}$, where $Hull(E)$ denotes the convex hull of a set E.

To investigate successive refinement, we computed bottlenecks on synthetic examples, with randomly chosen $p(X, Y)$. We used the Lagrangian formulation of the IB [5] - and thus conventionally denote, from now on, β the trade-off parameter - and the corresponding Blahut-Arimoto algorithm [5], combined with reverse deterministic annealing (as in [6]), starting at $\beta = \infty$ from the

[1] Our definition is distinct, but equivalent to the asymptotic definition of [3] - as a consequence of Theorem 1 in [3].

[2] Characterisation (i) - analogous to that for classic rate-distortion problems [1] - is a direct consequence of Theorem 1 in [3], whereas characterisation (ii) is the main result here.

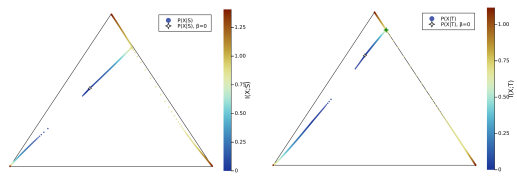

Figure 1: Example trajectory of the $q_\beta(X|T)$.

(generically) optimal distribution $T = X$. We test our geometric characterisation by focusing on very low-dimensional distributions: $|\mathcal{X}| = |\mathcal{Y}| = 3$. We can then plot the trajectories of $q_\beta(X|T)$, on the simplex $\Delta_\mathcal{X}$.

Figure 1 provides two representative examples. On these plots, we can visualise the bifurcations described in [6], which correspond to points where some $q_\beta(Y|t)$, and equivalently some $q_\beta(X|t)$, splits into two symbols. Moreover, between each bifurcation, the bottleneck "trajectories" seem to target specific points, which are the $p([x])$ for some merges $[x]$ of source symbol probabilities. More importantly for us here, we observe that the conditional probabilities $q_\beta(X|t)$ do have a general tendency to contain the $q(X|t_{\beta'})$ for $\beta' > \beta$. For instance, this is usually 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 Figure 1) should be linear segments. But it is clearly not the case for the right plot: on the latter, the first trajectory segment is non-negligibly curved - a relatively typical behavior. Thus, even though there are many pairs of β and $\beta' < \beta$ for which the successive refinement condition holds, these numerical results suggest that this is not generally the case.

3 Dense information coarsening

In the case of sequential information processing, if X is the original source, and T_1, T_2 are two representations at successive stages, we *must* have the Markov chain $X - T_1 - T_2$ - and the question is rather: how much optimality do we loose as compared to a one stage processing? Moreover it is natural to assume that each stage of the information processing is a bottleneck - where the relevancy Y keeps being the same. In the case of cognitive systems, the number of processing stages stemming from the sensory interfaces might be extremely large, and thus it is natural to consider the limit of an infinite number of processing stages. We chose to focus on the case where the processing is highly distributed among stages: in other words, each stage performs only a small amount of compression.

We formalised this situation by considering an original source $T_0 = X$, a fixed relevancy variable Y, and a sequence of bottlenecks $(T_k)_{k\in\mathbb{N}^*}$, where T_k is bottleneck with source T_{k-1} and relevancy Y. We chose the sequence $(\beta_k)_{k\in\mathbb{N}^*}$ with small increments, such that $I(X;T_1) = I(X;Y)$ and $I(X;T_k) \to 0$ when $k \to +\infty$. We will refer to the described method as the "dense information

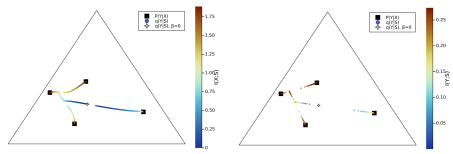

Figure 2: Trajectories of the $q_\beta(Y|T)$, for the IB (left) and the DIC (right). The black square are the $p(Y|x)$. HCsps are even precisely the $\beta = +\infty$ points of the regular IB with constrained cardinality, but this is was not always the case.

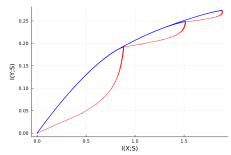

Figure 3: Information curves of relevancy $I(Y;T)$ versus complexity $I(X;T)$. Blue curve: IB; red curve: DIC (computed with the same $p(X,Y)$ as that of Figure 2).

coarsening" (DIC for short).

We then plotted, in Figure 2, the $q_{\beta_k}(Y|T_k)$ trajectories, on the $\Delta_\mathcal{Y}$ simplex, for a representative example, comparing them to the trajectories of a regular IB with source X and relevancy Y. We see that the two trajectories present similarities: they seem to "target", before reaching a bifurcation point, the same. probabilities on the simplex, which are some convex combinations of the $p(Y|x)$ with weights given by the corresponding $p(x)$ - this corresponds to the similar phenomenon seen on the $\Delta_\mathcal{X}$ simplex in Figure 1. The main difference here is that the DIC trajectories move strictly along linear segments. But more crucial are the differences exhibited by the information curves (Figure 3). Differently from the IB information curve, the DIC information curve is not concave. At the contrary, it clearly displays cusps - which correspond to the bifurcation one can observe on the simplex. These bifurcations are points at which the DIC information curve is the closest to the IB information curve.

The overall feature which is interesting to us here, and was present in all the synthetic examples we observed, is that along the flow of the DIC of information, there seems to be crucial points at which the non-optimality of the trade-off, resulting from the fact that the information has successively degraded along the various stages of processing, is minimal. In particular, it seems that at least in some cases, there is virtually no loss of optimality at these crucial, bifurcation points. Thus these points are advantageous to sit on within the information processing of any cognitive system, and it seems reasonable to hypothesise that living organisms have evolved towards organising their information-processing around those critical points.

References

[1] W.H.R. Equitz and T.M. Cover. "Successive refinement of information". In: *IEEE Transactions on Information Theory* 37.2 (1991), pp. 269–275. DOI: 10.1109/18.75242.

[2] Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. "An Information Theoretic Tradeoff between Complexity and Accuracy". In: *Lecture Notes in Computer Science* (July 2003). DOI: 10.1007/978-3-540-45167-9_43.

[3] Mohammad Mahvari, Mari Kobayashi, and Abdellatif Zaidi. *On the Relevance-Complexity Region of Scalable Information Bottleneck*. Nov. 2020.

[4] B. Rimoldi. "Successive refinement of information: characterization of the achievable rates". In: *IEEE Transactions on Information Theory* 40.1 (1994), pp. 253–259. DOI: 10.1109/18.272493.

[5] Naftali Tishby, Fernando Pereira, and William Bialek. "The Information Bottleneck Method". In: *Proceedings of the 37th Allerton Conference on Communication, Control and Computation* 49 (July 2001).

[6] Noga Zaslavsky and Naftali Tishby. "Deterministic annealing and the evolution of Information Bottleneck representations". In: (Aug. 2019).