Deterministic annealing and the evolution of Information Bottleneck representations

Noga Zaslavsky[1] and Naftali Tishby[1,2]

[1]Edmond and Lily Safra Centre for Brain Sciences, The Hebrew University of Jerusalem
[2]Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem

Abstract

The Information Bottleneck (IB) framework provides a principled and broadly applicable approach for studying efficient compressed representations in artificial and biological systems. However, a comprehensive mathematical understanding of the optimal IB representations and the structural phase transitions they undergo via deterministic annealing exists only in a few limited cases. Here, we address the case of symbolic, or discrete, representations, which is particularly relevant to the emergence of language and abstract representations more generally. We characterize the structural changes in the IB representations as they evolve via a deterministic annealing process; derive an algorithm for finding critical points; and explore numerically the types of bifurcations and related phenomena that occur in IB. This work extends the theoretical grounds for understanding optimal representations within the IB framework.

1 Introduction

The Information Bottleneck (IB) framework [1] provides a principled approach for studying efficient compressed representations in artificial and biological systems. In this view, efficient representations should compress their inputs by maintaining the minimal amount of information on the input that is required for making accurate predictions about a target variable. In the past several years, there has been a surge of evidence for the wide applicability of IB in multiple fields, including deep learning [2, 3, 4, 5], and machine learning in general [6, 7], neuroscience [8, 9, 10], language [11, 12], and music [13]. However, a comprehensive mathematical understanding of the structure and evolution of the IB representations exists only in very few cases, usually when Gaussian assumptions are made [7, 14].

The goal of this work is to extend this understanding to the case of discrete random variables, which induce symbolic IB representations. This setting is particularly relevant to the emergence of language [12] and, more broadly, abstract representations. Structural phase

transitions[1] have previously been studied in related settings, such as clustering and classification [15, 16], and to some extent also in the case of IB [17]. The present work goes beyond these previous studies by (1) introducing order parameters that capture the evolution of the IB representations; (2) deriving a novel algorithm for finding critical points in which the representations undergo a phase transition; and (3) exploring numerically the types of phase transitions and related phenomena that occur in IB.

The remainder of this paper is structured as follows. In Section 2 we formulate the notion of efficient compressed representations and ground it in the IB principle. In Section 3 we characterize the evolutionary process of the IB representations and the structural phase transitions they undergo. In Section 4 we present numerical simulations that demonstrate these phenomena.

2 Efficient compressed representations

2.1 Setting

Let $X \in \mathcal{X}$ be a source random variable, $Y \in \mathcal{Y}$ a target variable, and $p(x, y)$ their joint distribution. We assume $p(x, y)$ is known, although in practical applications this distribution is often estimated from data (see [6] for confidence bounds). For simplicity, assume that $\mathcal{X}$ and $\mathcal{Y}$ are finite sets with sizes m and n respectively. For any two random variables, denote by $\triangle(\mathcal{X})$ the $(m-1)$-dimensional simplex of distributions over the elements of $\mathcal{X}$, and by $\triangle(\mathcal{Y})^{\mathcal{X}}$ the set of conditional distributions of Y given X. That is, $\triangle(\mathcal{Y})^{\mathcal{X}} = \triangle(\mathcal{Y}) \times \cdots \times \triangle(\mathcal{Y})$ is the m-ary product of $\triangle(\mathcal{Y})$. We are interested in characterizing efficient representations of X.

Definition 1. *A representation $\hat{X} \in \hat{\mathcal{X}}$ is a stochastic function of X, defined by a conditional distribution $p(\hat{x}|x) \in \triangle(\hat{\mathcal{X}})^{\mathcal{X}}$. If $\hat{\mathcal{X}}$ is a discrete set of arbitrary symbols, then we say that $\hat{X}$ is a symbolic representation of X.*

In this work we consider symbolic representations, where $|\hat{\mathcal{X}}|$ is finite. From an information-theoretic perspective, $p(\hat{x}|x)$ is a stochastic *encoder* and $\hat{\mathcal{X}}$ is the *code alphabet*. In addition, Definition 1 implies that $\hat{X}$ obeys the Markov chain $Y - X - \hat{X}$.

This general setup is broadly applicable. For example, in supervised learning settings [e.g., 11, 6, 2], X would be an input of a classifier, Y would be a target label, and $\hat{X}$ would be an intermediate representation employed by the classifier. In unsupervised learning, this setting corresponds to distributional clustering [e.g., 18, 19], namely assignment of the points $p(y|x) \in \triangle(\mathcal{Y})$ to clusters $\hat{x} \in \hat{\mathcal{X}}$. In statistics, Y may be an unknown parameter of a distribution $p_y(x) = p(x|y)$, in which case X would be a sample from this distribution, and $\hat{X}$ would be a statistic of the sample. In the case of semantic systems [12], Y would be a set of relevant features in the environment, X would be a referent defined by a distribution over features,

[1] We use the term "phase transitions" a bit loosely. Strictly speaking, the phenomena we study are bifurcations, which are not necessary phase transitions in the physical sense.

i.e.$p(y|x)$, and $\hat{X}$ would be a word that is used to communicate the referent.

2.2 The Information Bottleneck method

In all of the settings mentioned above, we may ask: what would be an optimal representation? Intuitively, a good representation should require minimal resources, while achieving maximal predictive power. This intuition is formalized by the Information Bottleneck (IB) principle [1]. According to IB, the complexity of the representation is measured by $I_p(X;\hat{X})$, which is roughly the number of bits that are required for representing X using $\hat{X}$. The informativeness, or accuracy, of the representation is measured by $I_p(\hat{X};Y)$, which is the amount of *relevant information* about Y preserved by the representation. The optimal IB representations minimize $I_p(X;\hat{X})$, such that $I_p(\hat{X};Y)$ remains sufficiently high. Formally, this constrained optimization problem can be solved by minimizing the Lagrangian

$$\mathcal{F}_\beta[p(\hat{x}|x)] = I_p(X;\hat{X}) - \beta I_p(\hat{X};Y)\,, \tag{1}$$

where $\beta \geq 0$ is the Lagrange multiplier for the constraint on $I_p(\hat{X};Y)$. β can also be considered as a tradeoff parameter, or inverse-temperature in analogy to statistical mechanics [15]. Given β, denote the optimal value of the IB objective by $\mathcal{F}^*_\beta$, and the optimal complexity and accuracy by $I_\beta(X;\hat{X})$ and $I_\beta(\hat{X};Y)$ respectively. The IB theoretical limit is defined by the Pareto optimal tradeoffs $(I_\beta(X;\hat{X}), I_\beta(\hat{X};Y))$ as a function of β. This parametric curve [20] is called the *information curve* (see Figure 1A for example).

Tishby et al. [1] showed that a necessary condition for $p_\beta(\hat{x}|x)$ to be a stationary point of $\mathcal{F}_\beta$ is that it satisfies the following self-consistent equations:

$$\begin{cases} p_\beta(\hat{x}|x) &= \dfrac{p_\beta(\hat{x})}{Z_\beta(x)} \exp\left(-\beta D[p(y|x)\|p_\beta(y|\hat{x})]\right) \\ p_\beta(\hat{x}) &= \displaystyle\sum_{x\in\mathcal{X}} p(x)p_\beta(\hat{x}|x) \\ p_\beta(y|\hat{x}) &= \displaystyle\sum_{x\in\mathcal{X}} p(y|x)p_\beta(x|\hat{x}) \end{cases}, \tag{2}$$

where $Z_\beta(x)$ is the normalization factor, also known as the partition function, and $p_\beta(x|\hat{x})$ is obtained by applying Bayes' rule with respect to $p_\beta(\hat{x}|x)$ and $p(x)$. We refer to representations that satisfy (2) as *IB representations*. These representations can be found via the IB method (Algorithm 1), which is a variant of the Blahut–Arimoto algorithm [21, 22].

2.3 Effective cardinality

The cardinality of an IB representation $K(p_\beta)$ is defined by the cardinality of its support, $Supp(p_\beta) = \{\hat{x} \in \hat{\mathcal{X}} : p_\beta(\hat{x}) > 0\}$. That is, $K(p_\beta) = |Supp(p_\beta)|$. The following proposition

Algorithm 1: IB [Tishby et al., 1999]

Input: $p(x, y)$, initial mapping $p_0(\hat{x}|x)$, and tradeoff $\beta \geq 0$
Output: Fixed point of $\mathcal{F}_\beta$
$p(\hat{x}|x) \leftarrow p_0(\hat{x}|x)$
while $p(\hat{x}|x)$ *not converged* **do**
  $p(\hat{x}) \leftarrow \sum_x p(x)p(\hat{x}|x)$
  $p(y|\hat{x}) \leftarrow \sum_x p(y|x)p(x|\hat{x}(\hat{x}))$
  $p(\hat{x}|x) \leftarrow \frac{p(\hat{x})}{Z(x)} \exp\left(-\beta D[p(y|x)\|p(y|\hat{x})]\right)$
return $p(\hat{x}|x)$

shows that there may be a simple transformation that reduces the cardinality of the representation without compromising its optimality given β.

Proposition 1. *If $p_\beta(\hat{x}|x)$ is an IB representation with cardinality K, and there are $\hat{x}_1, \hat{x}_2 \in Supp(p_\beta)$ such that $p_\beta(y|\hat{x}_1) = p_\beta(y|\hat{x}_2)$, then there exists an IB representation $\tilde{p}_\beta(\hat{x}|x)$ with cardinality $K-1$ such that $\mathcal{F}_\beta[\tilde{p}_\beta] = \mathcal{F}_\beta[p_\beta]$.*

Proof. We construct a representation $\tilde{p}_\beta(\hat{x}|x)$ by merging $\hat{x}_1$ and $\hat{x}_2$. For all x and $\hat{x} \neq \hat{x}_1, \hat{x}_2$, let $\tilde{p}_\beta(\hat{x}|x) = p_\beta(\hat{x}|x)$. For $\hat{x}_2$ let $\tilde{p}_\beta(\hat{x}_2|x) = 0$, and for $\hat{x}_1$ let $\tilde{p}_\beta(\hat{x}_1|x) = p_\beta(\hat{x}_1|x) + p_\beta(\hat{x}_2|x)$. Given this construction, it is easy to verify that $\tilde{p}_\beta$ satisfies the IB equations (2), and that $\mathcal{F}_\beta[p_\beta] = \mathcal{F}_\beta[\tilde{p}_\beta]$. In addition, since $\tilde{p}_\beta(\hat{x}_2) = 0$, it holds that $Supp(\tilde{p}_\beta) = Supp(p_\beta)\backslash\{\hat{x}_2\}$, which implies that $K(\tilde{p}_\beta) = K - 1$, and this concludes the proof.

□

p_β and $\tilde{p}_\beta$ are equivalent representations in the sense that they keep the same information about X and Y. More generally, we define the equivalence class of p_β by the set of all representations $\tilde{p}_\beta$ that satisfy the IB equations (2) for the same value of β, and for which there exist mappings $\varphi : \hat{\mathcal{X}} \to \hat{\mathcal{X}}$ and $\psi : \hat{\mathcal{X}} \to \hat{\mathcal{X}}$ such that $\tilde{p}_\beta(y|\varphi(\hat{x})) \equiv p_\beta(y|\psi(\hat{x}))$. In other words, the equivalence class of p_β is determined by the set of distributions over $\mathcal{Y}$ that it induces, i.e.,

$$\{p(y) \in \triangle(\mathcal{Y}) : \exists \hat{x},\, p(y) \equiv p_\beta(y|\hat{x})\}\,. \tag{3}$$

Denote this equivalence class by $[p_\beta]$. Here, we focus on representations with minimal cardinality within their equivalence class.

Definition 2. *The effective cardinality of an IB representation p_β is*

$$k(p_\beta) = \min_{\tilde{p}_\beta \in [p_\beta]} K(\tilde{p}_\beta).$$

We say that $p_\beta(\hat{x}|x)$ is a canonical IB representation if $k(p_\beta) = K(p_\beta)$.

In the remainder of this paper we assume that the IB representations are canonical, unless stated otherwise. In particular, this implies that $p_\beta(y|\hat{x}_1) \neq p_\beta(y|\hat{x}_2)$ for all $\hat{x}_1 \neq \hat{x}_2$.

Algorithm 2: Reverse Deterministic Annealing for IB (RDA-IB)

Input: $p(x,y)$, scheduling $\beta_t > \beta_{t-1} > \cdots > \beta_1 \geq 0$
Output: Fixed points for all β_i
$p_0(\hat{x}|x) \leftarrow I_m$ (initialize)
for $i = t, t-1, \ldots, 1$ **do**
  $p_i(\hat{x}|x) \leftarrow \mathrm{IB}\left(p(x,y), p_{i-1}(\hat{x}|x), \beta_i\right)$ (initialize IB with the previous f.p.)
return $\{p_i(\hat{x}|x)\}_{i=1}^{t}$

Notice that for $\beta = 0$, the global optimum is trivial, and any $\hat{X}$ that is independent of X will attain the minimum $\mathcal{F}_0^* = 0$. In fact, this holds for all $\beta \in [0,1]$, because $I(X;\hat{X}) \geq I(\hat{X};Y)$ due to the Data Processing Inequality [23]. A canonical representation in this case is a constant $\hat{x}$, and so the effective cardinality is $k = 1$. As $\beta \to \infty$, the optimal mapping from X to $\hat{X}$ becomes deterministic, and the effective cardinality would be maximal. In particular, if $|\hat{\mathcal{X}}| \geq |\mathcal{X}|$, then the global optimum is attained by any one-to-one mapping from $\mathcal{X}$ to $\hat{\mathcal{X}}$.[2] In between these two extremes, as β gradually increases, the IB representations undergo a sequence of structural changes, also called phase transitions or bifurcations, in which the effective cardinality changes.

Intuitively, we can think of $I_\beta(X;\hat{X})$ as the logarithm of the effective cardinality because

$$k(p_\beta) \approx 2^{I_\beta(X;\hat{X})}\,. \tag{4}$$

This follows from the same typicality argument that Shannon applied in Rate-Distortion theory [24], which implies that $I_\beta(X;\hat{X})$ is roughly the minimal number of bits that are needed for encoding X using $\hat{X}$.

2.4 Reverse deterministic annealing

The IB optimization problem is non-convex, and thus Algorithm 1 is prone to converge to local minima of $\mathcal{F}_\beta$. A common approach for mitigating this problem is based on the notion of deterministic annealing [15, 16, and see also 25]. A deterministic annealing optimization procedure starts with an initial solution for a low value of β, e.g., $\beta = 0$, for which finding a globally optimal solution is trivial. Then, the solution is refined by invoking the iterative algorithm while gradually increasing β (cooling down the system) according to some annealing schedule. This process attempts to track the optimal solution as β increases from 0 to ∞.

Here, we are not only interested in the solution for $\beta \to \infty$, but rather in the whole trajectory which captures the evolution of the IB representations. In fact, if $|\hat{\mathcal{X}}| = |\mathcal{X}|$, then the solution for $\beta \to \infty$ is straight forward, as mentioned earlier. This suggests a *reverse deterministic annealing* procedure, which starts with a bijective representation and a large value of β, and

[2] We assume here that a non-trivial minimal sufficient statistics (MSS) of X for Y does not exists. If it does, then at the limit $\hat{X}$ would be isomorphic to the MSS.

then gradually decreases β. This procedure is summarized in Algorithm 2. The numerical simulations in Section 4 are based on reverse deterministic annealing because we found it to be more numerically stable than deterministic annealing, while yielding overall similar results.

3 Characterizing the evolution of the IB representations

In this section, we present several tools for characterizing the evolution of the IB representations and the structural phase transitions they undergo. More specifically, we propose several measures that reflect these structural changes as β varies, and introduce an algorithm for finding critical values of β. In addition, we analyze the structural phase transitions in the special case where Y is a deterministic function of X.

3.1 Bifurcations in IB

Bifurcation diagrams are a powerful method for observing qualitative changes in the fixed points of a dynamical system that occur when varying a bifurcation parameter [26]. In our case, the dynamics is defined by the iterative process of Algorithm 1, the fixed points of this process are the IB representations, and the bifurcation parameter is β. Typically, bifurcation diagrams show the fixed points of the system as a function of the bifurcation parameter. However, the IB fixed points are usually high-dimensional distributions, and so it is not always clear how to observe their bifurcations. Here we discuss how to address this issue in two cases: (1) when Y is binary, and (2) in the more general case of discrete variables.

3.1.1 Centroid bifurcations

We have shown in Section 2.3 that the set $\{p(y) \in \triangle(\mathcal{Y}) : \exists \hat{x},\, p(y) \equiv p_\beta(y|\hat{x})\}$ defines the equivalence class $[p_\beta]$. This implies that it is sufficient to consider $p_\beta(y|\hat{x})$ as a function of β, instead of $p_\beta(\hat{x}|x)$. In the case in which Y is binary, this reduces to a single parameter for each $\hat{x}$, namely $p_\beta(y=1|\hat{x})$. We refer to this type of bifurcation diagram as the *centroid bifurcation diagram*, because $p_\beta(y|\hat{x})$ can be viewed as the cluster centroids of the points $p(y|x) \in \triangle(\mathcal{Y})$ under the clustering $p_\beta(\hat{x}|x)$.

Figure 1D shows an example of this type of bifurcation diagram. For $\beta = 1$ there is only one possible value, $p_\beta(y=1|\hat{x}) = 0.5$, which corresponds to the prior distribution $p(y)$. This fixed point remains stable (and optimal) also for β greater than 1, but smaller from some critical value β_0. The first bifurcation occurs at β_0, when the prior centroid splits into two centroids and the effective cardinality increases. It is easy to verify that $p_\beta(y|\hat{x}) = p(y)$ remains a fixed point of the IB equations even for $\beta > \beta_0$, by simply substituting this solution in (2). However, this fixed point loses its stability at β_0 and is no longer globally optimal after that point. This type of phase transition is analogous to a pitchfork bifurcation [26]. A second critical point can

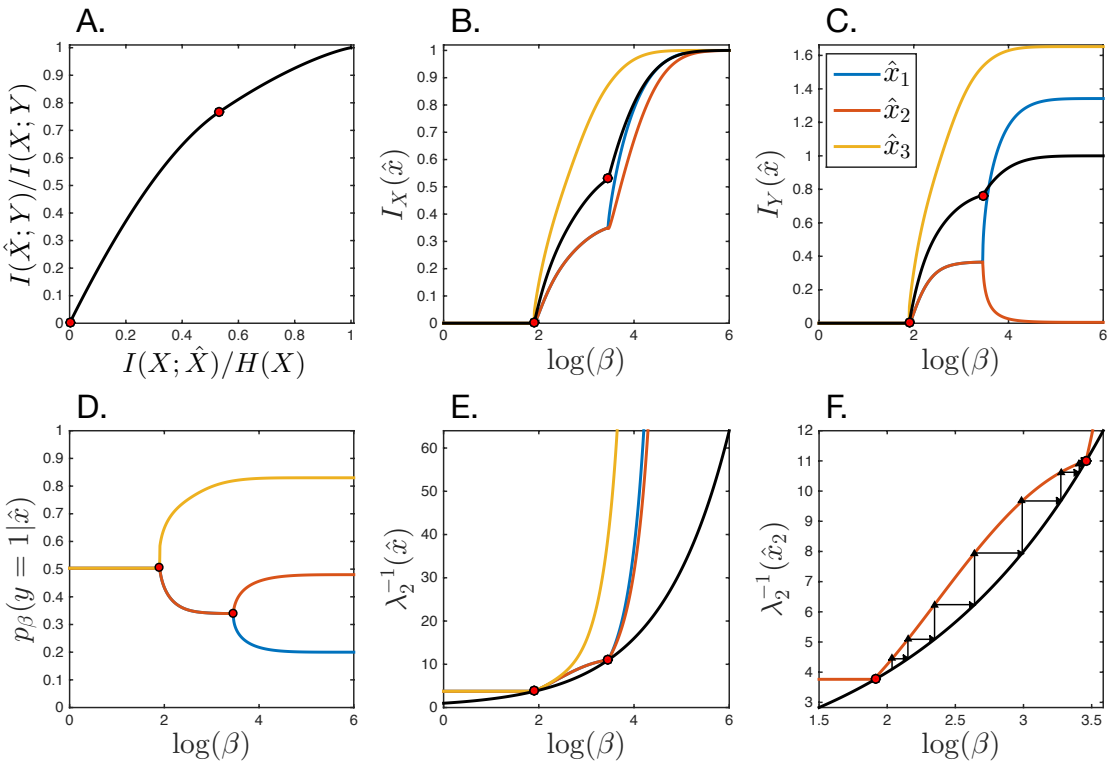

Figure 1. Characterization of the evolution of IB representations in an illustrative example. Here, Y is binary, X and $\hat{X}$ are trinary, $p(x)$ is uniform, and $p(y|x)$ is shown by the leafs of the bifurcation tree in panel D. The red points in all panels correspond to the two critical points that were found by Algorithm 3. **A.** Normalized information curve. **B-C.** Bifurcations of the (normalized) order parameters $I_X(\hat{x})$ and $I_Y(\hat{x})$ respectively. The expected values over $\hat{x}$, i.e. I_X and I_Y, are shown by the black curves. **D.** Centroid bifurcation diagram. **E.** Evolution of $\lambda_2^{-1}(\hat{x})$ for each $\hat{x}$, where $\lambda_2(\hat{x})$ is the second largest eigenvalue of $C_Y^{\beta,\hat{x}}$. The black curve shows β. **F.** Closer view of $\lambda_2^{-1}(x_2)$ (red curve in panel E) and β (black curve in panel E) near the two critical points. Black arrows show iterations of Algorithm 3 in which λ_2 is computed given β (vertical arrows) and then β is updated from λ_2 (horizontal arrows) until convergence. See main text for more detail.

also be seen in Figure 1D, in which another split occurs. The effective cardinality after this split is $K = 3$, and since in this case $|\hat{\mathcal{X}}| = 3$, another bifurcation after this point is impossible.

3.1.2 Informational bifurcations

Centroid bifurcation diagrams are useful when Y is binary, but are difficult to visualize when $|\mathcal{Y}| > 2$. Therefore, we propose an alternative approach that can be applied in the more general case of discrete variables. To this end, we define two informational measures that reflect the structural changes in IB as a function of β.

Definition 3. *Given an IB representation p_β for some $\beta \geq 0$, the point-wise information of*

$\hat{x} \in Supp(p_\beta)$ *about* X *or* Y*, denoted by* $I_X^\beta(\hat{x})$ *or* $I_Y^\beta(\hat{x})$ *respectively, are defined as*

$$I_X^\beta(\hat{x}) = D\left[p_\beta(x|\hat{x})\|p(x)\right] \tag{5}$$
$$I_Y^\beta(\hat{x}) = D\left[p_\beta(y|\hat{x})\|p(y)\right] , \tag{6}$$

where $D[\cdot\|\cdot]$ *is the Kullback-Leibler divergence. These measures are undefined for* $\hat{x} \notin Supp(p_\beta)$.

Notice that before the first phase transition, i.e. for $\beta < \beta_0$, if these measures are defined then necessarily $I_X^\beta(\hat{x}) = 0$ and $I_Y^\beta(\hat{x}) = 0$. At a critical point $\beta_{\hat{x}}$ after which $\hat{x} \in Supp(p_\beta)$, these two informational measures become non-negative. We refer to these measures as *order parameters*, as they are indicative of structural changes in the representation. Note that

$$I_\beta(\hat{X};Y) = \mathop{\mathbb{E}}_{\hat{x}\sim p_\beta(\hat{x})}[I_Y^\beta(\hat{x})] \tag{7}$$

and similarly, $I_\beta(X;\hat{X}) = \mathbb{E}[I_X^\beta(\hat{x})]$. In this sense, these two informational order parameters compose the information curve.

Figure 1B and Figure 1C show the changes of these order parameters as a function of $\log(\beta)$, for the same illustrative example considered in Section 3.1.1. We refer to these types of diagrams as *informational bifurcation diagrams*. The structural changes of the IB representations, directly observed in Figure 1D, are also reflected in the informational bifurcation diagrams, in which cardinality changes are accompanied by an emergence of an order parameter. This order parameter corresponds to the $\hat{x}$ that has been added to $Supp(p_\beta)$. These structural changes are also reflected in the expected values of the order parameters (black curves in Figure 1B-C), i.e., $I_\beta(X;\hat{X})$ and $I_\beta(\hat{X};Y)$, which have discontinuous derivatives with respect to β at the critical points. The following proposition shows that these discontinuities occur exactly at the same values of β.

Proposition 2. $\frac{\partial}{\partial\beta} I_\beta(X;\hat{X}) = \beta \frac{\partial}{\partial\beta} I_\beta(\hat{X};Y)$.

Proof. Substituting the explicit form of $p_\beta(\hat{x}|x)$, as given by (2), in $I_\beta(X;\hat{X})$ gives

$$\begin{aligned} I_\beta(X;\hat{X}) &= \sum_{x,\hat{x}} p(x)p_\beta(\hat{x}|x)\left(-\beta D[p(y|x)\|p_\beta(y|\hat{x})] - \log Z_\beta(x)\right) \\ &= -\mathop{\mathbb{E}}_x\left[\log Z_\beta(x)\right] - \beta\left(I(X;Y) - I_\beta(\hat{X};Y)\right) , \end{aligned}$$

where the second step follows from Lemma 1 in the Appendix. Therefore, the derivative with respect to β is

$$\frac{\partial}{\partial\beta} I_\beta(X;\hat{X}) = -\frac{\partial}{\partial\beta}\mathop{\mathbb{E}}_x\left[\log Z_\beta(x)\right] - \left(I(X;Y) - I_\beta(\hat{X};Y)\right) + \beta\frac{\partial}{\partial\beta} I_\beta(\hat{X};Y).$$

Lemma 2 in the Appendix shows that $\frac{\partial}{\partial\beta}\mathbb{E}_x\left[\log Z_\beta(x)\right] = I_\beta(\hat{X};Y) - I(X;Y)$. Substituting this in the equation above concludes the proof.

□

Another implication of Proposition 2 is that the discontinuities in $\frac{\partial}{\partial\beta}I_\beta(X;\hat{X})$ and $\frac{\partial}{\partial\beta}I_\beta(\hat{X};Y)$ coincide with Ehrenfest's definition of second-order phase transitions [27]. According to Ehrenfest, a second-order phase transition occurs if the second order derivative of the free energy $\mathcal{F}^*_\beta$ is discontinuous, but not the first order derivative. The following corollary shows that the n-th order derivative of $\mathcal{F}^*_\beta$ is given by the $(n-1)$-th order derivative of $-I_\beta(\hat{X};Y)$.

Corollary 1. $\frac{\partial}{\partial\beta}\mathcal{F}^*_\beta = -I_\beta(\hat{X};Y)$.

Proof. This follows directly from Proposition 2 because taking the derivative of $\mathcal{F}^*_\beta$ with respect to β gives

$$\frac{\partial}{\partial\beta}\mathcal{F}^*_\beta = \frac{\partial}{\partial\beta}I_\beta(X;\hat{X}) - \beta\frac{\partial}{\partial\beta}I_\beta(\hat{X};Y) - I_\beta(\hat{X};Y)\,.$$

□

Therefore, if the first-order derivative of $I_\beta(\hat{X};Y)$ is discontinuous, then so is the second-order derivative of $\mathcal{F}^*_\beta$. If $I_\beta(\hat{X};Y)$ is continuous in β, then this corresponds to Ehrenfest's second-order phase transition, and otherwise to a first-order phase-transition. Furthermore, proposition 2 and corollary 1 suggest that in practice it is sufficient to consider only $I^\beta_Y(\hat{x})$ as the order parameter. This conclusion is further supported by lemmas 3 and 4 in the Appendix, which show more precisely how the two order parameters and their derivatives are related.

3.2 Finding critical points

Thus far we have showed that the evolution of IB representations is reflected in a set of order parameters, $\mathcal{O} = \{I^\beta_Y(\hat{x}) : \hat{x} \in Supp(p_\beta), \beta \geq 0\}$. These parameters capture the evolutionary trajectory and the critical values of β in which second order phase transitions occur. A natural question is then: Given a joint distribution $p(x,y)$, what are the values of these critical points? To address this question, we propose an algorithm for finding such points. We refer to this algorithm as *Criticality Search* (Algorithm 3). First, we derive a necessary condition for a second-order phase transition, which will form the basis of the algorithm.

Following a similar argument as in [15], we consider small perturbations of the IB representation near a critical point. At a critical point in which a cluster splits continuously, there exist non-trivial perturbations $h_{\tilde{\beta}}(x,\hat{x})$ such that for all $\tilde{\beta} \geq \beta$, in a small vicinity of β, it holds that $\tilde{p}_\beta(\hat{x}|x) = p_\beta(\hat{x}|x) + h_{\tilde{\beta}}(x,\hat{x})$ satisfies the IB equations (2) for $\tilde{\beta}$. Assuming that the right derivatives of $h_{\tilde{\beta}}(x,\hat{x})$ and $p_{\tilde{\beta}}(\hat{x}|x)$ with respect to $\tilde{\beta}$ exist and are non-zero at $\tilde{\beta} = \beta$, then $\nabla_{h_{\tilde{\beta}}} \log p_{\tilde{\beta}}(\hat{x}|x)\big|_{\tilde{\beta}=\beta}$ is well-defined, and so are these derivatives for $\log p_{\tilde{\beta}}(x|\hat{x})$ and $\log p_{\tilde{\beta}}(y|\hat{x})$. As in [15], we neglect the influence of inter-cluster interactions, which yields

in our case the approximation

$$\mathbf{u}_{\hat{x},\beta}[x] \triangleq \sum_{x'} \frac{\partial \log p_\beta(x|\hat{x})}{\partial h_\beta(x',\hat{x})} \approx \beta \sum_y p(y|x) \sum_{x'} \frac{\partial \log p_\beta(y|\hat{x})}{\partial h_\beta(x',\hat{x})} \tag{8}$$

$$\mathbf{v}_{\hat{x},\beta}[y] \triangleq \sum_{x'} \frac{\partial \log p_\beta(y|\hat{x})}{\partial h_\beta(x',\hat{x})} = \sum_x \frac{p(y|x)p_\beta(x|\hat{x})}{p_\beta(y|\hat{x})} \sum_{x'} \frac{\partial \log p_\beta(x|\hat{x})}{\partial h_\beta(x',\hat{x})} \,. \tag{9}$$

The coupled equations (8)-(9) can be re-organized and simplified as follows:

$$\mathbf{u}_{\hat{x},\beta}[x] \approx \beta \sum_y p(y|x) \sum_{x'} \frac{p(y|x')p_\beta(x'|\hat{x})}{p_\beta(y|\hat{x})} \mathbf{u}_{\hat{x},\beta}[x'] \tag{10}$$

$$\mathbf{v}_{\hat{x},\beta}[y] \approx \beta \sum_x \frac{p(y|x)p_\beta(x|\hat{x})}{p_\beta(y|\hat{x})} \sum_{y'} p(y'|x)\mathbf{v}_{\hat{x},\beta}[y'] \,. \tag{11}$$

This gives two non-linear eigenvector conditions for a cluster split,

$$(\beta^{-1} I - C_X^{\beta,\hat{x}})\mathbf{u}_{\hat{x},\beta} = 0 \tag{12}$$

$$(\beta^{-1} I - C_Y^{\beta,\hat{x}})\mathbf{v}_{\hat{x},\beta} = 0 \,, \tag{13}$$

where $C_X^{\beta,\hat{x}}$ is a $|\mathcal{X}| \times |\mathcal{X}|$ matrix defined by

$$C_X^{\beta,\hat{x}}[x,x'] = \sum_y \frac{p(y|x)p(y|x')p_\beta(x'|\hat{x})}{p_\beta(y|\hat{x})} \,,$$

and $C_Y^{\beta,\hat{x}}$ is a $|\mathcal{Y}| \times |\mathcal{Y}|$ matrix defined by

$$C_Y^{\beta,\hat{x}}[y,y'] = \sum_x \frac{p(y|x)p_\beta(x|\hat{x})p(y'|x)}{p_\beta(y|\hat{x})} \,.$$

For brevity, we simplify the notation by omitting the explicit reference to β and $\hat{x}$ when their actual values are implied or can be arbitrary. It follows that under our assumptions, a necessary (approximated) condition for a second-order phase transition that involves $\hat{x}$ is that β^{-1} is an eigenvalue of $C_X^{\beta,\hat{x}}$ and $C_Y^{\beta,\hat{x}}$. We note that the condition on C_X is closely related to the bifurcation analysis of [17]. Next, we show that both C_X and C_Y are stochastic matrices with the same non-zero eigenvalues.

Proposition 3. *C_Y and C_X have the same non-zero eigenvalues, and their largest eigenvalue is always 1 with $\mathbf{1}$ as an eigenvector.*

Proof. The first part follows from the fact that for any two $m \times n$ real matrices, A and B, it

holds that $AB^\top$ and $A^\top B$ have the same eigenvalues. For any given $\beta \geq 0$ and $\hat{x} \in \hat{\mathcal{X}}$, let

$$\begin{aligned} A[x,y] &= p(y|x) \\ B[x,y] &= \frac{p(y|x)p_\beta(x|\hat{x})}{p_\beta(y|\hat{x})} \,. \end{aligned}$$

It is easy to verify that $C_X = AB^\top$ and $C_Y = B^\top A$. Next, we will show that C_X and C_Y are stochastic matrices. All the values in these matrices are clearly positive, and so it remains to show that the rows sum up to 1. Notice that $B[x,y] = p_\beta(x|\hat{x},y)$, and thus

$$\begin{aligned} \sum_{x'} C_X[x,x'] &= \sum_{x'}\sum_{y} p(y|x)p_\beta(x'|\hat{x},y) = 1 \\ \sum_{y'} C_Y[y,y'] &= \sum_{y'}\sum_{x} p(y'|x)p_\beta(x|\hat{x},y) = 1 \,. \end{aligned}$$

It follows from the Perron–Frobenius Theorem that for both C_X and C_Y, the largest eigenvalue is always 1 with eigenvector $\mathbf{1}$. □

An immediate conclusion from Proposition 3 is that it is sufficient to find the eigenvalues only for the lower dimensional matrix, which is typically C_Y. Furthermore, this criticality condition becomes particularly simple when Y is binary.

Corollary 2. *Assume $|\mathcal{Y}| = 2$, then a necessary condition for a phase transition at β is that there is some $\hat{x} \in \hat{\mathcal{X}}$ for which $\beta = \det(C_Y^{\beta,\hat{x}})^{-1}$.*

Proof. For 2×2 stochastic matrices, the first eigenvalue is $\lambda_1 = 1$ and the second eigenvalue λ_2 is given by the determinant. Therefore, for a binary Y it holds that $\lambda_2(\hat{x}) = \det(C_Y^{\beta,\hat{x}})$, which implies that a necessary condition for (13) is $\beta = \det(C_Y^{\beta,\hat{x}})^{-1}$. □

Another conclusion from Proposition 3 is that the criticality condition cannot hold for $\beta < 1$, because the largest eigenvalue is always 1. This is consistent with the fact that the first critical point β_{c_0} is necessarily greater or equal than 1 (see Section 2.3). For $1 \leq \beta \leq \beta_{c_0}$, any trivial representation for which $p(x|\hat{x}) = p(x)$ and $p(y|\hat{x}) = p(y)$ is optimal, yielding $C_Y^0[y,y'] = \sum_x p(x|y)p(y'|x)$ which is independent of β and $\hat{x}$. Therefore, finding β_{c_0} amounts to finding the eigendecomposition of C_Y^0. For $\beta > \beta_{c_0}$, $C_Y^{\beta,\hat{x}}$ may vary with β resulting in the self-consistent condition $\beta^{-1} \in \mathrm{Eig}(C_Y^{\beta,\hat{x}})$ for criticality, where $\mathrm{Eig}(C_Y^{\beta,\hat{x}})$ is the set of eigenvalues of $C_Y^{\beta,\hat{x}}$. Therefore, finding critical points after β_{c_0} is no longer a simple eigendecomposition problem. To address this problem, we propose the Criticality Search algorithm.

3.2.1 Criticality Search

Criticality Search (Algorithm 3) is an iterative algorithm that finds candidate values of β that satisfy the self-consistent criticality condition. It starts with an initial guess $\beta_c(\hat{x}) = \beta_0$, computes the eigenvalues of C_Y (assuming Y is the lower-dimensional variable), and then checks

the criticality condition. If the condition is not met, the algorithm picks another candidate by making an educated guess:

$$\beta_c^{new}(\hat{x}) = \min\{\lambda^{-1}(\hat{x}) : \lambda \in \mathrm{Eig}(C_Y^{\beta_c,\hat{x}}), \lambda \neq 1\}\,. \tag{14}$$

When Y is binary, this guess simply becomes $\beta_c^{new}(\hat{x}) = \det(C_Y^{\beta,\hat{x}})^{-1}$. This process is repeated for each $\hat{x}$ until a point $\beta_c(\hat{x})$ that satisfies the condition is found, or when β is large enough such that a maximally-informative point is reached, i.e. when $I_\beta(\hat{X};Y) = I(X;Y)$.

The algorithm is demonstrated by the simulations of Figure 1. The red points in all panels correspond to the two critical points found by the algorithm. It can be seen that these points correspond to the structural phase transitions observed in the centroid bifurcation diagram (Figure 1D) and in the informational bifurcation diagrams (Figure 1B-C). The iterations of the algorithms are demonstrated in Figure 1F. This figure shows a run that was initialized with β_0 slightly larger than the first critical point. It converged to the second critical point for $\hat{x}_2$ by iterating between the red curve, which corresponds to $\lambda_2^{-1}(\hat{x})$, and the black curve, which corresponds to β. The fixed points of this iterated map are precisely the points in which the criticality condition is met. While our criticality condition only approximates a necessary condition for a phase transition, in all our numerical simulations the algorithm converged to actual critical points. This suggests that the condition we derived is a good approximation. In addition, we conjecture that while it is possible that the condition is met at non-critical points, these points might be unstable fixed points of the algorithm.

3.3 The deterministic case

To complete our characterization of the IB phase transitions, we discuss the special case in which Y is a deterministic function of X. This case exhibits qualitatively different behavior compared to cases in which $p(y|x) > 0$ for all x and y, and has recently been explored in the context of deep learning [28].

First, we argue that we can consider without loss of generality the case in which $p(y|x)$ is deterministic and defines a one-to-one mapping from X to Y. That is, for every x there is a unique value $y(x)$ such that $p(y'|x) = \delta_{y',y(x)}$. Otherwise, if there exist x_1, x_2 such that $y(x_1) = y(x_2)$, we can replace both of them by a single value $x_{1,2}$ such that $y(x_{1,2}) = y(x_1)$ and $p(x_{1,2}) = p(x_1) + p(x_2)$. This does not change the structure of the problem, that is, the IB clustering problem discussed in Section 3.1.1 remains the same. This also implies that we may assume without loss of generality that $|\mathcal{X}| = |\mathcal{Y}|$.

In this case, $I(X;\hat{X}) = I(\hat{X};Y)$ and the IB objective function becomes

$$\mathcal{F}_\beta[p] = (1-\beta) I_p(\hat{X};Y)\,.$$

There are three different regimes for β in this case: (i) when $\beta < 1$, the solution is the same

Algorithm 3: Criticality Search

Input: $p(x, y)$, initial $p_0(\hat{x}|x)$, and β_0
Output: Candidate critical points
for $\hat{x} \in \hat{\mathcal{X}}$ **do**
  $p(\hat{x}|x) \leftarrow p_0(\hat{x}|x)$ (initialize)
  $\beta_c(\hat{x}) \leftarrow \beta_0$
  $\lambda_c \leftarrow 0$
  while $\beta_c(\hat{x}) \neq \lambda_c^{-1}$ **do**
    $p(\hat{x}|x) \leftarrow \mathrm{IB}\left(p(x, y), p(\hat{x}|x), \beta_c(\hat{x})\right)$ (update encoder)
    $C_Y \leftarrow B^\top A$ (update C_Y)
    $U, D = \mathrm{EVD}(C_Y)$ (eigendecomposition of C_Y)
    $L \leftarrow \{\lambda_i : \lambda_i = D_{ii}, \forall i = 1, \ldots, n\} \setminus \{1\}$
    if $\exists \lambda \in L$ *such that* $\beta_c(\hat{x}) = \lambda^{-1}$ **then**
      $\lambda_c \leftarrow \lambda$ (found a candidate for $\hat{x}$)
    else if $I_p(\hat{X}; Y) = I(X; Y)$ **then**
      $\beta_c(\hat{x}) \leftarrow \infty$ (no candidates were found for $\hat{x}$)
      continue
    else
      $\beta_c(\hat{x}) \leftarrow \min_{\lambda \in L} \lambda^{-1}$ (educated guess for the next iteration)
return $\beta_c(\hat{x})$, $\forall t \in \hat{\mathcal{X}}$

as in the general case, i.e. it is the trivial solution for which $I(\hat{X}; Y) = 0$; (ii) when $\beta = 1$, $\mathcal{F}_\beta[p] = 0$ for all $p(\hat{x}|x)$, which means that any representation $p(\hat{x}|x)$ would be equally good; (iii) when $\beta > 1$, minimizing $\mathcal{F}_\beta[p]$ becomes equivalent to maximizing $I_p(\hat{X}; Y)$. The solution in this regime is equivalent to the solution when $\beta \to \infty$, and so the optimal representation would be a deterministic mapping from X to $\hat{X}$. Therefore, in this regime, the parameter that shapes the optimal representations is the hard constraint on $|\hat{\mathcal{X}}|$ rather than β.

Because β^{-1} is the slope of the information curve [20], the curve in the deterministic case is linear with slope 1 (or piecewise linear, as noted also in [28], if we relax the assumption that $y(x)$ is a bijective function, in which case the curve becomes flat once $H(Y)$ is reached). We identify, contra to [28], a sequence of structural phase transitions along this line, which are characterized by the solutions to the following optimization problems for $K = 1, \ldots, |\mathcal{X}|$:

$$\begin{aligned} \max_{p} \quad & I_p(\hat{X}; Y) \\ \text{such that} \quad & Supp(p) \leq K \,. \end{aligned}$$

These problems are NP-Hard, although in some cases (e.g., $K = |\hat{\mathcal{X}}|$) they are tractable.

4 Numerical examples

In this section we explore numerically (a) the types of structural phase transitions that may occur in IB; (b) related phenomena such as critical slowing down; and (c) the influence of $p(x,y)$ on the evolutionary trajectory of the representations. We do so by considering several numerical examples that are designed to be as simple as possible and at the same time convey important insight about the evolutionary trajectory of the IB representations.

4.1 Sensitivity to the source distribution

We begin by exploring the influence of the source distribution $p(x)$ on the evolution of the representations. To this end, we fix $p(y|x)$ and vary only $p(x)$. We take $Y \in \{0,1\}$ and trinary X and $\hat{X}$. We define $p(y|x)$ by $p(y=1|x_1)=0.25$, $p(y=1|x_2)=0.48$, and $p(y=1|x_3)=0.75$. The choice of $p(y|x_2)$ is deliberately meant to break the symmetry in this example. The symmetric case will be explored in the next section. We consider four joint distributions defined by $p(y|x)$ and the following source distributions:

$$
\begin{aligned}
p_1(x) &= \begin{pmatrix} 0.45 & 0.1 & 0.45 \end{pmatrix} \\
p_2(x) &= \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \\
p_3(x) &= \begin{pmatrix} 0.18 & 0.64 & 0.18 \end{pmatrix} \\
p_4(x) &= \begin{pmatrix} 0.1 & 0.8 & 0.1 \end{pmatrix} .
\end{aligned}
$$

For each joint distribution $p_i(x,y) = p_i(x)p(y|x)$, we evaluated the evolutionary trajectory of the IB representations via Algorithm 2, the corresponding centroid bifurcation diagram, and the evolution of the second eigenvalue of $C_Y^{\beta,\hat{x}}$ for all $\hat{x}$. The results are shown in Figure 2. It can be seen that in all four cases there are two critical points. At these points, the effective cardinality increases, which is reflected in the emergence of a new distinct value in the centroid bifurcation diagrams (Figure 2A). Note that the effective cardinality corresponds to $Supp(p_\beta)$ only when the representation is canonical. For p_3 and p_4, all the representations found by Algorithm 2 are canonical, and therefore at the critical points $Supp(p_\beta)$ changes. This can be seen in Figure 2B, where $p_\beta(\hat{x})$ becomes positive for some $\hat{x}$.

Figure 2C shows that, as expected based on the theoretical analysis of Section 3, $\lambda_2(\hat{x})$ coincides with β^{-1} at critical points in which centroids splits continuously (e.g., the first phase transition for p_1). Interestingly, Figure 2A reveals that not all phase transitions correspond to continuous centroid splits (e.g., the second phase transition for p_1). However, even in these discontinuous cases $\lambda_2(\hat{x})$ seems to be indicative of the phase transition because it tends to reache β^{-1} at those critical points.

Finally, we observe a *critical slowing down* phenomenon near the phase transitions, in which the convergence time of the IB iterations diverges (Figure 2D). This phenomena has

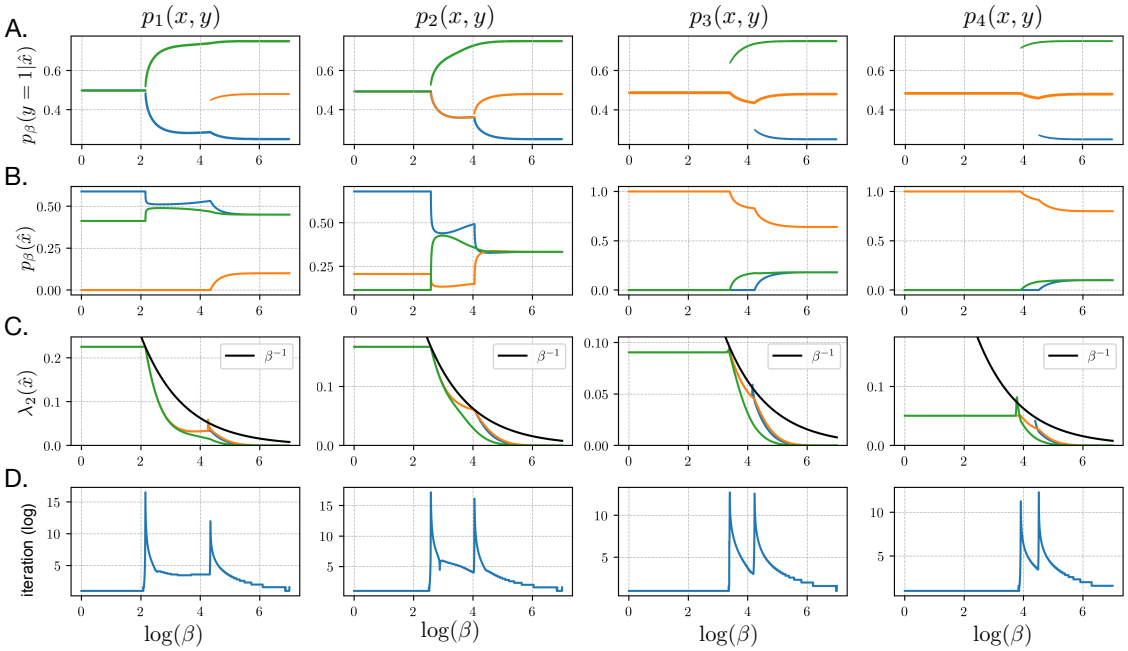

Figure 2. Numerical simulations with *asymmetric* distributions. The i-th column corresponds to the set of results for $p_i(x, y)$. Colored curves (blue, orange, green) in panels A-C correspond to different values of $\hat{x}$. A. Centroid bifurcation diagrams. B. $p_\beta(\hat{x})$ as a function of $\log(\beta)$. C. The evolution of the second eigenvalue $\lambda_2(\hat{x})$ as a function of $\log(\beta)$. D. Log convergence time of Algorithm 1, i.e., the number of IB iterations, as a function of $\log(\beta)$.

been known to happen near phase transitions in other settings [29, 30], and further analysis of this phenomena in the case of IB is left to future research.

This numerical exploration shows that the source distribution may have substantial influence on the location of the critical points, as well as their type. For example, bifurcations that appear as continuous splits, similar to pitchfork bifurcations, may change to what appears as a discontinuous emergence of a new centroid. In addition, our simulations suggest that the IB phase transitions may also be characterized by critical slowing down, in addition to the characterization of Section 3.

4.2 Symmetric distributions

Next, we repeat the same analysis with symmetric distributions. We constructed these distributions by taking the four asymmetric distributions from before and changing $p(y = 1|x_2) = 0.5$. Figure 3 shows the results in this case. Not surprisingly, the bifurcation diagrams are symmetric for these distributions (Figure 3A). In addition, these examples demonstrate that $p(x)$ may influence not only the type of bifurcations but also their number. For p_1 and p_2 there are two critical points, as before, however for p_3 and p_4 there is only one critical point. Furthermore, for p_3 and p_4 we observe a trinary split in which the effective cardinality jumps from $k = 1$ to $k = 3$. This appears to happen either via a continuous split (as in p_3) or via a discontinuous

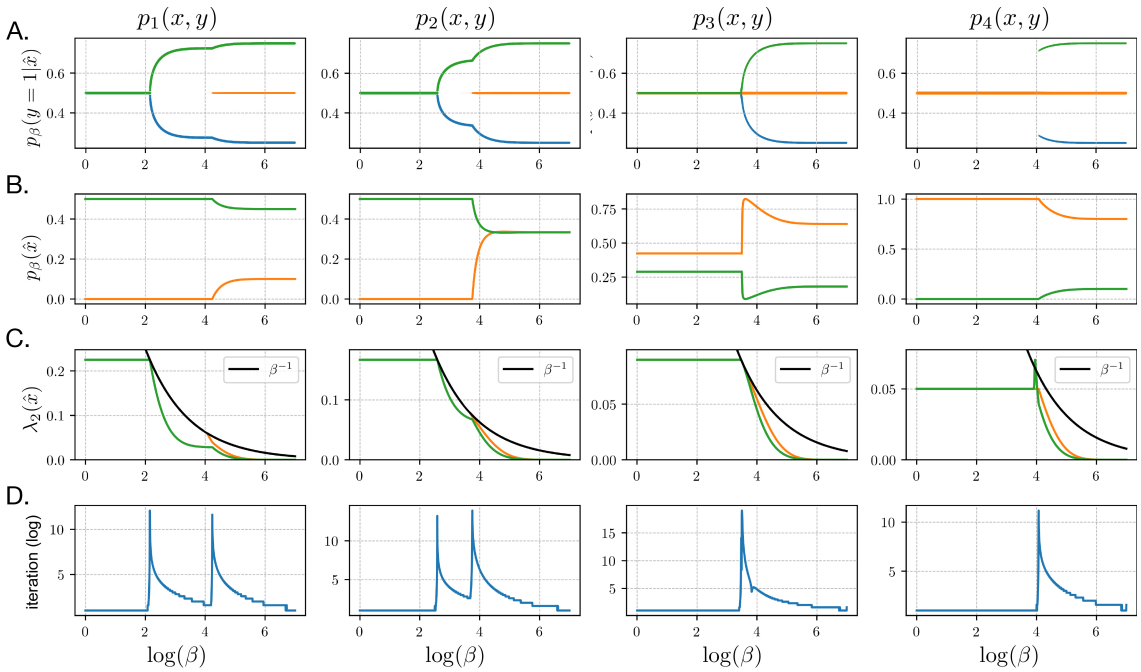

Figure 3. Numerical simulations with *symmetric* distributions. The i-th column corresponds to the set of results for $p_i(x, y)$. Colored curves (blue, orange, green) in panels A-C correspond to different values of $\hat{x}$. In some cases the blue and green curves overlap. A. Centroid bifurcation diagrams. B. $p_\beta(\hat{x})$ as a function of $\log(\beta)$. C. The evolution of the second eigenvalue $\lambda_2(\hat{x})$ as a function of $\log(\beta)$. D. Log convergence time of Algorithm 1, i.e., the number of IB iterations, as a function of $\log(\beta)$.

emergence of a new value (as in p_4). In the continuous case, which corresponds to the assumptions of our criticality condition, $\lambda_2(\hat{x}) = \beta_c^{-1}$ for all three clusters at the same critical point (intersection of the colored curves with the black curve in Figure 3C, p_3). This behavior is less clear is the discontinuous case (Figure 3C, p_4). In both cases, however, we observe critical slowing down near the phase transition (Figure 3D).

4.3 Water filling in Bayesian networks

In this final example, we extend our analysis to the multivariate case and illustrates a potential application of our approach to design principles for neural network architectures. Specifically, we use the methods of Section 3, but instead of the standard IB method we apply its multivariate extension [31, Multivariate IB (MVIB)]. MVIB takes the multi-information, which is defined for a set of random variables $\mathbf{Z} = (Z_1, \ldots, Z_n) \sim p(z_1, \ldots, z_n)$ by

$$\mathcal{I}(\mathbf{Z}) = D\left[p(z_1, \ldots, z_n) \middle\| \prod_{i=1}^{n} p_i(z_i)\right] , \tag{15}$$

as a natural extension of mutual information in the multivariate case. The MVIB objective function is then $\mathcal{I}(\mathbf{X}, \hat{\mathbf{X}}) - \beta\mathcal{I}(\hat{\mathbf{X}}, \mathbf{Y})$, where $\mathbf{X}$, $\mathbf{Y}$ and $\hat{\mathbf{X}}$ are multivariate variables and the statistical dependencies between them are defined by a Bayesian network.

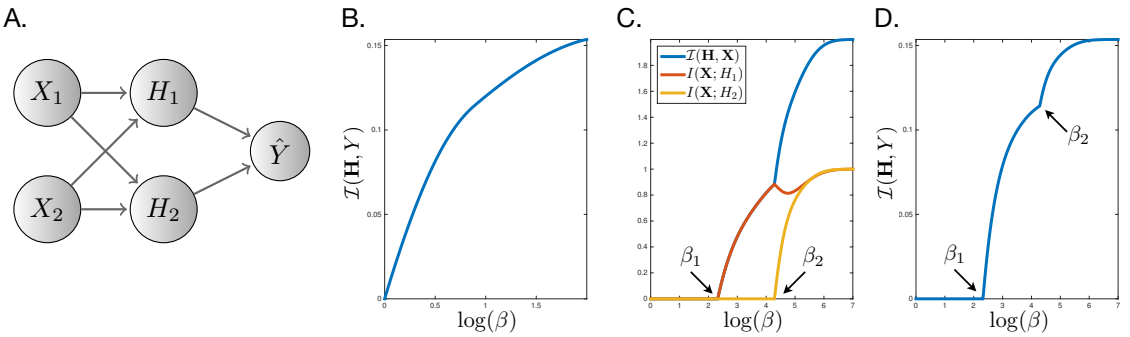

Figure 4. Numerical simulations in the multivariate case. A. The Bayesian network used in our simulations. B. The multivariate information curve. C. Information that the hidden representation maintains about the input. Note that for every β it holds that $\mathcal{I}(\mathbf{X},\hat{\mathbf{X}}) = I(\mathbf{X};H_1) + I(\mathbf{X};H_2)$. D. Information about the ground truth Y extracted by the hidden representation.

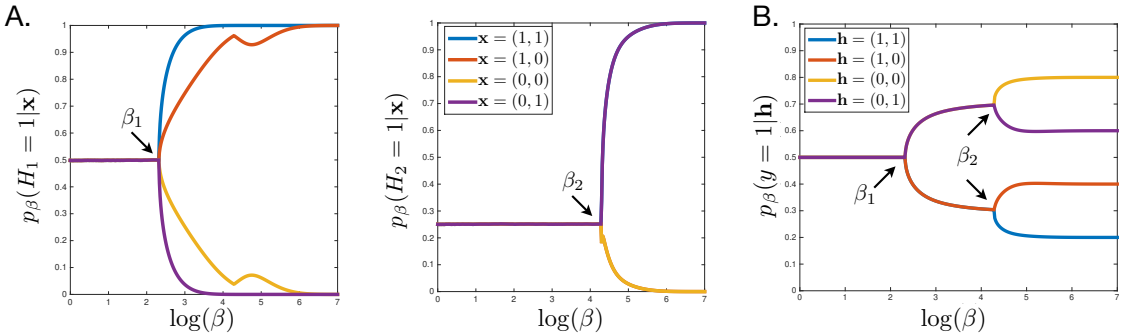

Figure 5. Evolution of the hidden representation. A. Bifurcation diagrams for the H_1 encoder (left) and H_2 encoder (right). B. Centroid bifurcation diagram.

To demonstrate our approach numerically, we consider the Bayesian network of Figure 4A, where $\mathbf{X} = (X_1, X_2)$ is the input, $\hat{\mathbf{X}} = \mathbf{H} = (H_1, H_2)$ is the hidden layer of the network, and $\hat{Y}$ is the network's prediction defined such that $p(\hat{Y} = y|\hat{\mathbf{x}}) = p(Y = y|\hat{\mathbf{x}})$. For simplicity, we assume that all variables — X_1, X_2, H_1, H_2, Y, and $\hat{Y}$ — are binary. We take $p(\mathbf{x})$ to be uniform, and define $p(y|\mathbf{x})$ by $p(y = 1|\mathbf{x} = (0,0)) = 0.8$, $p(y = 1|\mathbf{x} = (0,1)) = 0.6$, $p(y = 1|\mathbf{x} = (1,0)) = 04$, and $p(y = 1|\mathbf{x} = (1,1)) = 0.2$.

Figure 4 shows the multivariate information curve for this example, and the information that the hidden representation maintains about the input $\mathbf{X}$ and the desired output (or ground truth) Y. It is easy to verify that in this case $\mathcal{I}(\mathbf{X},\mathbf{H}) = I(\mathbf{X};H_1) + I(\mathbf{X};H_2)$. Figure 5 gives a more detailed view of the evolution of the hidden representation as a function of β.

These result demonstrate a water filling phenomenon for the hidden units of the networks, analogous to the water-filling phenomena in rate–distortion theory [23]. When $\beta < \beta_1$, both hidden units are independent of the input (Figure 5A), and do not maintain any information about $\mathbf{X}$ or Y (Figure 4C-D). The prediction of the network (Figure 5B) in this regime is based on the prior $p(y)$, which is uniform in this case. This means that the canonical hidden representation is constant, and thus both hidden units are redundant. When $\beta_1 < \beta < \beta_2$, only H_1 keeps

information about the input and output. In this case H_2 is redundant and can be eliminated from the network. When $\beta > \beta_2$, both units are informative, and their contribution is complementary. Namely, H_1 evolves to represent X_1 and H_2 evolves to represent X_2. Therefore, in this regime both units are necessary for the optimal architecture uses both of them.

5 Conclusions

In this work, we have cast the notion of efficient compressed representations in terms of IB, and characterized how these efficient representations evolve via a deterministic annealing process. The main contributions of this work are: (1) introduction of order parameters that capture the evolution of the IB representations and the structural phase transitions that they undergo; (2) derivation of an algorithm for finding critical points; and (3) numerical exploration of the phase transitions and related phenomena that occur in IB. Important directions for future research include an extension of our analysis to continuous variables; characterization of the critical slowing-down phenomenon in IB, and possibly methods for overcoming the computational problem this phenomenon raises. In addition, while the examples we considered here are merely illustrative, they demonstrate general principles that may apply to several fields. For example, some of these methods have already been applied to language evolution [12] and deep neural networks [2, 4]. This work lays out some of the theoretical grounds for extending these applications, as well as applying this approach more broadly.

Acknowledgments

This study was supported by the Gatsby Charitable Foundation. Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing.

References

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck method," in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 1999.

[2] N. Tishby and N. Zaslavsky, "Deep learning and the Information Bottleneck principle," in *IEEE Information Theory Workshop (ITW)*, 2015.

[3] A. Alemi, I. Fischer, J. Dillon, and K. Murphy, "Deep variational Information Bottleneck," in *ICLR*, 2017.

[4] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[5] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *Journal of Machine Learning Research*, vol. 19, no. 50, pp. 1–34, 2018.

[6] O. Shamir, S. Sabato, and N. Tishby, "Learning and generalization with the Information Bottleneck," *Theoretical Computer Science*, vol. 411, no. 29-30, pp. 2696–2711, 2010.

[7] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information Bottleneck for Gaussian variables," *Journal of Machine Learning Research*, vol. 6, 2005.

[8] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, "Predictive information in a sensory population," *Proceedings of the National Academy of Sciences*, vol. 112, no. 22, pp. 6908–6913, 2015.

[9] J. Rubin, N. Ulanovsky, I. Nelken, and N. Tishby, "The representation of prediction error in auditory cortex," *PLOS Computational Biology*, vol. 12, no. 8, pp. 1–28, 2016.

[10] S. Wang, A. Borst, N. Zaslavsky, N. Tishby, and I. Segev, "Efficient encoding of motion is mediated by gap junctions in the fly visual system," *PLOS Computational Biology*, vol. 13, no. 12, pp. 1–22, 2017.

[11] N. Slonim and N. Tishby, "The power of word clusters for text classification," in *23rd European Colloquium on Information Retrieval Research*, 2001.

[12] N. Zaslavsky, C. Kemp, T. Regier, and N. Tishby, "Efficient compression in color naming and its evolution," *Proceedings of the National Academy of Sciences*, vol. 115, no. 31, pp. 7937–7942, 2018.

[13] N. Jacoby, N. Tishby, and D. Tymoczko, "An information theoretic approach to chord categorization and functional harmony," *Journal of New Music Research*, vol. 44, no. 3, pp. 219–244, 2015.

[14] M. Rey and V. Roth, "Meta-Gaussian Information Bottleneck," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1916–1924.

[15] K. Rose, E. Gurewitz, and G. C. Fox, "Statistical mechanics and phase transitions in clustering," *Phys. Rev. Lett.*, vol. 65, pp. 945–948, Aug 1990.

[16] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," in *Proceedings of the IEEE*, 1998, pp. 2210–2239.

[17] T. Gedeon, A. E. Parker, and A. G. Dimitrov, "The mathematical structure of information bottleneck methods," *Entropy*, vol. 14, no. 3, pp. 456–479, 2012.

[18] F. Pereira, N. Tishby, and L. Lee, "Distributional clustering of English words," in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 1993, pp. 183–190.

[19] N. Slonim and N. Tishby, "Document clustering using word clusters via the Information Bottleneck method," in *Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, 2000, pp. 208–215.

[20] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Proceedings of the 16th Annual Conference on Learning Theory (COLT)*, 2003.

[21] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[22] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[23] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley New York, 1991.

[24] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, 1948.

[25] G. Elidan and N. Friedman, "Learning hidden variable networks: The Information Bottleneck approach," *Journal of Machine Learning Research*, vol. 6, pp. 81–127, 2005.

[26] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview press, 1994.

[27] G. Jaeger, "The ehrenfest classification of phase transitions: Introduction and evolution," *Archive for history of exact sciences*, vol. 53, no. 1, pp. 51–81, 1998.

[28] A. Kolchinsky, B. D. Tracey, and S. V. Kuyk, "Caveats for Information Bottleneck in deterministic scenarios," in *International Conference on Learning Representations*, 2019.

[29] D. S. Fisher, "Scaling and critical slowing down in random-field Ising systems," *Physical Review Letter*, vol. 56, pp. 416–419, 1986.

[30] A. A. Middleton, "Critical slowing down in polynomial time algorithms," *Physical Review Letter*, vol. 88, p. 017202, 2001.

[31] N. Slonim, N. Friedman, and N. Tishby, "Multivariate Information Bottleneck," *Neural Computation*, vol. 18, no. 8, pp. 1739–1789, 2006.

Appendix

In this section we prove several technical lemmas that were used in our main analysis.

Lemma 1. *Let $Y - X - \hat{X}$ be a Markov chain such that $p(y, x, \hat{x}) = p(x, y)p(\hat{x}|x)$, and let $p(y|\hat{x})$ be the corresponding conditional distribution of Y given $\hat{X}$. Then*

$$\mathop{\mathbb{E}}_{x,\hat{x}} \left[D[p(y|x) \| p(y|\hat{x})] \right] = I(X;Y) - I(\hat{X};Y) \,.$$

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_{x,\hat{x}} \left[D[p(y|x) \| p(y|\hat{x})] \right] &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x) \left[D[p(y|x) \| p(y|\hat{x})] \right] \\
&= \sum_{x,\hat{x},y} p(x)p(\hat{x}|x) \log \frac{p(y|x)p(y)}{p(y|\hat{x})p(y)} \\
&= \sum_{x,\hat{x},y} p(x)p(\hat{x}|x) \log \frac{p(y|x)}{p(y)} - \sum_{x,\hat{x},y} p(x)p(\hat{x}|x) \log \frac{p(y|\hat{x})}{p(y)} \\
&= I(X;Y) - I(\hat{X};Y) \,.
\end{aligned}$$

□

Lemma 2. $\frac{\partial}{\partial \beta} \mathbb{E}_x \left[\log Z_\beta(x) \right] = I_\beta(\hat{X};Y) - I(X;Y)$.

Proof.

$$\begin{aligned}
\frac{\partial}{\partial \beta} \mathop{\mathbb{E}}_{x} \left[\log Z_\beta(x) \right] &= \sum_x p(x) \frac{\partial}{\partial \beta} \log Z_\beta(x) \\
&= \sum_x p(x) \frac{1}{Z_\beta(x)} \frac{\partial}{\partial \beta} \left(\sum_{\hat{x}} p_\beta(\hat{x}) e^{-\beta D[p(y|x) \| p_\beta(y|\hat{x})]} \right) \\
&= \sum_{x,\hat{x}} p(x) \frac{p_\beta(\hat{x}) e^{-\beta D[p(y|x) \| p_\beta(y|\hat{x})]}}{Z_\beta(x)} \left(\frac{\partial}{\partial \beta} \log p_\beta(\hat{x}) - \beta \frac{\partial}{\partial \beta} D[p(y|x) \| p_\beta(y|\hat{x})] \right) \\
&\quad - \sum_{x,\hat{x}} p(x) \frac{p_\beta(\hat{x}) e^{-\beta D[p(y|x) \| p_\beta(y|\hat{x})]}}{D[p(y|x) \| p_\beta(y|\hat{x})]} \\
&= \sum_{x,\hat{x}} p(x) p_\beta(\hat{x}|x) \left(\frac{\partial}{\partial \beta} \log p_\beta(\hat{x}) - \beta \frac{\partial}{\partial \beta} D[p(y|x) \| p_\beta(y|\hat{x})] \right) \\
&\quad - \sum_{x,\hat{x}} p(x) p_\beta(\hat{x}|x) D[p(y|x) \| p_\beta(y|\hat{x})] \,.
\end{aligned}$$

The first term is zero (assuming $p_\beta(\hat{x})$ is differentiable w.r.t. β) because

$$\mathop{\mathbb{E}}_{x,\hat{x}} \left[\frac{\partial}{\partial \beta} \log p_\beta(\hat{x}) \right] = \sum_{\hat{x}} p_\beta(\hat{x}) \frac{\partial}{\partial \beta} \log p_\beta(\hat{x}) = \sum_{\hat{x}} \frac{\partial}{\partial \beta} p_\beta(\hat{x}) = 0 \,,$$

and so is the second term, for similar reasons:

$$\begin{aligned}\mathop{\mathbb{E}}_{x,\hat{x}}\left[\frac{\partial}{\partial\beta}D[p(y|x)\|p_\beta(y|\hat{x})]\right] &= -\sum_{x,\hat{x},y}p(x)p_\beta(\hat{x}|x)p(y|x)\frac{\partial}{\partial\beta}\log p_\beta(y|\hat{x}) \\ &= -\sum_{\hat{x}}p_\beta(\hat{x})\sum_y\left(\sum_x p_\beta(x|\hat{x})p(y|x)\right)\frac{\partial}{\partial\beta}\log p_\beta(y|\hat{x}) \\ &= -\sum_{\hat{x}}p_\beta(\hat{x})\sum_y p_\beta(y|\hat{x})\frac{\partial}{\partial\beta}\log p_\beta(y|\hat{x}) = 0\,.\end{aligned}$$

It follows that

$$\frac{\partial}{\partial\beta}\mathop{\mathbb{E}}_x\left[\log Z_\beta(x)\right] = -\mathop{\mathbb{E}}_{x,\hat{x}}\left[D[p(y|x)\|p_\beta(y|\hat{x})]\right]\,,$$

and applying Lemma 1 to the right hand side of this equation concludes the proof.

□

Lemma 3. *Let p_β be a canonical IB representation and $\hat{x} \in Supp(p_\beta)$, then*

$$I_X^\beta(\hat{x}) = \beta I_Y^\beta(\hat{x}) - \sum_x p_\beta(x|\hat{x})\left[\log Z_\beta(x) + \beta I_Y(x)\right]\,, \tag{16}$$

where $I_Y(x) = D[p(y|x)\|p(y)]$.

Proof. This follows from substituting (2) in the definition of $I_X^\beta(\hat{x})$, i.e.

$$\begin{aligned}I_X^\beta(\hat{x}) &= \sum_x p_\beta(x|\hat{x})\left(-\beta D[p(y|x)\|p_\beta(y|\hat{x})] - \log Z_\beta(x)\right) \\ &= \sum_x p_\beta(x|\hat{x})\left[\beta\left(\sum_y p(y|x)\log\frac{p_\beta(y|\hat{x})}{p(y)} - I_Y(x)\right) - \log Z_\beta(x)\right] \\ &= \beta\sum_{x,y}p_\beta(x|\hat{x})p(y|x)\log\frac{p_\beta(y|\hat{x})}{p(y)} - \sum_x p_\beta(x|\hat{x})\left[\beta I_Y(x) + \log Z_\beta(x)\right] \\ &= \beta I_Y^\beta(\hat{x}) - \sum_x p_\beta(x|\hat{x})\left[\beta I_Y(x) + \log Z_\beta(x)\right]\,.\end{aligned}$$

□

Lemma 4. *Let p_β be a canonical IB representation and $\hat{x} \in Supp(p_\beta)$, then*

$$\frac{\partial}{\partial\beta}I_X^\beta(\hat{x}) = \beta\frac{\partial}{\partial\beta}I_Y^\beta(\hat{x}) + g_\beta(\hat{x})\,,$$

where

$$g_\beta(\hat{x}) = I_Y^\beta(\hat{x}) - \frac{\partial}{\partial\beta}\left(\sum_x p_\beta(x|\hat{x})\left[\log Z_\beta(x) + \beta I_Y(x)\right]\right)$$

and $\mathbb{E}_{\hat{x}}\left[g_\beta(\hat{x})\right] = 0$.

Proof. The first part follows from differentiating (16) with respect to β. For the second part, notice that Proposition 2 implies that

$$\mathop{\mathbb{E}}_{\hat{x}}\left[\frac{\partial}{\partial\beta}I_X^\beta(\hat{x})\right] = \beta \mathop{\mathbb{E}}_{\hat{x}}\left[\frac{\partial}{\partial\beta}I_Y^\beta(\hat{x})\right] ,$$

and therefore $\mathbb{E}_{\hat{x}}\left[g_\beta(\hat{x})\right] = 0$.

□