Critical Slowing Down Near Topological Transitions in Rate-Distortion Problems

Shlomi Agmon∗†, Etam Benger ∗†, Or Ordentlich†, and Naftali Tishby†‡
†School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel
‡Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel
Email: {shlomi.agmon, etam.benger, or.ordentlich, naftali.tishby}@mail.huji.ac.il

Abstract—In rate-distortion (RD) problems one seeks reduced representations of a source that meet a target distortion constraint. Such optimal representations undergo topological transitions at some critical rate values, when their cardinality or dimensionality change. We study the convergence time of the Arimoto-Blahut alternating projection algorithms, used to solve such problems, near those critical points, both for the rate-distortion and information bottleneck settings. We argue that they suffer from critical slowing down—a diverging number of iterations for convergence—near the critical points. This phenomenon can have theoretical and practical implications for both machine learning and data compression problems.

I. INTRODUCTION

Given a source \( X \sim p(x) \) on a finite alphabet \( \mathcal{X} \), a representation alphabet \( \hat{\mathcal{X}} \), and a distortion measure \( d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}^+ \), the rate-distortion function (RDF) is defined as \( R(D) = \min I(X; \hat{X}) \), where the minimization is with respect to all test channels \( p(\hat{x}|x) \) satisfying the distortion constraint \( \mathbb{E}[d(X, \hat{X})] \leq D \). The distortion-rate function \( D(R) \) is merely the inverse of \( R(D) \). An analytic expression for \( R(D) \) (or \( D(R) \)) involves solving the minimization above, and is only known for some special cases. However, it is possible to obtain a numerical solution using different algorithms, including the Arimoto-Blahut (AB) algorithm [3], [4].

Clearly, for \( R = 0 \), the test channel that maps all \( x \in \mathcal{X} \) to \( \text{argmin}_{\hat{x}} \mathbb{E}[d(X, \hat{x})] \) is optimal, whereas for \( R = H(X) \) the channel that maps any \( x \in \mathcal{X} \) to \( \text{argmin}_{\hat{x}} d(x, \hat{x}) \) is optimal. Thus, the cardinality (support size) of the optimal \( \hat{X} \) attaining \( D(R) \) changes as we increase/decrease \( R \). Typically, the cardinality of the optimal \( \hat{X} \) decreases gradually from \( |\hat{\mathcal{X}}| = |\mathcal{X}| \) to \( |\hat{\mathcal{X}}| = 1 \) as we decrease \( R \) from \( H(X) \) to 0, but there are also examples where the cardinality of the optimal \( \hat{X} \) behaves non-monotonically [1, Section 2.7]. We refer to these changes of the representation cardinality as topological or phase transitions and the values of \( R \) where they occur as critical. This paper studies the algorithmic difficulty of computing \( D(R) \) near such critical values of \( R \).

In the context of data compression, the main quantity of interest is the representatives’ distribution, \( p(\hat{x}) \), at a given value of \( R \), for which an optimal code is constructed and the convergence to the optimal \( D(R) \) can be obtained, as the blocklength increases. In applications of RD to machine learning and statistical physics, however, there is more interest in the nature of the optimal channel, or representation encoder, \( p(\hat{x}|x) \). The reason is that in machine learning the similar features of the source patterns \( x \) that are mapped to specific representations \( \hat{x} \) determine the relevant order parameters and topology of the problem. At the critical points, the neighborhoods of patterns can merge or split during the learning process, and the induced topology of the patterns—which patterns are neighbours—can significantly alter. Understanding such topological changes and the nature of the encoder is critical in representation learning [5] and has gained recent interest also in statistical physics [6].

Similar topological transitions occur also in the closely related approach of the information bottleneck (IB) [7], which aims to achieve maximal compression of \( X \) while preserving most relevant information about another correlated variable \( Y \). This approach has recently drawn attention due to its possible relation to the learning dynamics of deep neural networks (e.g. [8]–[10]). Specifically, there is evidence to suggest that the representations of the layers in such networks converge to successively refineable points near the IB curve, which may be related to the critical points where such transitions occur [9], [11]. Moreover, the critical points represent changes in the nature of the optimal solutions to the RD or IB problems, when considering the size of the representation alphabet as an additional constraint.

In this work we show that solutions to RD problems lose their stability at the critical points. As a result, we prove that the AB algorithm slows down dramatically near such critical points. This phenomenon, in which systems’ dynamics slow down near phase transitions, is known in statistical physics as critical slowing down (CSD). Finally, we show that similar slowing down occurs also for the extended AB algorithm, used to solve IB problems numerically, near critical points.

II. RELATED WORK

The convergence of the AB algorithm for finite/countable reconstruction alphabets was established in [3], [4], [12]. Bouchris [13] further derived an upper bound on the convergence rate, which shows that the gap between the value of the Lagrangian defined below in (1) under the AB output and the optimal solution decreases at least inversely proportional to the number of iterations. Several papers have analyzed the convergence rate of the AB algorithm for capacity computation,
and have demonstrated that the algorithm converges exponentially fast whenever the support of a capacity achieving input distribution is full [14]–[16].

Phase transitions in the optimal test-channel attaining $D(R)$, as $R$ changes, were already discussed by Berger [1]. In fact, Berger also showed that if the support of the optimal $X$ is known, the computation of the optimal test-channel simplifies. In a sense, this already shows that finding an optimal test-channel at critical points, where the support of the optimal solution changes, can be computationally challenging. Two decades later, Rose [17] demonstrated that even when the reconstruction alphabet is continuous, the optimal $X$ typically has finite support, which grows with $R$.

The extended version of the AB algorithm for the IB problem (henceforth, the IB algorithm), in which the representors are optimized as well, was proven in [7] to converge, although not necessarily to a unique minimum as the convexity is lost. In [18], the authors address the difficulty to identify the topological transitions in the IB framework using the method of deterministic annealing [19]. To solve this difficulty, Parker et al. [20], [21] study the bifurcation structure of solutions to the IB and other RD-like problems using bifurcation theory. They focus mainly on the first critical point – where they argue that the trivial solution ($\hat{x} = 1$) loses its stability and structure begins to emerge. More recently, phase transitions in the IB have been studied from a representation learning perspective, shown to relate to learning of new features in the data [22], [23], and algorithms for finding the critical points were presented [24]. Analytical expression for the location of the critical points is known in the Gaussian IB case [11], [25], where the topological transitions correspond to changes in the dimension of the Gaussian distribution of $X$.

### III. CRITICAL POINTS IN RATE-DISTORTION THEORY

The constrained optimization problem of RD is solved by introducing a positive Lagrange multiplier, $\beta$, to impose the expected distortion constraint, and minimizing the Lagrangian [2]

\[
I(X; X) + \beta \mathbb{E}[d(X, \hat{X})],
\]

with respect to all test channels $p(\hat{x}|x)$. The Lagrange multiplier $\beta$, of a role similar to that of inverse temperature in statistical physics, determines the topological structure of the optimizing channel as well as the trade-off between compression and distortion. A solution to the RD problem at a given value of $\beta$, denoted by $p_\beta(\hat{x}|x)$, that is, a minimizer of (1), must (self-consistently) satisfy both equations [1]–[3]:

\[
p_\beta(\hat{x}|x) = \frac{p_\beta(\hat{x})e^{-\beta d(\hat{x}, \hat{x})}}{Z(x, \beta)}
\]

and

\[
p_\beta(\hat{x}) = \sum_x p_\beta(\hat{x}|x)p(x),
\]

where $Z(x, \beta) := \sum_x p_\beta(\hat{x})e^{-\beta d(\hat{x}, \hat{x})}$ is a normalization (partition) function. Since $p_\beta(\hat{x}|x)$ is determined by $p_\beta(\hat{x})$ when $\beta$ and the distortion function are given, we may consider $p_\beta(\hat{x})$ as the optimization variable. Moreover, to simplify the

![Graph](https://example.com/graph.png)

Fig. 1. A simple RD problem: the sets $X, \hat{X} \in \mathbb{R}^2$ shown in orange and purple, correspondingly; $p(x) = (0.4, 0.3, 0.2, 0.1)\top$; the distortion function is defined as $d(x, \hat{x}) = \frac{1}{2}\|x - \hat{x}\|^2_2$, where $\mu = \max_{x, \hat{x} \in X, \hat{X}} \|x - \hat{x}\|^2_2$ is a normalization factor. B. Values of $p_\beta(\hat{x})$, the solutions to the RD problem at different values of $\beta$; note the three phase transitions at $\beta_1, \beta_2, \beta_3$ (marked with dashed lines) – until $\beta_1$ only $p(\hat{x}_4)$ > 0, then at $\beta_1$ the representor $\hat{x}_2$ starts to gain mass, and at $\beta_2$ also $\hat{x}_3$, finally $\hat{x}_1$ starts to gain mass at $\beta_3$. C. Eigenvalues of $A$ at $p_\beta$: note that near each of the phase transitions an eigenvalue of $A$ approaches 0. D. Number of iterations until convergence ($\epsilon = 10^{-9}$, logarithmic scale) using uniform initial conditions and reverse annealing: a slowing down of approximately an order of magnitude is clearly noticed at each critical point. E. The relation between the number of iterations until convergence and $\lambda_0$, the smallest nonzero eigenvalue of $A$, computed using reverse annealing with various values of $\epsilon$; as $\epsilon$ decreases, the relation approaches the limit formula in Theorem 5 (diagonal line); the artifact at the center, consisting of a few vertically arranged points, corresponds to the slowing down to the left of $\beta_1$ (we do not fully understand this phenomenon, which is more prominent in the uniform initial conditions setting).
discussion, we assume throughout that \( p_\beta(\hat{x}) \) is unique for all values of \( \beta \), and consider it as a vector \( p_\beta \) in the sequel.

**Definition 1:** Denote the support of a solution \( p_\beta \) by \( \text{supp} p_\beta = \{ \hat{x} \in X : p_\beta(\hat{x}) > 0 \} \). We say that \( \beta_c \) is critical and that the RD problem has a topological transition at \( \beta_c \) if

\[
| \text{supp} p_{\beta^-} | \neq | \text{supp} p_{\beta^+} |
\]

(4)

for all \( \beta^- < \beta_c < \beta^+ \) in some (small) neighborhood of \( \beta_c \).

We restrict our discussion to local stability analysis, and thus assume that \( \beta \rightarrow p_\beta \) is continuous. As a result, we tackle only what is known in statistical physics as phase transitions of second order or higher.

Equations (2) and (3) above can be written concisely as

\[
F(p(\beta))_x := p(\hat{x}) - \sum_x p(x)p(\hat{x})e^{-\beta d(x,\hat{x})} Z(x,\beta) .
\]

(5)

In what follows, we study the properties of \( F \) and its Jacobian \( \nabla F = \partial F / \partial p(\hat{x}) \) around critical points, and provide a characterization of the phase transitions of RD problems in terms of \( \nabla F \). This characterization will play a pivotal role, as we show next that \( \nabla F \) is closely related to the Jacobian of a single step of the Arimoto-Blahut algorithm and its spectrum governs the algorithm’s convergence rate.

Assume there is a topological transition at \( \beta_c \), such that \( | \text{supp} p_{\beta^-} | < | \text{supp} p_{\beta^+} | \) in the above notation, then there exists \( \hat{x}_0 \) such that \( p_{\beta^-}(\hat{x}_0) = 0 \) and \( p_{\beta^+}(\hat{x}_0) > 0 \). Consider the representation space \( \hat{X} = X \setminus \{ \hat{x}_0 \} \) and the distortion function \( d' \), which is the restriction of \( d \) to \( \hat{X} \). Clearly, the solution \( p_{\beta} \) of the RD problem defined by \( d' \) is identical to \( p_{\beta} \) (restricted to \( \hat{X} \)) at a left neighborhood of \( \beta_c \), but they must differ to the right. Nevertheless, the extension of \( q_{\beta} \) over the original problem satisfies \( F(q_{\beta}, \beta) = 0 \) also to the right of \( \beta_c \), if one sets \( q_{\beta}(\hat{x}_0) = 0 \). While this would not be a solution to the RD problem, it shows the number of solutions to \( F = 0 \) changes at critical values of \( \beta \), or that \( F = 0 \) has a bifurcation at \( \beta_c \).

Although the extension of \( q_{\beta} \) is a fixed point of the AB algorithm, it is not a stable solution, as adding a small perturbation to \( q_{\beta}(\hat{x}_0) \) leads to convergence to the optimal solution \( p_{\beta} \) instead. Hence, at \( \beta_c \) the existing solution to the RD problem loses its stability with respect to the AB algorithm, and a new stable solution emerges.

Notice that \( \nabla F \) must be singular at critical values of \( \beta \). Otherwise, by the implicit function theorem, there must exist a unique solution to \( F = 0 \) as a function of \( \beta \) in the vicinity of \( \beta_c \) [26]. Unfortunately, \( \nabla F \) is trivially singular when \( p_{\beta}(\hat{x}) = 0 \) for some representing \( \hat{x} \), so this is not a useful characterization of the critical points. However, we show next that the Jacobian becomes “more singular” with each transition.

Let \( A \) denote the transposed Jacobian matrix of \( F \) at a solution \( p_{\beta} \), then [27, Appendix A]

\[
A_{\hat{x}\hat{x}'} := [(\nabla F)^T]_{\hat{x}\hat{x}'} = \sum_x p_{\beta}(\hat{x}')p_{\beta}(x)\hat{x}
\]

(6)

\[
= p_{\beta}(\hat{x}') \sum_x p(x)e^{-\beta(d(x,\hat{x})+d(x,\hat{x}'))} Z(x,\beta)^2 ,
\]

(7)

and we have the following result:

**Theorem 1:** Let \( m = |\hat{X}| \) and \( k = |\text{supp} p_{\beta}| \), then at \( p_{\beta} \)

\[
\dim \ker A = m - k.
\]

**Corollary 1:** Topological transitions of a RD problem occur exactly at values of \( \beta \) where the dimension of \( \ker A \), at the solutions \( p_{\beta} \), changes.

The proof consists of two steps [27, Appendix B]. First, we show that \( \dim \ker A \geq m - k \):

**Lemma 2:** Let \( e_x \in \mathbb{R}_m \) denote the standard basis vector with 1 at the \( x \) coordinate and 0 elsewhere, then \( p_{\beta}(\hat{x}) = 0 \iff e_x \in \ker A \).

Second, assume for simplicity that \( d \) is finite and without loss of generality \( d(\cdot, \hat{x}_1) \neq d(\cdot, \hat{x}_2) \) for all \( \hat{x}_1 \neq \hat{x}_2 \). Then there is always a standard basis of \( \ker A \), resulting in the converse inequality \( \dim \ker A \leq m - k \):

**Proposition 3:** Let \( v \in \ker A \), such that exactly \( r \geq 1 \) of its coordinates, denoted \( \hat{x}_1, \ldots, \hat{x}_r \), are nonzero; then all the corresponding standard basis vectors \( e_{\hat{x}_i} \), for \( 1 \leq i \leq r \), belong to \( \ker A \).

**Theorem 1** refers to the geometric multiplicity of the eigenvalue 0 of \( A \), and shows that it changes at critical points. However, to ensure that near such points \( A \) must have a small positive eigenvalue we need to establish a similar statement for the algebraic multiplicity. This will follow as a corollary of the next theorem [27, Appendix C]:

**Theorem 4:** The matrix \( A \) is diagonalizable with real non-negative eigenvalues.

Together with **Theorem 1** above, we obtain:

**Corollary 4.1:** Let \( m = |\hat{X}| \) and \( k = |\text{supp} p_{\beta}| \), then at \( p_{\beta} \)

the algebraic multiplicity of the eigenvalue 0 of \( A \) is exactly \( m - k \). Consequently, topological transitions of a RD problem occur exactly at values of \( \beta \) where the algebraic multiplicity of the eigenvalue 0 of \( A \) changes.

Figure 1 demonstrates this result on a simple RD problem, consisting of 4 points in \( \mathbb{R}^2 \) (Figure 1A). The solutions to the problem undergo 3 transitions, at \( \beta_1, \beta_2 \) and \( \beta_3 \), where the cardinality of \( \hat{X} \) increases from 1 (trivial solution) to 2, 3 and 4, correspondingly (Figure 1B). Figure 1C shows that at each critical \( \beta \) another eigenvalue of \( A \) reaches 0.

Finally, as the matrix \( A \) is row-stochastic, all its eigenvalues are inside the unit circle [28], and by **Theorem 4**, they are in \( [0, 1] \), as required in the next section.

**IV. SLOWING DOWN OF ARIMOTO-BLAHUT**

The numerical computation of solutions to RD problems is usually performed using the Arimoto-Blahut (AB) algorithm [3], [4]. It consists of an alternating minimization, applying Equations (2) and (3) repeatedly, starting from some initial distribution \( p_0 \) in the interior of the simplex \( \Delta \hat{X} \). The \( k \)-th iteration \( p_k \) of the AB algorithm is said to have \( \varepsilon \)-converged to a RD solution \( p_\beta \) if

\[
\|p_k - p_\beta\| < \varepsilon .
\]

(8)

Given a value of \( \beta \), define the operator \( AB : \Delta \hat{X} \rightarrow \Delta \hat{X} \)

to be the result of applying a single step of the AB algorithm.

\[1\]The exact norm used at (8) is of little importance, as \( |\hat{X}| \) is finite and we are typically interested in small values of \( \varepsilon \). For convenience, we have chosen \( L^1 \) in **Theorem 5** below, and \( L^\infty \) in the Figures 1 and 2.
Solutions to the RD problem are fixed points of the algorithm, that is \( AB p_\beta = p_\beta \) or \((I - AB) p_\beta = 0\), where \( I \) is the identity operator. Using (5) it follows that at the solutions of the RD problem \( AB = I - F \), hence \( \nabla AB |_{p_\beta} = I - A^\top \), where \( \nabla AB |_{p_\beta} \) is the Jacobian matrix of \( AB \) at \( p_\beta \).

Let \( \delta p_k = p_k - p_\beta \) be the deviation from \( p_\beta \), then to first order in \( \delta p_k \),
\[
AB p_k \approx p_\beta + \nabla AB |_{p_\beta} \delta p_k \tag{9}
\]
\[
\Rightarrow \delta p_{k+1} = AB p_k - p_\beta \approx (I - A^\top) \delta p_k .
\]

Hence,
\[
\delta p_k \approx (I - A^\top)^k \delta p_0 . \tag{10}
\]

As a result, the convergence rate of the algorithm is governed by the largest eigenvalue of \( I - A^\top \) inside the unit circle, denoted \( \lambda_{\text{max}} \). If the initial deviation \( \delta p_0 \) has a nonzero component in the eigenspace of \( \lambda_{\text{max}} \), then
\[
\| \delta p_k \| < \varepsilon \quad \Rightarrow \quad k \approx -\frac{\log \varepsilon + \text{const}}{-\log |\lambda_{\text{max}}|} . \tag{11}
\]

For the asymptotic convergence rate we consider
\[
\lim_{\varepsilon \to 0} \frac{k}{-\log \varepsilon}; \quad \text{to avoid dependence on the particular choice of initial conditions, via the constant at (11).}
\]

This argument is made precise by the next theorem, proven in [27, Appendix D]. Recall that \( A \) is diagonalizable (Theorem 4) and its eigenvalues are in \([0, 1]\). Therefore, \( \nabla AB = I - A^\top \) is also diagonalizable and its eigenvalues are in \([0, 1]\). When \( p_\beta \) has a full support, the eigenvalues of \( \nabla AB \) are in \([0, 1]\) (Theorem 1). Consequently, we have \( \lambda_{\text{max}} = 1 - \lambda_0 < 1 \), where \( \lambda_0 > 0 \) is the smallest nonzero eigenvalue of \( A \).

**Theorem 5:** Let \( p_\beta \) be a RD solution with \( p_\beta(\hat{x}) > 0 \) for all \( \hat{x} \), and \( \lambda_{\text{max}} = 1 - \lambda_0 < 1 \) the largest eigenvalue smaller than 1 of \( \nabla AB \) at \( p_\beta \). Denote by \( k(p_\beta, \varepsilon) \) the number of iterations required for an initial distribution \( p_0 \) to \( \varepsilon \)-converge to \( p_\beta \), and define \( B(\delta) := \{ p \in \Delta \hat{X} : \| p - p_\beta \|_1 \leq \delta \} \). Then, for any \( a > 0 \),
\[
\Pr_{p_0 \sim U(S)} \left( \lim_{\varepsilon \to 0^+} \frac{k(p_\beta, \varepsilon)}{-\log \varepsilon} - \frac{1}{-\log \lambda_{\text{max}}} < a \right) \to 1 , \tag{12}
\]

where \( U(S) \) denotes the uniform distribution on \( S \).

A lower bound cannot be expected to hold for every initial condition in the vicinity of \( p_\beta \). Indeed, even if the linearization in (9) were exact, the lower bound (11) holds for all but a zero-measure set of initial conditions – those with no component in the \( \lambda_{\text{max}} \) eigenspace. In the general case, where the dynamics are nonlinear, the fraction of applicable initial conditions increases in the vicinity of \( p_\beta \) as \( \delta \to 0 \), (12). The rate at which this fraction approaches 1 depends on the desired accuracy \( a \).

Consider a topological transition of the RD problem at \( \beta_c \), such that \( |\text{supp } p_\beta | < |\text{supp } p_{\beta^+}| \). According to Corollary 4.1, the algebraic multiplicity of the eigenvalue 0 of \( A \) is greater at \( \beta^- \) than at \( \beta^+ \). Since \( A \) is continuous in \( p_\beta \) and \( p_\beta \) is assumed continuous in \( \beta \), there exists an eigenvalue \( \lambda_0 > 0 \) of \( A \) such that \( \lambda_0 \to 0 \) as \( \beta \) approaches \( \beta_c \) from above. When coordinates \( \hat{x} \) outside \( \text{supp } p_{\beta^+} \) are initialized at 0, as in reverse annealing (see below), AB effectively coincides with its restriction to \( \text{supp } p_{\beta^+} \). Consequently, by Theorem 5, the AB algorithm experiences a significant slowing down as it gets closer to the critical point.

This is clearly observed in simulations (Figure 1D) where the number of iterations until convergence is plotted in two different settings: reverse annealing and uniform initial conditions.\(^2\) \( p_0(\hat{x}) = 1/|\hat{x}| \). In reverse annealing the algorithm is run for decreasing values of \( \beta \), starting from the converged solution at the previous \( \beta \). In both settings there is a noticeable slowing down to the right of the critical points, as expected from Theorem 5, given the small nonzero eigenvalues of \( A \) in those areas (Figure 1C).

In addition, it can be seen that the reverse annealing method always converges faster than uniform initial distributions. In the latter, there is an overall increase in the iterations baseline as \( \beta \) decreases, since more eigenvalues reach 0.

Finally, Figure 1E shows that the number of iterations required to converge to a solution within accuracy \( \varepsilon \) gets closer to the bound in Theorem 5 as \( \varepsilon \) decreases. Smaller \( \varepsilon \) are needed when approaching a topological transition, as can be seen by examining the second order terms (the details are omitted).

**V. Critical Slowing Down in the IB Framework**

Given a pair of random variables \((X, Y) \sim p(x, y)\), such that \( I(X; Y) > 0 \), the information bottleneck approach aims to find a channel \( p(\hat{x}|x) \) that minimizes \( I(\hat{X}; X) \), while preserving as much information \( I(\hat{X}; Y) \) as possible [7]. While the IB problem can be viewed as a noisy source coding problem where the reconstruction alphabet is \( \Delta \hat{X} \), the AB algorithm does not directly apply, since \( \hat{X} \) is continuous. Nevertheless, it is known [29] that for each \( \beta \), taking at most \(|X| \) points of the simplex \( \Delta \hat{X} \) suffices. If one were given those points of the simplex, the problem would be reduced to a standard RD problem on a finite reconstruction alphabet. However, since those are not known \( a \text{ priori} \), the IB problem can be thought of as an envelope of many different RD problems, making the optimization problem non-convex [10].

As in RD, the constrained optimization problem of the IB is solved by minimizing a Lagrangian,
\[
I(X; \hat{X}) - \beta I(\hat{X}; Y) , \tag{13}
\]
over all channels from \( X \) to \( \hat{X} = \{1, \ldots, |X|\} \). Consequently, solutions to the IB problem follow a set of self consistent equations, similar to those of RD (Equations (3) and (2)) with the addition of the decoder equation
\[
p_{\beta}(y|\hat{x}) = \sum_x p(y|x)p(\hat{x}|x)p_\beta(\hat{x}) , \tag{14}
\]
and the distortion function in (2) is given by \( d_{IB}(x, \hat{x}) = D_{KL}[p(y|x)||p_{\beta}(y|\hat{x})] \). Notice that this distortion depends (indirectly) on the encoder distribution and on \( \beta \). Moreover, here \( p_{\beta}(\hat{x}) \) does not capture the entire solution, as in RD, which must be described by \( p_{\beta}(\hat{x}|x) \). As a result, the analysis of the IB topological transitions is somewhat more complicated.

\(^2\) Other similar choices in the interior of \( \Delta \hat{X} \), e.g. sampling according to the symmetric Dirichlet distribution \( p_0 \sim \text{Dir}(1) \), do not yield significantly different results.
In addition, the transitions in the IB framework do not consist only of changes in the size of the support, \( \text{supp} p_\beta \). In fact, various representors may share the same decoder distribution \( p_\beta(y|x) \), making them essentially equivalent. The relevant quantity that changes in topological transitions of the IB is the effective cardinality [24], defined as the number of different non-empty decoder distributions, \( \{ p_\beta(y|x) : p_\beta(x) > 0 \} \).

Let \( q_\beta \) be the solution to the RD problem defined by \( \hat{X}^+ \) and \( d^* \). Since \( p_\beta \) is optimal at \( \beta^- \), being the IB solution there, we know that \( \text{supp} q_{\beta^-} = \hat{X}^- \); similarly \( \text{supp} q_{\beta^+} = \hat{X}^+ \). Notice that if the IB problem undergoes a transition at \( \beta^* \), then the effective cardinality of \( p_{\beta^-} \) differs from that of \( p_{\beta^*} \), and therefore by Definition 1, the tangent RD problem above must have a topological transition at some \( \beta^- < \beta < \beta^+ \).

Taking the limits \( \beta^- \to \beta^* \) from below and \( \beta^+ \to \beta^* \) from above, we say that the RD problem defined by \( \hat{X}^+ \) and \( d^* \) is the corresponding tangent RD problem to the original IB problem at \( \beta^* \). Consequently, if the IB problem has a topological transition at \( \beta^* \), then the corresponding tangent RD problem at \( \beta_c \) must also have a critical transition at \( \beta_c \) (see Figure 2C).

Finally, by its definition, the solution to the tangent RD problem at a given \( \beta \) already achieves the optimal decoder distribution at that point. However, since the IB algorithm has to iterate additionally over the decoder distributions in order to converge to the IB solution at \( \beta \), it is expected to perform at least the same number of iterations as the AB algorithm for the tangent RD solution there. Therefore, when \( \beta \) is close enough to some critical point of the IB, the IB algorithm experiences similar slowing down there, as shown in Figure 2B,E.

\textbf{VI. DISCUSSION}

The AB algorithm for rate-distortion problems is known to converge uniformly to the optimal RD function in times that are \( O(1/\varepsilon) \) [13]. Our results deal with the ratio between the number of iterations until \( \varepsilon \)-convergence and \( -\log \varepsilon \), and show that this ratio increases significantly near critical points. While the \( O(1/\varepsilon) \) bound is independent of \( \beta \), that is, the constant does not increase near critical points, it corresponds to a much slower convergence, for sufficiently small \( \varepsilon \). Moreover, our results address the convergence of the encoder \( \hat{p}(\hat{x}|x) \), not just the rate \( R(D) \). At the critical points there might be different competing optimal solutions at the same \( R(D) \).

For similar reasons, variational approximations to either RD or IB (e.g. [30]) may not suffer from CSD, as they can be too far from the optimal encoder or use different dynamics. It is not clear if other local converging algorithms, such as stochastic gradient decent, should also exhibit CSD near topological representation transitions, but we know that near the critical points the Hessian matrix of the Lagrangian at the optimum becomes singular. Thus any gradient based optimization is susceptible to CSD if it has components in the flat dimensions of the minima.

The implications of our results to local representation learning, when effective annealing is obtained through complexity regularization, are intriguing. In such cases, deep learning in particular, the critical points can determine the location of the final representations along the optimal RD or IB curves.
REFERENCES