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 $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$, 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 $\mathbb{E}[d(X, \hat{X})] \leq D$, [1], [2]. 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 $\operatorname{argmin}_{\hat{x}} \mathbb{E}[d(X, \hat{x})]$ is optimal, whereas for $R = H(X)$ the channel that maps any $x \in \mathcal{X}$ to $\operatorname{argmin}_{\hat{x}} d(x, \hat{x})$ 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 representors' 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]. Boukris [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,
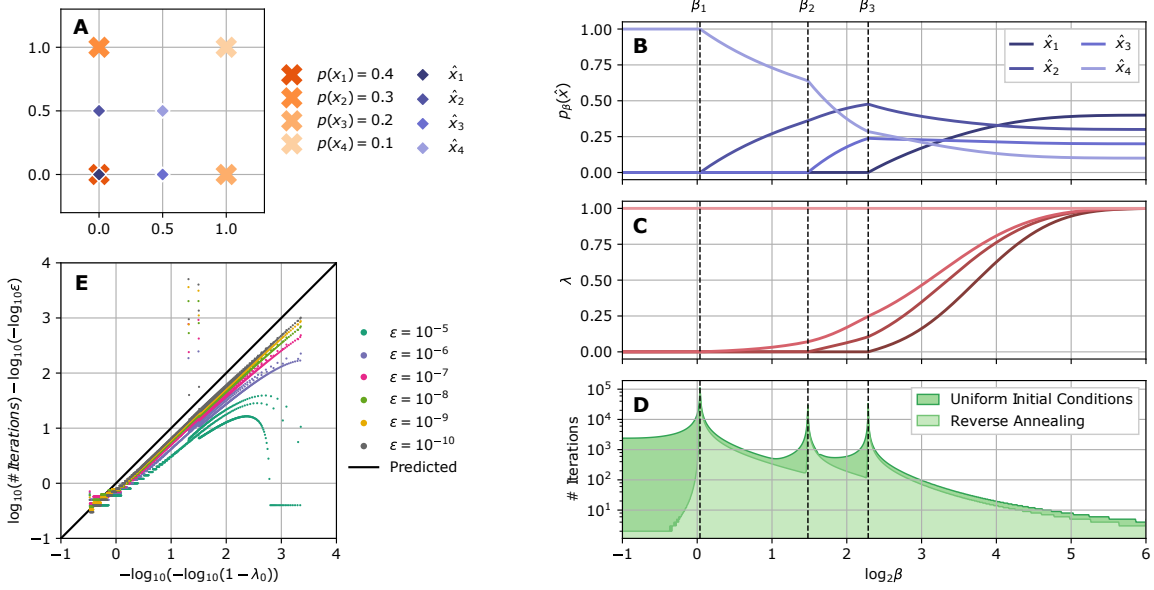

Fig. 1. **A.** A simple RD problem: the sets $\mathcal{X}, \hat{\mathcal{X}} \in \mathbb{R}^2$ shown in orange and purple, correspondingly; $p(x) = (0.4, 0.3, 0.2, 0.1)^\mathsf{T}$; the distortion function is defined as $d(x, \hat{x}) = \frac{1}{\mu} ||x - \hat{x}||_2^2$, where $\mu = \max_{x' \in \mathcal{X}, \ \hat{x}' \in \hat{\mathcal{X}}} ||x' - \hat{x}'||_2^2$ is a normalization factor. **B.** Values of $p_\beta(\hat{x})$, the solutions to the RD problem at different values of β: note the three phase transitions at β_1, β_2, β_3 (marked with dashed lines) – until β_1 only $p(\hat{x}_4) > 0$, then at β_1 the representor $\hat{x}_2$ starts to gain mass, and at β_2 also $\hat{x}_3$, finally $\hat{x}_1$ starts to gain mass at β_3. **C.** Eigenvalues of A at p_β: note that near each of the phase transitions an eigenvalue of A approaches 0. **D.** Number of iterations until convergence ($\varepsilon = 10^{-9}$, 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 λ_0, the smallest nonzero eigenvalue of A, computed using reverse annealing with various values of ε: as ε 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 β_1 (we do not fully understand this phenomenon, which is more prominent in the uniform initial conditions setting).

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 $\hat{X}$ 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 $\hat{X}$ 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 (i.e. $|\hat{\mathcal{X}}| = 1$) 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 $\hat{X}$.

III. CRITICAL POINTS IN RATE-DISTORTION THEORY

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

$$I(X; \hat{X}) + \beta \, \mathbb{E}[d(X, \hat{X})], \tag{1}$$

with respect to all test channels $p(\hat{x}|x)$. The Lagrange multiplier β, 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 β, denoted by $p_\beta(\hat{x}|x)$, that is, a minimizer of (1), must (self-consistently) satisfy both equations [1]–[3]:

$$p_\beta(\hat{x}|x) = \frac{p_\beta(\hat{x})e^{-\beta d(x, \hat{x})}}{Z(x, \beta)} \tag{2}$$

$$\text{and} \quad p_\beta(\hat{x}) = \sum_x p_\beta(\hat{x}|x)p(x) , \tag{3}$$

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

discussion, we assume throughout that $p_\beta(\hat{x})$ is unique for all values of β, and consider it as a vector p_β in the sequel.

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

$$|\operatorname{supp} p_{\beta^-}| \neq |\operatorname{supp} p_{\beta^+}| \tag{4}$$

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

We restrict our discussion to local stability analysis, and thus assume that $\beta \mapsto p_\beta$ 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 $F = 0$, where for all $\hat{x} \in \hat{\mathcal{X}}$

$$\left[F(p, \beta)\right]_{\hat{x}} := p(\hat{x}) - \sum_x p(x) \frac{p(\hat{x}) e^{-\beta d(x, \hat{x})}}{Z(x, \beta)} . \tag{5}$$

In what follows, we study the properties of F and its Jacobian $\nabla F = \partial[F(p, \beta)]_{\hat{x}}/\partial p(\hat{x}')$ around critical points, and provide a characterization of the phase transitions of RD problems in terms of ∇F. This characterization will play a pivotal role, as we show next that ∇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 β_c, such that $|\operatorname{supp} p_{\beta^-}| < |\operatorname{supp} p_{\beta^+}|$ 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 $\hat{\mathcal{X}}' = \hat{\mathcal{X}} \setminus \{\hat{x}_0\}$ and the distortion function d', which is the restriction of d to $\hat{\mathcal{X}}'$. Clearly, the solution q_β of the RD problem defined by d' is identical to p_β (restricted to $\hat{\mathcal{X}}'$) at a left neighborhood of β_c, but they must differ to the right. Nevertheless, the extension of q_β over the original problem satisfies $F(q_\beta, \beta) = 0$ also to the right of β_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 β, or that $F = 0$ has a *bifurcation* at β_c.

Although the extension of q_β 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_β instead. Hence, at β_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 ∇F must be singular at critical values of β. Otherwise, by the implicit function theorem, there must exist a unique solution to $F = 0$ as a function of β in the vicinity of β_c [26]. Unfortunately, ∇F is trivially singular when $p_\beta(\hat{x}) = 0$ for some representor $\hat{x}$, 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_β, then (see Appendix A)

$$A_{\hat{x}\hat{x}'} := \left[(\nabla F)^\mathsf{T}\right]_{\hat{x}\hat{x}'} = \sum_x p_\beta(\hat{x}'|x) p_\beta(x|\hat{x}) \tag{6}$$

$$= p_\beta(\hat{x}') \sum_x p(x) \frac{e^{-\beta\left(d(x,\hat{x}) + d(x,\hat{x}')\right)}}{Z(x, \beta)^2} , \tag{7}$$

and we have the following result:

Theorem 1: Let $m = |\hat{\mathcal{X}}|$ and $k = |\operatorname{supp} p_\beta|$, then at p_β $\dim \ker A = m - k$.

Corollary 1.1: Topological transitions of a RD problem occur exactly at values of β where the dimension of $\ker A$, at the solutions p_β, changes.

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

Lemma 2: Let $\mathbf{e}_{\hat{x}} \in \mathbb{R}^m$ denote the standard basis vector with 1 at the $\hat{x}$ coordinate and 0 elsewhere, then $p_\beta(\hat{x}) = 0 \iff \mathbf{e}_{\hat{x}} \in \ker A$.

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 $\mathbf{e}_{\hat{x}_i}$, 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 (see 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 $m = |\hat{\mathcal{X}}|$ and $k = |\operatorname{supp} p_\beta|$, then at p_β 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 β 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 β_1, β_2 and β_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 β another eigenvalue of A reaches 0.

Finally, as the matrix A is row-stochastic, all its eigenvalues are inside the unit circle [27], 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 $\Delta\hat{\mathcal{X}}$. The k-th iteration p_k of the AB algorithm is said to have ε-*converged* to a RD solution p_β if [1]

$$\|p_k - p_\beta\| < \varepsilon . \tag{8}$$

Given a value of β, define the operator $AB : \Delta\hat{\mathcal{X}} \to \Delta\hat{\mathcal{X}}$ 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 $|\hat{\mathcal{X}}|$ is finite and we are typically interested in small values of ε. For convenience, we have chosen L^1 in Theorem 5 below, and L^∞ 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 $\nabla AB|_{p_\beta} = I - A^\mathsf{T}$, where $\nabla AB|_{p_\beta}$ is the Jacobian matrix of AB at p_β.

Let $\delta p_k = p_k - p_\beta$ be the deviation from p_β, then to first order in δp_k,

$$AB\, p_k \approx p_\beta + \nabla AB|_{p_\beta} \delta p_k \qquad (9)$$
$$\Rightarrow\ \delta p_{k+1} = AB\, p_k - p_\beta \approx (I - A^\mathsf{T})\delta p_k\ .$$

Hence,

$$\delta p_k \approx (I - A^\mathsf{T})^k \delta p_0\ . \qquad (10)$$

As a result, the convergence rate of the algorithm is governed by the largest eigenvalue of $I - A^\mathsf{T}$ inside the unit circle, denoted λ_{max}. If the initial deviation δp_0 has a nonzero component in the eigenspace of λ_{max}, then

$$\|\delta p_k\| < \varepsilon \quad \overset{(10)}{\Longrightarrow} \quad k \approx \frac{-\log \varepsilon + \text{const}}{-\log |\lambda_{max}|}\ . \qquad (11)$$

For the asymptotic convergence rate we consider $\lim_{\varepsilon \to 0} \frac{k}{-\log \varepsilon}$, 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 Appendix D. Recall that A is diagonalizable (Theorem 4) and its eigenvalues are in $[0,1]$. Therefore, $\nabla AB = I - A^\mathsf{T}$ is also diagonalizable and its eigenvalues are in $[0,1]$. When p_β has a full support, the eigenvalues of ∇AB are in $[0,1)$ (Theorem 1). Consequently, we have $\lambda_{max} = 1 - \lambda_0 < 1$, where $\lambda_0 > 0$ is the smallest nonzero eigenvalue of A.

Theorem 5: Let p_β be a RD solution with $p_\beta(\hat{x}) > 0$ for all $\hat{x}$, and $\lambda_{max} = 1 - \lambda_0 < 1$ the largest eigenvalue smaller than 1 of ∇AB at p_β. Denote by $k(p_0, \varepsilon)$ the number of iterations required for an initial distribution p_0 to ε-converge to p_β, and define $B(\delta) := \{p \in \Delta\hat{\mathcal{X}} : \|p - p_\beta\|_1 \le \delta\}$. Then, for any $a > 0$,

$$\Pr_{p_0 \sim \mathcal{U}(B(\delta))} \left(\left| \lim_{\varepsilon \to 0^+} \frac{k(p_0, \varepsilon)}{-\log \varepsilon} - \frac{1}{-\log \lambda_{max}} \right| < a \right) \xrightarrow[\delta \to 0]{} 1\ , \qquad (12)$$

where $\mathcal{U}(S)$ denotes the uniform distribution on S.

A lower bound cannot be expected to hold for *every* initial condition in the vicinity of p_β. 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 λ_{max} eigenspace. In the general case, where the dynamics are nonlinear, the fraction of applicable initial conditions increases in the vicinity of p_β 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 β_c, such that $|\operatorname{supp} p_{\beta-}| < |\operatorname{supp} p_{\beta+}|$. According to Corollary 4.1, the algebraic multiplicity of the eigenvalue 0 of A is greater at β^- than at β^+. Since A is continuous in p_β and p_β is assumed continuous in β, there exists an eigenvalue $\lambda_0 > 0$ of A such that $\lambda_0 \to 0$ as β approaches β_c from above. When coordinates $\hat{x}$ outside $\operatorname{supp} p_{\beta+}$ are initialized at 0, as in reverse annealing (see below), AB effectively coincides with its restriction to $\operatorname{supp} p_{\beta+}$. 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 β, starting from the converged solution at the previous β. 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 β decreases, since more eigenvalues reach 0.

Finally, Figure 1E shows that the number of iterations required to converge to a solution within accuracy ε gets closer to the bound in Theorem 5 as ε decreases. Smaller ε 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 $I(X; \hat{X})$, 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 $\Delta\hat{\mathcal{X}}$, the AB algorithm does not directly apply, since $\hat{\mathcal{X}}$ is continuous. Nevertheless, it is known [28] that for each β, taking at most $|\mathcal{X}|$ points of the simplex $\Delta\hat{\mathcal{X}}$ 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 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,

$$I(X; \hat{X}) - \beta\, I(\hat{X}; Y) \qquad (13)$$

over all channels from $\mathcal{X}$ to $\hat{\mathcal{X}} = \{1, \ldots, |\mathcal{X}|\}$. 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

$$p_\beta(y|\hat{x}) = \sum_x p(y|x) \frac{p_\beta(\hat{x}|x)p(x)}{p_\beta(\hat{x})}\ , \qquad (14)$$

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 β. 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 $\Delta\hat{\mathcal{X}}$, e.g. sampling according to the symmetric Dirichlet distribution $p_0 \sim Dir(\mathbf{1})$, 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 $p_\beta(y|\hat{x})$, 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, $|\{p_\beta(y|\hat{x}) : p_\beta(\hat{x}) > 0\}|$.
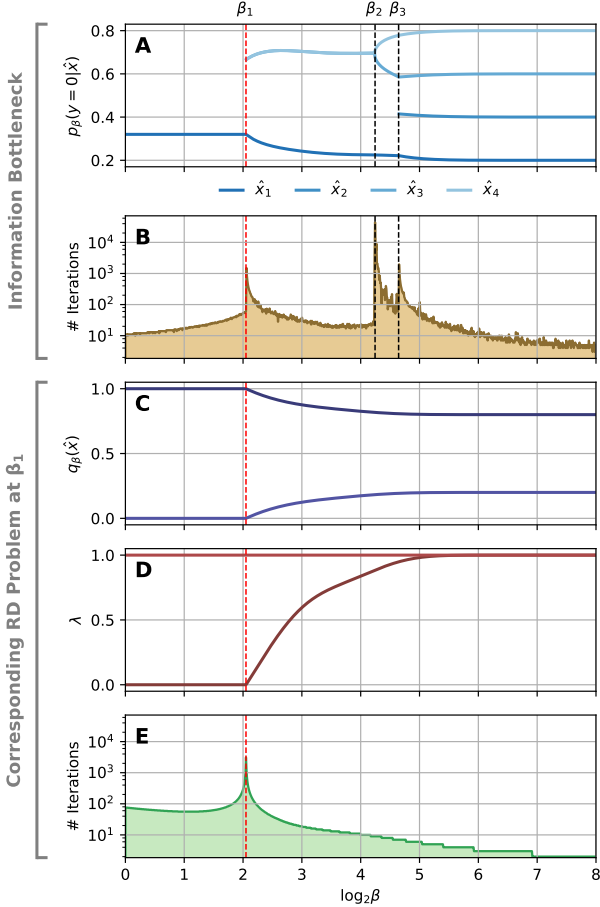

Fig. 2. A simple IB problem with a binary Y, defined by $p(x) = (0.7, 0.1, 0.1, 0.1)$ and $p(y = 0 \mid x) = (0.2, 0.4, 0.6, 0.8)$, as well as its corresponding RD problem. **A.** The decoder distribution, $p_\beta(y = 0 \mid \hat{x})$: notice the three phase transitions of the IB problem at β_1, β_2, β_3 (marked with dashed lines). **B.** Number of iterations until convergence of the IB solution ($\varepsilon = 10^{-7}$, logarithmic scale): a slowing down of at least one order of magnitude is clearly observed at each critical point. **C.** Values of $q_\beta(\hat{x})$ of the corresponding (tangent) RD problem at β_1 (red dashed line): its topological transition is at β_1, as in the IB. **D.** Eigenvalues of A at q_β. **E.** Number of iterations until convergence of the RD solution ($\varepsilon = 10^{-7}$, logarithmic scale): notice the single slowing down at β_1.

IB problems can be solved numerically using a modified version of the AB algorithm [7], which iterates also the decoder equation (14). Based on our RD analysis, we show that the IB alternating projection algorithm also exhibits critical slowing down near topological transitions (see Figure 2A,B).

While it can be analyzed directly, we can rely on the fact that the IB curve is an envelope of tangent RD curves [10]. Consider an IB solution $p_{\beta^*}(\hat{x}|x)$ at some value β^* and let $\beta^- < \beta^* < \beta^+$ in some small neighborhood. Let $\hat{\mathcal{X}}^-$ be a set of representors of the effective cardinality of p_{β^-}, such

that all their decoder distributions $p_{\beta^-}(y|\hat{x})$ are different, and let $\hat{\mathcal{X}}^+$ be defined accordingly for p_{β^+}. Define $\hat{\mathcal{X}}^*$ to be the (formal) disjoint union $\hat{\mathcal{X}}^- \sqcup \hat{\mathcal{X}}^+$, and the (fixed) distortion function $d^* : \mathcal{X} \times \hat{\mathcal{X}}^* \to \mathbb{R}^+$ as

$$d^*(x, \hat{x}) := \begin{cases} D_{KL}[p(y|x)||p_{\beta^-}(y|\hat{x})] & \hat{x} \in \hat{\mathcal{X}}^- \\ D_{KL}[p(y|x)||p_{\beta^+}(y|\hat{x})] & \hat{x} \in \hat{\mathcal{X}}^+ \end{cases} \quad (15)$$

Let q_β be the solution to the RD problem defined by $\hat{\mathcal{X}}^*$ and d^*. Since p_{β^-} is optimal at β^-, being the IB solution there, we know that $\text{supp}\, q_{\beta^-} = \hat{\mathcal{X}}^-$; similarly $\text{supp}\, q_{\beta^+} = \hat{\mathcal{X}}^+$. Notice that if the IB problem undergoes a transition at β^*, then the effective cardinality of p_{β^-} differs from that of p_{β^+}, 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 $\hat{\mathcal{X}}^*$ and d^* is the *corresponding tangent RD problem to the original IB problem at β^**. Consequently, if the IB problem has a topological transition at β_c then the corresponding tangent RD problem at β_c must also have a critical transition at β_c (see Figure 2C).

Finally, by its definition, the solution to the tangent RD problem at a given β 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 β, it is expected to perform at least the same number of iterations as the AB algorithm for the tangent RD solution there. Therefore, when β is close enough to some critical point of the IB, the IB algorithm experiences similar slowing down there, as shown in Figure 2B,E.

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 ε-convergence and $-\log \varepsilon$, and show that this ratio increases significantly near critical points. While the $O(1/\varepsilon)$ bound is independent of β, that is, the constant does not increase near critical points, it corresponds to a much slower convergence, for sufficiently small ε. Moreover, our results address the convergence of the encoder $p(\hat{x}|x)$, 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. [29]) 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

[1] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.

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

[3] 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.

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

[5] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[6] A. Gordon, A. Banerjee, M. Koch-Janusz, and Z. Ringel, "Relevance in the renormalization group and in information theory," *arXiv preprint arXiv:2012.01447*, 2020.

[7] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *The 37th annual Allerton Conference on Communication, Control, and Computing*, 1999, pp. 368–377.

[8] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *IEEE Information Theory Workshop*, 2015, pp. 1–5.

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

[10] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 595–609.

[11] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 19–38, 2020.

[12] I. Csiszar, "On the computation of rate-distortion functions (Corresp.)," *IEEE Transactions on Information Theory*, vol. 20, no. 1, pp. 122–124, 1974.

[13] P. Boukris, "An upper bound on the speed of convergence of the Blahut algorithm for computing rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 19, no. 5, pp. 708–709, 1973.

[14] Y. Yu, "Squeezing the Arimoto–Blahut algorithm for faster convergence," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3149–3157, 2010.

[15] K. Nakagawa, Y. Takei, S. Hara, and K. Watabe, "Analysis of the convergence speed of the Arimoto-Blahut algorithm by the second order recurrence formula," *arXiv preprint arXiv:2009.08780*, 2020.

[16] G. Matz and P. Duhamel, "Information geometric formulation and interpretation of accelerated Blahut-Arimoto-type algorithms," in *IEEE Information Theory Workshop*, 2004, pp. 66–70.

[17] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1939–1952, 1994.

[18] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds., vol. 12. MIT Press, 2000, pp. 617–623.

[19] K. Rose, E. Gurewitz, and G. Fox, "A deterministic annealing approach to clustering," *Pattern Recognition Letters*, vol. 11, pp. 589–594, 1990.

[20] A. E. Parker, T. Gedeon, and A. G. Dimitrov, "Annealing and the rate distortion problem," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15. MIT Press, 2003, pp. 993–976.

[21] 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.

[22] T. Wu, I. Fischer, I. L. Chuang, and M. Tegmark, "Learnability for the information bottleneck," in *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, ser. Proceedings of Machine Learning Research, R. P. Adams and V. Gogate, Eds., vol. 115. PMLR, 2020, pp. 1050–1060.

[23] T. Wu and I. Fischer, "Phase transitions for the information bottleneck in representation learning," in *International Conference on Learning Representations*, 2020.

[24] N. Zaslavsky, "Information-theoretic principles in the evolution of semantic systems," Ph.D. dissertation, The Hebrew University of Jerusalem, 2019. [Online]. Available: https://www.nogsky.com/publication/phd-thesis/ZaslavskyPhDthesis.pdf

[25] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," *Journal of Machine Learning Research*, vol. 6, no. 6, p. 165–188, 2005.

[26] H. Kielhöfer, *Bifurcation Theory An Introduction with Applications to Partial Differential Equations*, 2nd ed. Springer, 2012.

[27] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge University Press, 2012.

[28] P. Harremoës and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *IEEE International Symposium on Information Theory*, 2007, pp. 566–570.

[29] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *International Conference on Learning Representations*, 2017.

[30] G. B. Folland, "Remainder estimates in Taylor's theorem," *The American Mathematical Monthly*, vol. 97, no. 3, pp. 233–235, 1990.

APPENDIX A
DERIVATION OF A

Recall that $Z(x, \beta) := \sum_{\hat{x}} p_\beta(\hat{x}) e^{-\beta d(x, \hat{x})}$, therefore

$$\frac{\partial Z(x, \beta)}{\partial p_\beta(\hat{x})} = e^{-\beta d(x, \hat{x})} . \tag{16}$$

For $\hat{x}' \neq \hat{x}$ we have

$$\frac{\partial}{\partial p_\beta(\hat{x}')} \left(\frac{p_\beta(\hat{x})}{Z(x, \beta)} \right) = -\frac{p_\beta(\hat{x}) e^{-\beta d(x, \hat{x}')}}{Z(x, \beta)^2} , \tag{17}$$

and thus

$$\frac{\partial [F(p_\beta, \beta)]_{\hat{x}}}{\partial p_\beta(\hat{x}')} = p_\beta(\hat{x}) \sum_x p(x) \frac{e^{-\beta \left(d(x, \hat{x}) + d(x, \hat{x}') \right)}}{Z(x, \beta)^2} . \tag{18}$$

In contrast,

$$\frac{\partial}{\partial p_\beta(\hat{x})} \left(\frac{p_\beta(\hat{x})}{Z(x, \beta)} \right) = \frac{Z(x, \beta) - p_\beta(\hat{x}) e^{-\beta d(x, \hat{x})}}{Z(x, \beta)^2}$$
$$= \frac{1}{Z(x, \beta)} - \frac{p_\beta(\hat{x}) e^{-\beta d(x, \hat{x})}}{Z(x, \beta)^2} , \tag{19}$$

and then

$$\frac{\partial [F(p_\beta, \beta)]_{\hat{x}}}{\partial p_\beta(\hat{x})} = 1 - \sum_x p(x) \left(\frac{e^{-\beta d(x, \hat{x})}}{Z(x, \beta)} - \frac{p_\beta(\hat{x}) e^{-2\beta d(x, \hat{x})}}{Z(x, \beta)^2} \right)$$
$$= 1 - \sum_x \frac{p(x) e^{-\beta d(x, \hat{x})}}{Z(x, \beta)} + p_\beta(\hat{x}) \sum_x p(x) \frac{e^{-2\beta d(x, \hat{x})}}{Z(x, \beta)^2} . \tag{20}$$

When $p_\beta(\hat{x}) > 0$ we have by (2), applying Bayes' law,

$$p_\beta(x|\hat{x}) = \frac{p(x)}{p_\beta(\hat{x})} p(\hat{x}|x) = \frac{p(x) e^{-\beta d(x, \hat{x})}}{Z(x, \beta)} , \tag{21}$$

and therefore $\sum_x \frac{p(x) e^{-\beta d(x, \hat{x})}}{Z(x, \beta)} = 1$. Note, however, that the right hand side of (21) is well defined even when $p_\beta(\hat{x}) = 0$. Since we are interested only in the right derivative when $p_\beta(\hat{x}) = 0$ (as it is in the boundary of $\Delta \hat{\mathcal{X}}$), we can refer in that case to the limit, which also satisfies $\sum_x \frac{p(x) e^{-\beta d(x, \hat{x})}}{Z(x, \beta)} = 1$. Consequently, we have from (20)

$$\frac{\partial [F(p_\beta, \beta)]_{\hat{x}}}{\partial p_\beta(\hat{x})} = p_\beta(\hat{x}) \sum_x p(x) \frac{e^{-2\beta d(x, \hat{x})}}{Z(x, \beta)^2} , \tag{22}$$

which together with (18) gives the result in (7).

The formula in (6) is a straightforward result of (7) and (21). Although simpler, it is not defined when $p_\beta(\hat{x}) = 0$.

Proof of Lemma 2: If $p_\beta(\hat{x}) = 0$ then by (7) the corresponding column of A is zero and thus $A\mathbf{e}_{\hat{x}} = 0$. Conversely, if $A\mathbf{e}_{\hat{x}} = 0$, then by (7) we have particularly

$$p_\beta(\hat{x}) \sum_x p(x) \frac{e^{-2\beta d(x,\hat{x})}}{Z(x,\beta)^2} = 0 \ . \tag{23}$$

Therefore, either $p_\beta(\hat{x}) = 0$ or $d(x,\hat{x}) = \infty$ for all x. But the latter implies by (2) that $p_\beta(\hat{x}|x) = 0$ for all x, and so anyway $p_\beta(\hat{x}) = 0$. ∎

Proof of Proposition 3: We prove the proposition by induction on r.

The case $r = 1$ is trivial. Let $r = 2$ and assume by contradiction that $\mathbf{e}_{\hat{x}_1} \notin \ker A$. This means that also $\mathbf{e}_{\hat{x}_2} \notin \ker A$, otherwise $v - v_{\hat{x}_2}\mathbf{e}_{\hat{x}_2} \in \ker A$, which would imply that $\mathbf{e}_{\hat{x}_1} \in \ker A$. Therefore, according to Lemma 2, both $p_\beta(\hat{x}_1), p_\beta(\hat{x}_2) > 0$.

Now, since $v \in \ker A$, we have by (7) for all $\hat{x}$

$$0 = \sum_{\hat{x}'} p_\beta(\hat{x}') \sum_x p(x) \frac{e^{-\beta\left(d(x,\hat{x})+d(x,\hat{x}')\right)}}{Z(x,\beta)^2} v_{\hat{x}'} \tag{24}$$

$$= \sum_x p(x) \frac{e^{-\beta d(x,\hat{x})}}{Z(x,\beta)} \left(p_\beta(\hat{x}_1|x)v_{\hat{x}_1} + p_\beta(\hat{x}_2|x)v_{\hat{x}_2}\right) , \tag{25}$$

where the second equality follows from (2). Averaging over $p_\beta(\hat{x})$ gives $p_\beta(\hat{x}_1)v_{\hat{x}_1} + p_\beta(\hat{x}_2)v_{\hat{x}_2} = 0$, and thus $v_{\hat{x}_1} = -\frac{p_\beta(\hat{x}_2)}{p_\beta(\hat{x}_1)}v_{\hat{x}_2}$. Plugging this result back in (25) and dividing by $p_\beta(\hat{x}_2)v_{\hat{x}_2} \neq 0$ we get for all $\hat{x}$

$$\sum_x p(x) \frac{e^{-\beta d(x,\hat{x})}}{Z(x,\beta)^2} \left(e^{-\beta d(x,\hat{x}_2)} - e^{-\beta d(x,\hat{x}_1)}\right) = 0 , \tag{26}$$

where we used again (2). Substituting $\hat{x} = \hat{x}_1, \hat{x}_2$ in the last equation we have

$$\sum_x \frac{p(x)}{Z(x,\beta)^2} \left(e^{-\beta d(x,\hat{x}_1)}e^{-\beta d(x,\hat{x}_2)} - e^{-2\beta d(x,\hat{x}_1)}\right) = 0 \tag{27}$$

$$\sum_x \frac{p(x)}{Z(x,\beta)^2} \left(e^{-2\beta d(x,\hat{x}_2)} - e^{-\beta d(x,\hat{x}_1)}e^{-\beta d(x,\hat{x}_2)}\right) = 0 \tag{28}$$

and subtracting (27) from (28) gives

$$\sum_x \frac{p(x)}{Z(x,\beta)^2} \left(e^{-\beta d(x,\hat{x}_1)} - e^{-\beta d(x,\hat{x}_2)}\right)^2 = 0 \ . \tag{29}$$

Therefore, for all x we must have $d(x,\hat{x}_1) = d(x,\hat{x}_2)$, contradicting our assumption on the non-degeneracy of d. Consequently, $\mathbf{e}_{\hat{x}_1} \in \ker A$, and thus $v - v_{\hat{x}_1}\mathbf{e}_{\hat{x}_1} \in \ker A$, implying that also $\mathbf{e}_{\hat{x}_2} \in \ker A$.

Finally, let $r \geq 3$ and assume the proposition holds for all $1 \leq r' < r$. If there exists $1 \leq i \leq r$ such that $\mathbf{e}_{\hat{x}_i} \in \ker A$, then $u = v - v_{\hat{x}_i}\mathbf{e}_{\hat{x}_i} \in \ker A$. However, u has exactly $r-1 < r$ nonzero coordinates, namely $\hat{x}_j$ for $1 \leq j \leq r$, $j \neq i$, and therefore by the induction hypothesis all the corresponding $\mathbf{e}_{\hat{x}_j}$ also belong to $\ker A$. Together with $\mathbf{e}_{\hat{x}_i}$ this completes the induction step.

Conversely, assume by contradiction that $\mathbf{e}_{\hat{x}_i} \notin \ker A$ for all $1 \leq i \leq r$, then by Lemma 2 we have $p_\beta(\hat{x}_i) > 0$ for all $\hat{x}_i$.

This implies that $A_{\hat{x}\hat{x}_i} > 0$ for all $\hat{x}$ and $\hat{x}_i$, otherwise by (7) we would get for some $\hat{x}$ and $\hat{x}_i$ that $d(x,\hat{x}) + d(x,\hat{x}_i) = \infty$ for all x, contradicting our assumption on the finiteness of d. In particular, this means that $\sum_{i=2}^r A_{\hat{x}\hat{x}_i} > 0$.

Now, since $v \in \ker A$ we have $\sum_{i=1}^r A_{\hat{x}\hat{x}_i}v_{\hat{x}_i} = 0$ for all $\hat{x}$, and thus

$$0 = A_{\hat{x}\hat{x}_1}v_{\hat{x}_1} + \sum_{i=2}^r A_{\hat{x}\hat{x}_i}v_{\hat{x}_i} \tag{30}$$

$$= \sum_{i=2}^r \frac{A_{\hat{x}\hat{x}_i}A_{\hat{x}\hat{x}_1}v_{\hat{x}_1}}{\sum_{j=2}^r A_{\hat{x}\hat{x}_j}} + \sum_{i=2}^r A_{\hat{x},\hat{x}_i}v_{\hat{x}_i} \tag{31}$$

$$= \sum_{i=2}^r A_{\hat{x}\hat{x}_i} \left(\frac{A_{\hat{x}\hat{x}_1}v_{\hat{x}_1}}{\sum_{j=2}^r A_{\hat{x}\hat{x}_j}} + v_{\hat{x}_i}\right) \ . \tag{32}$$

Define the vector $u \in \mathbb{R}^m$ such that $u_{\hat{x}_i} = \frac{A_{\hat{x}\hat{x}_1}v_{\hat{x}_1}}{\sum_{j=2}^r A_{\hat{x}\hat{x}_j}} + v_{\hat{x}_i}$ for all $2 \leq i \leq r$, and all its other coordinates are 0. By (32) $u \in \ker A$ and it has at most $r-1 < r$ nonzero coordinates. If there exists $2 \leq i \leq r$ such that $u_{\hat{x}_i} \neq 0$ then by the induction hypothesis we would have $\mathbf{e}_{\hat{x}_i} \in \ker A$, contradicting our assumption. Therefore, for all $2 \leq i \leq r$ we must have

$$v_{\hat{x}_i} = -\frac{A_{\hat{x}\hat{x}_1}v_{\hat{x}_1}}{\sum_{j=2}^r A_{\hat{x}\hat{x}_j}} \ . \tag{33}$$

In particular this implies that $\operatorname{sgn} v_{\hat{x}_2} = \operatorname{sgn} v_{\hat{x}_3} = -\operatorname{sgn} v_{\hat{x}_1}$. Finally, we can perform the same analysis starting at (30) by setting aside $v_{\hat{x}_2}$ instead of $v_{\hat{x}_1}$, concluding with $\operatorname{sgn} v_{\hat{x}_1} = \operatorname{sgn} v_{\hat{x}_3} = -\operatorname{sgn} v_{\hat{x}_2}$. Together with the previous result, this means that $\operatorname{sgn} v_{\hat{x}_i} = 0$ for $i = 1, 2, 3$, or equivalently that $v_{\hat{x}_i} = 0$, contradicting our initial assumption and completing the induction step. ∎

Proof: First, we deal with the case in which all $p_\beta(\hat{x}) > 0$. Note from (7) that A can be written as the product of three matrices,

$$A = BB^\mathsf{T}C , \tag{34}$$

where $B_{\hat{x}x} = \frac{p(x)^{1/2}}{Z(x,\beta)}e^{-\beta d(x,\hat{x})}$ and C is a diagonal matrix with $p_\beta(\hat{x})$ in its diagonal. Therefore, we have

$$C^{1/2}AC^{-1/2} = C^{1/2}BB^\mathsf{T}C^{1/2} = (C^{1/2}B)(C^{1/2}B)^\mathsf{T}, \tag{35}$$

meaning that A is similar to a real Gram matrix, and thus diagonalizable with non-negative eigenvalues [27] .

Second, assume that $p_\beta(\hat{x}_i) = 0$ for $i = 1, \ldots, r$ – the first $r \geq 1$ coordinates – and $p_\beta(\hat{x}) > 0$ elsewhere. Let $\hat{X}' = \operatorname{supp} p_\beta$ and denote by p'_β and A' the solutions and matrix corresponding to the RD problem restricted to $\hat{X}'$. Note that $Z(x,\beta)$ depends only on the support, and thus

$$A = \begin{pmatrix} 0 & \cdots \\ 0 & A' \end{pmatrix} \ . \tag{36}$$

Since $p'_\beta(\hat{x}) > 0$ for all $\hat{x} \in \hat{X}'$, there exist, by the first part of the proof, an invertible matrix P and a non-negative diagonal matrix Λ such that $P^{-1}A'P = \Lambda$. Moreover, from

Theorem 1 we have $\dim \ker A' = 0$, hence none of the values in the diagonal of Λ (that is, the eigenvalues of A') is 0.

Now, we have

$$\begin{pmatrix} I_r & 0 \\ 0 & P^{-1} \end{pmatrix} A \begin{pmatrix} I_r & 0 \\ 0 & P \end{pmatrix}$$
$$= \begin{pmatrix} I_r & 0 \\ 0 & P^{-1} \end{pmatrix} \begin{pmatrix} 0 & \cdots \\ 0 & A' \end{pmatrix} \begin{pmatrix} I_r & 0 \\ 0 & P \end{pmatrix}$$
$$= \begin{pmatrix} 0 & \cdots \\ 0 & P^{-1}A'P \end{pmatrix} = \begin{pmatrix} 0 & \cdots \\ 0 & \Lambda \end{pmatrix} , \quad (37)$$

where I_r is the $r \times r$ identity matrix. This means that A is similar to an upper-triangular matrix, and thus all its eigenvalues appear on the diagonal of that matrix, repeated according to their respective algebraic multiplicities. Since Λ has no zeroes in its diagonal, we conclude that the algebraic multiplicity of the eigenvalue 0 of A is exactly r. However, according to Theorem 1 we have $\dim \ker A = r$, or equivalently, that the geometric multiplicity of the eigenvalue 0 of A is also r.

Finally, note that the standard basis row vector $e_{\hat{x}_{r+i}}^{\mathsf{T}}$ for $i \geq 1$ is a left eigenvector of the matrix in (37), associated with the eigenvalue Λ_{ii}. Therefore, also for all nonzero eigenvalues of A, the algebraic multiplicity must equal the geometric multiplicity. Consequently the matrix A is diagonalizable with real non-negative eigenvalues. ∎

APPENDIX D
PROOF OF THEOREM 5

Proof: Denote by δp_k the deviation vector $p_k - p_\beta$ of the k-th iterate from the fixed point p_β, $\delta p_k(\hat{x}) = p_k(\hat{x}) - p_\beta(\hat{x})$ for its $\hat{x}$-indexed entry. For convenience, we use the L^1 norm in the sequel, denoted $\|\cdot\|$. The expansion of AB around p_β is [30]

$$AB(p_\beta + \delta p_k) - p_\beta = \nabla AB\big|_{p_\beta} \delta p_k + O(\|\delta p_k\|^2). \quad (38)$$

That is, to first order, a single AB iteration amounts to an application of the linear operator $\nabla AB\big|_{p_\beta(\hat{x})} = I - A^{\mathsf{T}}$ to the deviation. Write $B(0, r)$ for the ball of radius r around the origin with respect to L^1. Then,

$$\delta p_{k+1} \in (I - A^{\mathsf{T}})\delta p_k + B(0, \tilde{c}\|\delta p_k\|^2), \quad (39)$$

where $\tilde{c} > 0$ is a constant bounding the expansion's remainder. By Theorem 4, A is diagonalizable, and so $I - A^{\mathsf{T}} = P\Lambda P^{-1}$ with Λ diagonal. Multiplying (39) by P^{-1},

$$P^{-1}\delta p_{k+1} \in P^{-1}\left(P\Lambda P^{-1}\right)\delta p_k + P^{-1}B(0, \tilde{c}\|\delta p_k\|^2). \quad (40)$$

Denote $\|\cdot\|_{op}$ for the operator norm with respect to L^1. By its definition, $P^{-1}B(0, r) \subset B(0, \|P^{-1}\|_{op}\, r)$. Thus, exchanging coordinates $\tilde{\delta p}_k := P^{-1}\delta p_k$ to a basis of eigenvectors,

$$\tilde{\delta p}_{k+1} \in \Lambda \tilde{\delta p}_k + B(0, c\|\tilde{\delta p}_k\|^2), \quad (41)$$

for $c := \tilde{c} \cdot \|P\|_{op}^2 \cdot \|P^{-1}\|_{op}$.

Denote by $\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ the eigenvalues of $I - A^{\mathsf{T}}$. As noted in Section IV, they are contained in $[0, 1)$ by our assumptions. Denote the i-th coordinate of $\tilde{\delta p}_k$ with respect to this basis by $\tilde{\delta p}_k^{(i)}$, $i = 1, \ldots, |\hat{\mathcal{X}}|$. For exposition's

simplicity, suppose that λ_1 is a simple eigenvalue, $\lambda_1 > \lambda_2$; the proof is similar otherwise[3].

Let $0 < a < \frac{1}{-\log \lambda_1}$. An upper bound for convergence is immediate, when $\lambda_1 < 1$. Choose $\mu := \exp\left(\left(\frac{1}{\log \lambda_1} - a\right)^{-1}\right)$. It satisfies $\frac{1}{-\log \mu} = \frac{1}{-\log \lambda_1} + a$, and $\lambda_1 < \mu < 1$. Then whenever $\|\tilde{\delta p}_k\| \leq \frac{1}{c}(\mu - \lambda_1)$ we have

$$\|\tilde{\delta p}_{k+1}\| \overset{(41)}{\leq} \lambda_1 \|\tilde{\delta p}_k\| + c\|\tilde{\delta p}_k\|^2 \leq \mu\|\tilde{\delta p}_k\|. \quad (42)$$

Since $\|\tilde{\delta p}_k\| \leq \|P\|_{op} \cdot \|\tilde{\delta p}_k\|$, this holds whenever

$$\|\delta p_k\| \leq \delta_1 := \frac{\|P\|_{op}}{c}(\mu - \lambda_1). \quad (43)$$

Therefore, at most

$$k \leq \frac{-\log \varepsilon + \log(\|P^{-1}\delta p_0\| \cdot \|P\|_{op})}{-\log \mu} \quad (44)$$

iterations are then required for ε-convergence of p_k. To capture the asymptotic convergence rate we divide by $-\log \varepsilon$ and take the limit to obtain

$$\lim_{\varepsilon \to 0^+} \frac{k}{-\log \varepsilon} \leq \lim_{\varepsilon \to 0^+} \frac{1 + \frac{\log(\|P^{-1}\delta p_0\| \cdot \|P\|_{op})}{-\log \varepsilon}}{-\log \mu}$$
$$= \frac{1}{-\log \mu} = \frac{1}{-\log \lambda_1} + a. \quad (45)$$

For a lower bound, choose $\eta := \exp\left(\left(\frac{1}{\log \lambda_1} + a\right)^{-1}\right)$. It satisfies $\frac{1}{-\log \eta} = \frac{1}{-\log \lambda_1} - a > 0$, and thus $0 < \eta < \lambda_1$. Define,

$$\rho(\tilde{\delta p}) := \frac{|\tilde{\delta p}^{(1)}|}{\|\tilde{\delta p}\|} \quad (46)$$

when $\tilde{\delta p}^{(1)} \neq 0$, $\rho_k := \rho(\tilde{\delta p}_k)$. We proceed by assuming

$$|\tilde{\delta p}_k^{(1)}| \geq \rho_0 \cdot \|\tilde{\delta p}_k\| > 0 \quad (47)$$

for all $k \geq 0$. That is, the relative weight of the first components cannot decrease beyond its initial value at $k = 0$. This shall be justified in the sequel. From (41),

$$|\tilde{\delta p}_{k+1}^{(1)}| \geq \lambda_1 |\tilde{\delta p}_k^{(1)}| - c\|\tilde{\delta p}_k\|^2 \overset{(47)}{\geq} \lambda_1 |\tilde{\delta p}_k^{(1)}| - c\frac{1}{\rho_0^2}|\tilde{\delta p}_k^{(1)}|^2$$
$$= |\tilde{\delta p}_k^{(1)}| \left[\lambda_1 - \frac{c}{\rho_0^2}|\tilde{\delta p}_k^{(1)}|\right]. \quad (48)$$

Thus, if $|\tilde{\delta p}_k^{(1)}| \leq \frac{\rho_0^2}{c}(\lambda_1 - \eta)$ then $|\tilde{\delta p}_{k+1}^{(1)}| \geq \eta|\tilde{\delta p}_k^{(1)}|$. If the above were to hold for all $k \geq 0$, then we obtain a lower bound

$$|\tilde{\delta p}_k^{(1)}| \geq \eta^k |\tilde{\delta p}_0^{(1)}|. \quad (49)$$

Since $|\tilde{\delta p}_k^{(1)}| \leq \|\tilde{\delta p}_k\| \leq \|P^{-1}\|_{op} \cdot \|\delta p_k\|$, the condition $|\tilde{\delta p}_k^{(1)}| \leq \frac{\rho_0^2}{c}(\lambda_1 - \eta)$ can be replaced by the stricter

$$\|\delta p_k\| \leq \delta_2 := \frac{\rho_0^2}{c\|P^{-1}\|_{op}}(\lambda_1 - \eta). \quad (50)$$

---

[3]If λ_{max} is of multiplicity > 1, then take $\tilde{\delta p}_k^{(1)}$ to be a non-zero component along some normalized λ_{max}-eigenvector, and discard the other coordinates in the λ_{max}-eigenspace. The proof follows with minor modifications.

Since $\|\tilde{\delta p}_k\| \geq |\tilde{\delta p}_k^{(1)}|$, and $|\tilde{\delta p}_0^{(1)}| = \rho_0 \|\tilde{\delta p}_0\|$ by the definition (46), then (49) implies

$$\|P^{-1}\|_{op} \cdot \|\delta p_k\| \geq \|\tilde{\delta p}_k\| \geq \eta^k \cdot \rho_0 \|\tilde{\delta p}_0\| = \eta^k \cdot \rho_0 \|P^{-1}\delta p_0\|. \tag{51}$$

Thus, at least

$$k \geq \frac{-\log \varepsilon + \log(\rho_0 \frac{\|P^{-1}\delta p_0\|}{\|P^{-1}\|_{op}})}{-\log \eta} \tag{52}$$

iterations are required for ε-convergence of p_k. In a manner similar to before,

$$\lim_{\varepsilon \to 0^+} \frac{k}{-\log \varepsilon} \geq \lim_{\varepsilon \to 0^+} \frac{1 + \frac{\log(\frac{\rho_0 \|P^{-1}\delta p_0\|}{\|P^{-1}\|_{op}})}{-\log \varepsilon}}{-\log \eta}$$
$$= \frac{1}{-\log \eta} = \frac{1}{-\log \lambda_1} - a. \tag{53}$$

Next, we prove assumption (47) by induction. That is, that the relative weight of the first component cannot decrease beyond ρ_0. For $k = 0$ this is the definition of ρ_0. Assuming that (47) holds for k, we shall prove that it holds for $k + 1$. i.e., we shall prove

$$|\tilde{\delta p}_{k+1}^{(1)}| \geq \rho_0 \cdot \|\tilde{\delta p}_{k+1}\|. \tag{54}$$

To do so, it suffices to upper-bound $\rho_0 \cdot \|\tilde{\delta p}_{k+1}\|$ by some u, to lower-bound $|\tilde{\delta p}_{k+1}^{(1)}|$ by some l, and to provide a sufficient condition for $l \geq u$ to hold.

Notice that (48) provides a lower bound l to the left-hand side of (54), by using the induction assumption (47). To upper-bound the right-hand side of (54),

$$\|\tilde{\delta p}_{k+1}\| = |\tilde{\delta p}_{k+1}^{(1)}| + \sum_{i=2}^{n} |\tilde{\delta p}_{k+1}^{(i)}|$$

$$\overset{(41)}{\leq} \lambda_1 |\tilde{\delta p}_k^{(1)}| + \lambda_2 \sum_{i=2}^{n} |\tilde{\delta p}_k^{(i)}| + c\|\tilde{\delta p}_k\|^2$$

$$= \lambda_1 |\tilde{\delta p}_k^{(1)}| + \lambda_2 \left(\|\tilde{\delta p}_k\| - |\tilde{\delta p}_k^{(1)}| \right) + c\|\tilde{\delta p}_k\|^2$$

$$= (\lambda_1 - \lambda_2)|\tilde{\delta p}_k^{(1)}| + \lambda_2 \|\tilde{\delta p}_k\| + c\|\tilde{\delta p}_k\|^2 \tag{55}$$

Multiplying the latter by ρ_0 gives an upper bound u to $\rho_0 \cdot \|\tilde{\delta p}_{k+1}\|$. To prove (54), it remains to provide a sufficient condition for $l \geq u$ to hold,

$$|\tilde{\delta p}_k^{(1)}| \left[\lambda_1 - \frac{c}{\rho_0^2}|\tilde{\delta p}_k^{(1)}| \right]$$
$$\geq \rho_0 \left\{ (\lambda_1 - \lambda_2)|\tilde{\delta p}_k^{(1)}| + \lambda_2 \|\tilde{\delta p}_k\| + c\|\tilde{\delta p}_k\|^2 \right\}. \tag{56}$$

By the induction assumption (47), $\frac{1}{\rho_0}|\tilde{\delta p}_k^{(1)}| \geq \|\tilde{\delta p}_k\|$, which we use to upper-bound the right-hand side of (56). Thus, (56) is implied by the stricter,

$$\lambda_1 - \frac{c}{\rho_0^2}|\tilde{\delta p}_k^{(1)}| \geq \rho_0 \left\{ (\lambda_1 - \lambda_2) + \frac{\lambda_2}{\rho_0} + \frac{c}{\rho_0^2}|\tilde{\delta p}_k^{(1)}| \right\}. \tag{57}$$

This is equivalent to,

$$|\tilde{\delta p}_k^{(1)}| \leq \frac{\rho_0^2(1 - \rho_0)(\lambda_1 - \lambda_2)}{c(1 + \rho_0)}. \tag{58}$$

In a similar manner, the latter is implied by the stricter

$$\|\delta p_k\| \leq \delta_3 := \frac{\rho_0^2(1 - \rho_0)(\lambda_1 - \lambda_2)}{2c\|P^{-1}\|_{op}}, \tag{59}$$

where we have used $1 + \rho_0 \leq 2$, and $|\tilde{\delta p}_k^{(1)}| \leq \|P^{-1}\|_{op} \cdot \|\delta p_k\|$. That is, condition (59) is sufficient for the induction step (54) to hold.

Since our discussion focuses on the convergence of the Arimoto-Blahut algorithm, we may assume that $\|\delta p_{k+1}\| \leq \|\delta p_k\|$, for all $k \geq 0$, [2]. Therefore, it suffices to require that $\|\delta p_0\| \leq \delta_i$ for $i = 1, 2, 3$.

Finally, consider δ_1 (43), δ_2 (50) and δ_3 (59) as functions of ρ_0, $\delta_i = \delta_i(\rho_0)$, for $i = 1, 2, 3$. These are polynomials of zeroth, second and third order in ρ_0. They are strictly positive for $0 < \rho_0 < 1$, from their definitions. Given an initial condition p_0, $\delta_i(\rho(\tilde{\delta p}_0))$ is δ_i evaluated at the relative weight ρ of the first component (46), at the initial deviation $\tilde{\delta p}_0 := P^{-1}(p_0 - p_\beta)$. By (46), $0 \leq \rho(\tilde{\delta p}_0) \leq 1$ for any initial condition p_0, and so $\delta_i(\rho)$ are defined on the unit interval $[0, 1]$.

Let $B(\delta)$ be the ball of radius δ around p_β, and

$$\tilde{B}_i(\delta) := \left\{ p_0 \in B(\delta) : \|p_0 - p_\beta\| \leq \delta_i(\rho(\tilde{\delta p}_0)) \right\}, \tag{60}$$

for $i = 1, 2, 3$. Denote,

$$\tilde{B}(\delta) := \tilde{B}_1(\delta) \cap \tilde{B}_2(\delta) \cap \tilde{B}_3(\delta) \tag{61}$$

That is, $\tilde{B}(\delta)$ consists of those initial conditions p_0 for which the conditions (43, 50, 59) required along the proof are met. Clearly, $\tilde{B}(\delta) \subset B(\delta)$. We will show that $\tilde{B}(\delta)$ gradually fills the entire volume of $B(\delta)$ when $\delta \to 0$:

$$\lim_{\delta \to 0} \frac{\text{vol } \tilde{B}(\delta)}{\text{vol } B(\delta)} = 1, \tag{62}$$

where $\text{vol } S$ stands for the volume of a set S. It suffices to show this separately for each $\tilde{B}_i(\delta)$, $i = 1, 2, 3$.

Take $\tilde{B}_2(\delta)$ for example. We show that it contains a set whose volume approaches that of $B(\delta)$, as $\delta \to 0$. Consider initial conditions in the ball $B(\delta)$ by their value of $\rho(\tilde{\delta p}_0)$. Formally, we rewrite $\tilde{B}_2(\delta)$ as a disjoint union

$$\tilde{B}_2(\delta) = \bigcup_{0 \leq \rho \leq 1} \tilde{B}_2(\delta, \rho) \tag{63}$$

over the sets

$$\tilde{B}_2(\delta, \rho) := \left\{ p_0 \in \tilde{B}_2(\delta) : \rho(\tilde{\delta p}_0) = \rho \right\}. \tag{64}$$

These can be rewritten as,

$$\tilde{B}_2(\delta, \rho) =$$
$$\left\{ p_0 \in B(\delta) : \rho(\tilde{\delta p}_0) = \rho \wedge \|p_0 - p_\beta\| \leq \delta_2(\rho(\tilde{\delta p}_0)) \right\}$$
$$= \left\{ p_0 \in B(\delta) : \rho(\tilde{\delta p}_0) = \rho \wedge \|p_0 - p_\beta\| \leq \delta_2(\rho) \right\}$$
$$= \left\{ p_0 \in B(\min\{\delta, \delta_2\}) : \rho(\tilde{\delta p}_0) = \rho \right\} \tag{65}$$

where the first equality is by plugging in the definition (60) of $\tilde{B}_2(\delta)$.

Write (50) as $\delta_2(\rho) = C \cdot \rho^2$, for $C > 0$. It has a root at 0, and is otherwise positive. Thus, there are $\delta > 0$ with $\delta \leq \delta_2(\rho)$. For these δ, by (65)

$$\tilde{B}_2(\delta, \rho) = \left\{ p_0 \in B(\delta) : \rho(\tilde{\delta}p_0) = \rho \right\}. \tag{66}$$

Note that $\delta \leq \delta_2(\rho)$ is equivalent to $\sqrt{\delta/C} \leq \rho$. So by (63), $\tilde{B}_2(\delta)$ contains the set

$$\bigcup_{\sqrt{\delta/C} \leq \rho \leq 1} \left\{ p_0 \in B(\delta) : \rho(\tilde{\delta}p_0) = \rho \right\}. \tag{67}$$

If a particular δ value satisfies the above inequalities, then so does any smaller $\delta > 0$ value. At the limit $\delta \to 0$, $\tilde{B}_2(\delta)$ contains a union (67) over all ρ values, except for $\rho = 0$ which is of zero-measure. Since the coordinates transformation P is invertible, then the latter fills almost all the volume of $B(\delta)$ as $\delta \to 0$, as required for $\tilde{B}_2(\delta)$.

The argument for $\tilde{B}_1(\delta)$ and $\tilde{B}_3(\delta)$ is similar. ∎