Article

Not peer-reviewed version

Detecting Signatures of Criticality Using Divergence Rate

Tenzin Chan , De Wen Soh , Christopher Hillar *

Posted Date: 11 November 2024

doi: 10.20944/preprints202411.0663.v1

Keywords: Rate-Distortion; Optimal Coding; Criticality; Phase Transistion; Kullback-Leibler Divergence; Generalization

 doi:10.20944/preprints202411.0663.v1

Article

Detecting Signatures of Criticality Using Divergence Rate

Tenzin Chan [1] and De Wen Soh [2] and Christopher Hillar [3,*]

[1] Singapore University of Technology and Design; tenzin_chan@mymail.sutd.edu.sg
[2] Singapore University of Technology and Design; dewen_soh@sutd.edu.sg
[3] Algebraic, San Francisco, CA
* Correspondence: hillarmath@gmail.com

Abstract: Oftentimes in a complex system it is observed that, as a control parameter is varied, there are certain intervals during which the system undergoes dramatic change. Especially in biology, these signatures of criticality are thought to be connected with efficient computation and information processing. Guided by the classical theory of Rate-Distortion (RD) from information theory, we propose a measure for detecting and characterizing such phenomena from data. When applied to RD problems, the measure correctly identifies exact critical trade-off parameters emerging from the theory and allows for the discovery of new conjectures in the field. Other application domains include efficient sensory coding, machine learning generalization, and natural language. Our findings give support to the hypothesis that critical behavior is a signature of optimal processing.

Keywords: Rate-Distortion, Optimal Coding, Criticality, Phase Transistion, Kullback-Leibler Divergence, Generalization

1. Introduction

Criticality has been a hallmark of many fundamental processes in nature. For instance, in classical work, Landau [1] and others [2] investigated critical phase transitions between discrete states of matter. More generally, this concept has been used to describe a complex system that is in sharp transition between two different regimes of behavior as one varies a control parameter. Often, a critical point borders ordered and disordered behavior, such as what occurs in the classical 2D Lenz-Ising spin model of statistical physics at special temperatures [3]. For instance, if the system is too hot then usually disorder prevails; while too cold, and it freezes into a low entropy configuration. The tipping points between such regimes have been the subject of much study in science.

In the recent literature, the notion of criticality [4–6] has grown to encompass a computational element [7,8], with an emphasis on how to apply it to understand information processing in natural dynamical systems such as organisms [9,10]. In these applications, it is frequently argued that optimal properties for computation emerge [11–17] when systems are critically poised at some special set of control parameters. Of particular excitement is how criticality can help neuroscientists understand the brain, and also where such signatures have been observed [18–26]. Another exciting area where criticality may be present is in large neural networks which seem to generalize to certain problems at critical scales, training epochs, or dataset sizes [27,28].

Here, we leverage concepts from Rate-Distortion (RD) theory [29,30] to give insight into how normative principles governing a system [31], such as efficient coding [32,33], can lead to critical behavior. Consider a system A that evolves its internal steady-state dynamics through changes in a continuous control parameter t. The equilibrium behavior of A at t is modeled as a conditional distribution $p_t(y|x)$, which defines the probability that any given input x results in system state y. We say that A exhibits a critical phase transition about a special value t^* of the control variable if $p_t(y|x)$ changes dramatically near this parameter setting.

To quantify the change at a given t, we average over inputs x the Kullback-Leibler divergence (D_{KL}) between distributions $p_t(y|x)$ and $p_{t+\Delta t}(y|x)$, with Δt small. We call this a divergence rate for the evolution of the system's behavior along control parameter t. Given a sample of parameters t, we determine the significant peaks in the divergence rate, which we identify as critical control settings

 doi:10.20944/preprints202411.0663.v1

t^* for the system. As we shall see in several examples, local maxima in the divergence rate seem to coincide with locations of phase transitions in a system. Intruigingly, this measure can uncover higher-dimensional manifolds of criticality (Figure 3). This is the equivalent of finding the local maxima in the rate of change in evolving information-theoretic measures [34].

In the language of RD theory, these stochastic Input/Output systems $p_\beta(y|x)$ are called codebooks, and they determine – via a rate parameter β – a minimal communication rate R for encoding the state x as y given a desired average distortion D. More specifically, the control parameter $t = \beta$ indexes a pair (D_β, R_β) on the RD function, which separates possible from impossible encodings by a continuous curve in the positive orthant (see Figures 5 and 6).

For example, any lossy compression of a signal class has a particular rate and reconstruction error, which must necessarily lie on or above this curve. Intriguingly, the RD function contains a discrete set of special points $(D_{\beta^*}, R_{\beta^*})$ that correspond to critical control parameters β^* and codebooks $p_{\beta^*}(y|x)$ where some states y disappear or reappear in a coding. Intuitively, these are the behavioral phase transitions that must occur when traversing the RD curve from zero to maximal distortion.

In biology, the system A could be an organism that lossily encodes a signal through a channel bottleneck. For example, the retina communicates via the optic nerve to cortex, which is argued to be a several-fold compression of information capture from raw visual input [31]. We may also zoom out in scale to consider collections of organisms. In this setting, the theory of punctuated equilibrium in evolution [35] proposes that species are stable (in equilibrium) until some outside force requires them to change, during which species adapted quickly (punctuation). Here, the control parameter might correspond to some environmental variable in the habitat of the species, and sharp phase transitions [36,37] could emerge from some underlying normative dynamics [38–42].

We first validate our approach on classical RD problems with a number of states n small enough for comparison to mathematically exact calculations. We demonstrate that the critical β^* arising from theory match those that correspond to significant peaks in divergence rate. We also discover in a large-scale numerical experiment that there is an explicit relationship between counts of these critical control parameters and the number of Input/Output states. Namely, under mild assumptions, we conjecture that there are $n-1$ critical β^*, corresponding to $n-1$ critical codebooks $p_{\beta^*}(y|x)$ and critical pairs $(D_{\beta^*}, R_{\beta^*})$, for an RD function on n states (Conjecture 1). As another application, we show how experiments with our measure shed light on the Weak Universality Conjecture [43], which has implications for efficient systems in engineering and in nature.

We next explore several applications to more practical domains. In image processing, for instance, it is commonly desired to encode a picture with a small number of bits that nonetheless represents it faithfully enough for further computation down a channel. In human retina, this is thought to be accomplished by the firing patterns of ON and OFF ganglion neurons, which represent intensity values above or below a local mean, respectively. Using a standard database of natural images [44], we compute the RD function for ON/OFF encodings of small patches and study the structure of its critical β. We find that our measure uncovers several significant phase transitions for encoding these natural signals. Interestingly, the number of such critical points on the RD function seems to be significantly smaller than that predicted for a generic RD problem. This finding suggests that natural images define a special class of distributions [45] that might be exploited by visual sensory systems for efficient coding [46].

In machine learning, a common problem is to adapt model parameters for achieving optimal performance on a task. Clustering noisy data is one such challenge, and information theory provides tools for studying its solution [33,47]. We find that critical RD codebooks uncovered by divergence rate can reveal original cluster centers and their count. Another challenge is to store a large collection of patterns in a denoising autoencoder. We examine a specific example of robustly storing an exponential number of memories [48,49] in a Hopfield network [50] given only a small fraction of patterns as training input. We show that as the number of training samples increases, the divergence rate detects critical changes in dynamics, which allow the network to increase performance until generalizing

 doi:10.20944/preprints202411.0663.v1

to the full set of desired memories (graph cliques). Thus, we believe that the method of finding criticality described in this paper can be applied to understanding phase transitions in machine learning algorithms. For example, of particular interest is when large language models begin to be able to give the correct answers to certain types of questions [51]. This represents a type of generalization criticality in the space of model parameters.

As a final application, we study critical phenomena in writing. It has been pointed out that phase transitions arise in the geometry of language and knowledge [52–57]. We study agriculture during the 1800s in the United States using journal articles and uncover conceptual phase transitions across certain years. In particular, we find that an important shift in written expression occurs during the year 1840. Upon closer examination of the data, there were indeed significant language changes coinciding with the influences of war, religion, and commerce that were occurring at the time.

The outline of this paper is as follows. We give the requisite background for defining divergence rate in Section 2. Next, we explain in Section 3 findings from applying our measure to various domains such as RD theory, sensory coding, machine learning, and language. We close with a discussion in Section 5 and a conclusion in Section 6.

2. Background

We first define a measure of conditional distribution change over an independent control variable, which we call a divergence rate. As the control parameter varies, peaks in this divergence rate can be used to predict critical phase transitions in a system's behavior. Our methodology is inspired by the work of [54], who used the same measure to compute surprise in a sequence of successive debates of the French Revolution, and hence which topics tended to gain traction. We also give a brief background on rate-distortion theory, which provides a powerful class of examples to validate our approach and its utility as a tool for discovering new results in the field (e.g. Conjecture 1).

2.1. Definition of Measure

Let X be a set of data points with underlying distribution $q = (q_x)_{x\in X}$; that is, the probability of $x \in X$ is given by $q_x > 0$. Also, suppose for real parameters t and each $x \in X$, there are conditional distributions $p_t(y|x)$ specifying the probability of an output y given input x. The variable t could be a rate-distortion parameter β in RD theory, the size of a training dataset, the number of epochs for estimating a model, or simply time. In this work, we restrict our attention to discrete distributions so that X is a finite set.

Definition 1 (Divergence rate)**.** *The following nonegative quantity measures how much a system changes behavior from t to $t + \Delta t$:*

$$M_t = \frac{1}{\Delta t} \sum_{x\in X} q_x D_{KL}(p_t(y|x) \| p_{t+\Delta t}(y|x)).$$

The quantity $D_{KL}(u\|v) := \sum_y u_y \log(\frac{u_y}{v_y})$ is the Kullback-Leibler divergence of two distributions u and v, although other such functions could be used. The divergence rate is a simple proxy for the rate of change of the system's behavior. In particular, to detect critical t, we find local maxima in M_t across a range of values of the parameter t.

The following lemma validates the intuition that the divergence rate measures phase transitions.

Lemma 1. *Suppose that $p_t(y|x) \neq 0$ for any x, y and that $p_t(y|x)$ is differentiable at t. Then, the divergence rate M_t limits to zero as Δt goes to zero.*

 doi:10.20944/preprints202411.0663.v1

Proof. Set $\dot{p}_t(y|x) := \frac{dp_t(y|x)}{dt}$. Consider the quantity $\frac{1}{\Delta t} D_{KL}(p_t(y|x) \| p_{t+\Delta t}(y|x))$ with Δt small for a fixed x, which looks like:

$$\approx -\frac{1}{\Delta t} \sum_y p_t(y|x) \log\left(1 + \frac{\dot{p}_t(y|x)}{p_t(y|x)} \Delta t\right) \approx -\sum_y \frac{p_t(y|x)}{\Delta t} \frac{\dot{p}_t(y|x)}{p_t(y|x)} \Delta t = -\sum_y \dot{p}_t(y|x).$$

Since $\sum_y p_t(y|x) = 1$, we have $\sum_y \dot{p}_t(y|x) = 0$, so that the limit of M_t as Δt goes to zero is indeed zero. □

On the other hand, it is easy to verify by its definition that the divergence rate is infinite for any $t, \Delta t$ such $p_t(y|x) \neq 0$ for all x, y, but for which there is some state y with $p_{t+\Delta t}(y|x) = 0$.

2.2. Determining Critical Control Parameters

To locate critical control parameters of an evolving system $p_t(y|x)$, we find all significant local maxima produced by the measure. In practice, numerical imprecision or the presence of noise results in a divergence rate that has several local maxima which likely do not represent critical changes in system behavior. To filter out such false positives, we first normalize the set of divergence rates over samples by subtracting the mean value and then dividing by the standard deviation. This allows us to filter out peaks that are not significant (for example, peaks $\leq \alpha$ for some $\alpha > 0$). From here on, divergence rate will refer to this normalized divergence rate.

In practice, the divergence rate M_t is potentially non-concave and since t is sampled, we are left with finding peaks in a piecewise linear function. To this end, we utilize a discrete version of the Nesterov momentum algorithm [58] to find significant peaks. We find that the approach is fairly insensitive to hyperparameters when implemented. We provide explicit details of this method and its pseudocode in Section 4.

2.3. Rate-Distortion Theory

Rate-distortion (RD) theory [59, Chapter 10] has frequently been used to analyze and develop lossy compression methods [60] for images [61], video [62], and even memory devices [63]. It also offers some perspective on how biological organisms organize themselves based on their perceptions of the world around them [64]. For example, rate-distortion theory has been used to understand fidelity-efficiency tradeoffs in sensory decision-making [42] and has also been useful in understanding relationships between perception and memory [65]. And at a molecular level, it has been used to study the communication channels and coding properties of proteins [39,66].

Suppose that we have a system with n states $X = \{1, \ldots, n\}$ and an input probability distribution $q = (q_1, \ldots, q_n)$ defined on X. We seek a deterministic coding of x to y, with the overall error arising from this coding determined by a distortion function $d(x, y) \geq 0$, which expresses the cost in setting x to y (typically, we assume $d(x, x) = 0$ for all x).

Remarkably, provided only with q and d, the Blahut-Arimoto algorithm [67,68] produces the convex curve of minimal rate $R(D)$ over all deterministic codings having a fixed level of average distortion D. For example, when $D = 0$, so that error is not allowed in a coding, the optimal rate is the Shannon entropy $H(q)$ of the input distribution q. In general, when $D > 0$, the minimal rate achievable for a coding with average distortion D is less than $H(q)$ and is determined by a unique point on RD curve. This is intuitive as a sacrifice of some error in coding should yield a dividend in a better possible rate.

Although the RD curve determines a theoretical boundary between possible and impossible deterministic coding schemes, the curve itself is determined via an optimization over stochastic codings, which do not directly lend themselves to practical deterministic implementations. Nonetheless, computing the RD curve is still useful for understanding the complexity of a given problem and benchmarking. See Figures 5 and 6 for examples of RD curves in various settings.

 doi:10.20944/preprints202411.0663.v1

Mathematically, the RD curve is parameterized by a real number β which indexes an optimal stochastic coding of x to y as a codebook or matrix of conditionals $p_\beta(y|x) \geq 0$ (so that $\sum_y p_\beta(y|x) = 1$). The average distortion is given by $\sum_{x,y} q_x p_\beta(y|x) d(x,y)$. We also set $p_\beta(y) = \sum_x q_x p_\beta(y|x)$ to be the corresponding distribution for a given output y.

It is classical, using Lagrange multipliers, that points on the RD curve arise from solutions to a set of equations in the codebooks and output distributions. For expositional simplicity in the following, we drop β in subscripts and set $w = e^{-\beta}$. We also identify $p(y)$ and $p(y|x)$ with p_i and p_{ij}, respectively. The governing equations for the RD curve are then given by the following.

Definition 2. *The Blahut-Arimoto (BA) equations for the RD curve are:*

$$p_i = \sum_{j=1}^{n} q_j p_{ij},\ Z_j = \sum_{i=1}^{n} p_i w^{d_{ij}},\ p_{ij} = p_i w^{d_{ij}} / Z_j. \tag{1}$$

When the numbers d_{ij} are nonnegative integers, solutions to the BA equations form a real algebraic variety, which allows them to be computed symbolically using methods of computational algebra [69].

Example 1 (Two states)**.** *When $n = 2$, the three equations translate to these for p_1 and p_2:*

$$\begin{aligned} p_1 &= p_1 \left(\frac{q_1 w^{d_{11}}}{p_1 w^{d_{11}} + p_2 w^{d_{21}}} + \frac{q_2 w^{d_{12}}}{p_1 w^{d_{12}} + p_2 w^{d_{22}}} \right), \\ p_2 &= p_2 \left(\frac{q_1 w^{d_{21}}}{p_1 w^{d_{11}} + p_2 w^{d_{21}}} + \frac{q_2 w^{d_{22}}}{p_1 w^{d_{12}} + p_2 w^{d_{22}}} \right). \end{aligned} \tag{2}$$

More generally, set W to be the $n \times n$ matrix defined by $W_{ij} = w^{d_{ij}}$. (Note that W is invertible near $w = 0$.) Then, equations (1) can be combined into one, as follows:

$$p = p \odot \left[W \left(\frac{q}{W^\top p} \right) \right]. \tag{3}$$

Here, $\odot$ is the point-wise (Hadamard) product of two vectors, $\left(\frac{u}{v}\right) := (u_1 v_1^{-1}, \ldots, u_n v_n^{-1})$ is the element-wise ratio of the vectors u and v, and $W^\top$ is the transpose of W. Although somewhat surprising given its form, the right-hand side of (3) is indeed in the simplex.

Interestingly, when the BA equations are written in this form, we can identify the RD curve as a fixed-point using Browder's fixed point theorem [70,71]. We summarize this finding with the following useful characterization.

Proposition 1. *The RD curve is given by the following equivalent objects.*

1. *The minimization of the RD objective function.*
2. *Browder's fixed-point for the function on the right-hand side of (3).*
3. *Solutions to the BA equations.*

Proof. The equivalence of the first and third statement is classical theory. For the equivalence of the second and the third, consider the map from an interval cross the simplex to the simplex given by $f(w,p) = p \odot \left[W \left(\frac{q}{W^\top p} \right) \right]$. By Browder's fixed-point theorem, this has a fixed-point given by a curve p_w satisfying $f(w, p_w) = p_w$, which is equation (3). □

When some of the indeterminates p_i become zero at some β, then the BA equations still makes sense, we simply end up with fewer variables. Such equations involving inverses of linear forms also appear in work on the entropic discriminant [72] in algebraic geometry and maximum entropy graph distributions [73].

 doi:10.20944/preprints202411.0663.v1

When β is large (w near 0), optimal codebooks are close to the identity, corresponding to average distortion near zero. In this case, it can be shown directly using equations (3) that there is an explicit solution [30]. Set $1 = (1, \ldots, 1)$ to be the all-ones vector.

Lemma 2. *The exact solution for small w (large β) of the RD function is given by:*

$$p = W^{-\top}\left(\frac{q}{W^{-1}1}\right). \tag{4}$$

Proof. When β is very large, all p_i are nonzero. In this case, we may cancel p on both sides of equation (3) to give:

$$1 = W\left(\frac{q}{W^{\top}p}\right) \implies W^{-1}1 = \left(\frac{q}{W^{\top}p}\right) \implies \left(\frac{q}{W^{-1}1}\right) = W^{\top}p.$$

The lemma now follows by multiplying both sides of this last equation by $W^{-\top}$. □

It turns out that the rate-distortion curve has critical points on it where dramatic shifts in codebooks occur, an observation that was one of our inspirations. For the purposes of this work, these critical β^* are defined when some p_i goes from positive to zero (or visa-versa) as β varies.[1] Note that the third equation in (1) implies this is when p_{ij} also has such a change. Given this definition, Lemma 1 shows that divergence rate characterizes these critical points.

The first critical β^* can be determined using Lemma 2 as follows.

Corollary 1. *The first critical β^* on the RD curve going from $\beta = \infty$ to $\beta = \beta^*$ is the first value of β that makes some entry p_i of the vector in equation (4) equal to zero.*

Proof. Equation (4) describes the values of p_i along the RD function as long as we are allowed to cancel p from both sides of equation (3). From continuity, the first time a p_i becomes zero is governed by when (4) first has this happen. □

In the next section, we shall continue with our two state Example 1 above and compare this theoretical calculation from Corollary 1 to the first critical β found by our criticality measure (see Figure 1).

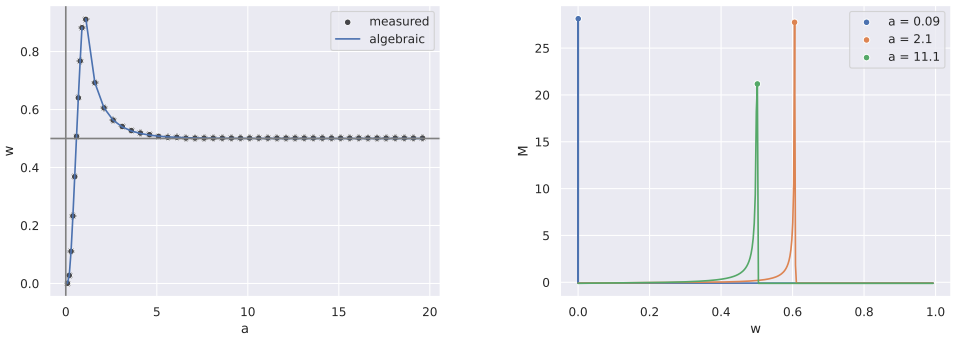

Figure 1. Critical RD parameters $w^* = \exp(-\beta^*)$ determined mathematically from the theory (algebraic) match up exactly with significant peaks in divergence rate (measured) (Left). Plotting M vs w for three different values of a shows the single peak in the divergence rate M (Right).

[1] We remark that in the literature, another equivalent definition is usually used, which is when the derivative of the RD curve has a discontinuity.

 doi:10.20944/preprints202411.0663.v1

3. Applications

We present applications of finding peaks (Section 2.2) in the divergence rate (Definition 1) to the discovery of phase transitions in various domains such as rate-distortion theory, natural signal modeling, machine learning, and language.

3.1. Rate-Distortion Theory

Recall that given a distribution q and a distortion function d, the rate-distortion curve characterizes the minimal rate of a coding given a fixed average distortion. A point on the curve (D_β, R_β) is specified by a parameter β and is determined by an underlying codebook $p_\beta(y|x)$. In particular, the setup for RD theory can be seen as a direct application of our general framework with the control parameter $t = \beta$.

In the following subsections, we shall validate our approach to finding critical phase transitions by comparing our peak finder estimates of critical β with those afforded by RD theory. In cases where the number of states is small, such as $n = 2$ and $n = 3$, we can compute explicit solutions to the RD equations (1) and compare them with those estimated by our divergence rate approach (see Figures 1, 2, 3).

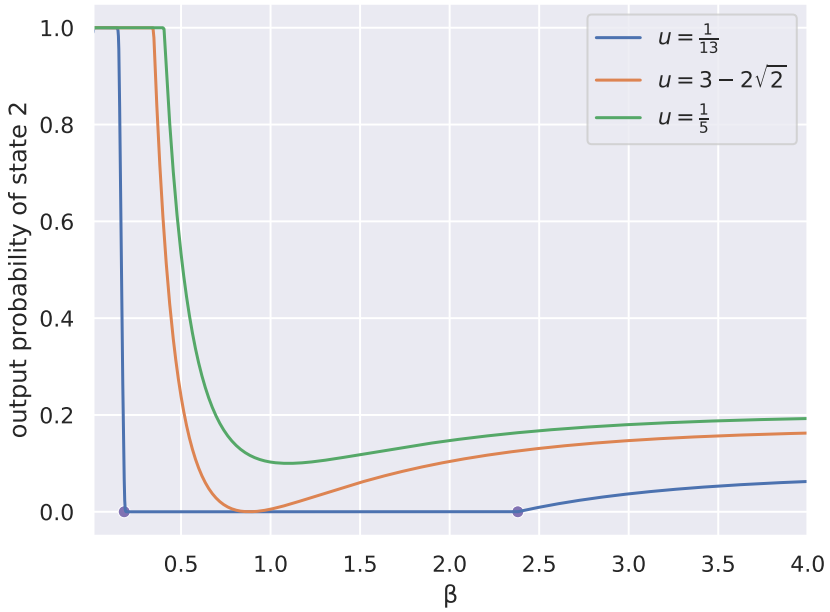

Figure 2. Consider $q = ((1-u)/2, u, (1-u)/2)$ for a parameter u and a fixed distortion d given by (5). For each $u = 1/13, 3 - 2\sqrt{2}, 1/5$, we plot the output probabilities of the second state as a function of β, determined by the exact theory. The circles were determined from finding significant peaks in divergence rate at $u = 1/13$ and match the vanishing and re-emergence of state 2.

 doi:10.20944/preprints202411.0663.v1

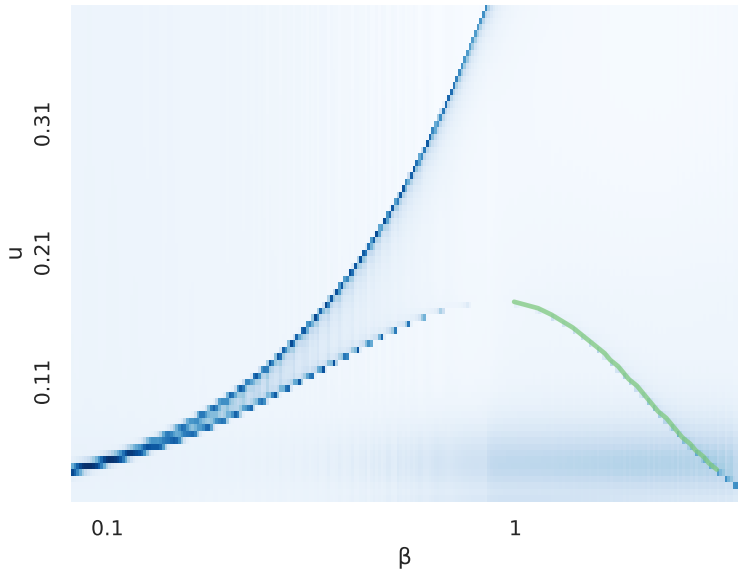

Figure 3. We plot a heatmap of the divergence rate as a function of u and β. We also plot the theoretical calculation matching the lower-right part of the curve in the heatmap in green.

After demonstrating that peaks in divergence rate agree with theory in these cases, we next turn our attention to exploring what our criticality tool can uncover theory-wise for the discipline. In particular, our experiments suggest new results for RD theory, such as that for a random RD problem, the number of critical β^* on n states goes as $n-1$ (Conjecture 1). We also use our framework to help shed light on a universality conjecture [43] inspired by tradeoffs in sensory coding for biology (Conjecture 2).

3.1.1. Exact RD on Two States

The case of $n = 2$ states with varying distortion is already interesting. Consider a source distribution $q = (1/2, 1/2)$ and distortion matrix for fixed $a \neq 0$ given by:

$$d = \begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix}.$$

In this case, using Lemma 2, one can explicitly solve the equations for the output distributions p_1, p_2 before the first critical point in terms of β (recall, $w = e^{-\beta}$):

$$p_1 = \frac{1 - 2w^a + w^{a+1}}{2(1-w)(1-w^a)},\ p_2 = \frac{1 - 2w + w^{a+1}}{2(1-w)(1-w^a)}.$$

From Corollary 1, the first critical β^* is determined when one of these expressions becomes zero. One can check that for $a > 1$, p_2 first has a zero value for some critical β. At this value, p_1 is automatically determined to be $p_1 = 1$. For $a < 1$, on the other hand, it is p_1 that becomes zero.

We consider the case $a > 1$, as the other is similar. Our main tool is the generalized version of Descartes' Rule of Signs [74,75]. This rule says that an expression in a positive indeterminate w, such as $1 - 2w + w^{a+1}$ from above, has a number of positive real zeroes bounded above by the number s of sign changes of its coefficients. Moreover, the true number of positive zeroes can only be s, $s-2$, ..., or $s - 2\lfloor s/2 \rfloor$.

 doi:10.20944/preprints202411.0663.v1

In our particular setting, we have $s = 2$ so that there can only be two or no positive zeroes. As $w = 1$ is a zero, it follows that there is some other positive number w^* making this expression zero. It follows that our first critical point is $\beta^* = -\ln(w^*)$. In particular, when $a = 2$, we have:

$$\beta^* = -\ln\left(\frac{\sqrt{5}-1}{2}\right).$$

Given these explicit calculations from the theory in hand, we would like to compare them with critical parameters determined from significant peaks in the divergence rate.

We summarize our findings in Figure 1. To produce the "measured" points in this plot, we vary the distortion measure using the parameter a, compute codebooks parameterized by β with the BA algorithm, and then find critical values of $w = \exp(-\beta)$ at which the divergence rate peaks. At all values of $a \neq 1$, the critical value of w found by the noisy peak-finder matches the corresponding exact value from the theory (the "algebraic" line). Note that near $a = 1$, the critical β^* approaches zero but is not defined there. Also from the plot, one can guess that β^* approaches $-\ln 2$ (resp. infinity) as a goes to infinity (resp. zero). This can also be verified directly from the theoretical calculations above.

3.1.2. Exact RD on Three States

We now consider a slightly more complicated example from Berger [76] with $n = 3$ states. In this case, we assume a varying source $q = ((1-u)/2, u, (1-u)/2)$ for a parameter $u \in (0,1)$, but we fix a distortion matrix:

$$d = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}. \tag{5}$$

Again using Lemma 2, we find that for large β, the output distributions on the RD curve are given by:

$$p = \frac{1}{2(w-1)^2}\begin{bmatrix} (u+1)w + u - 1 \\ w^2 + (u-1)w + u \\ (u+1)w + w - 1 \end{bmatrix}. \tag{6}$$

For example, when $u = 1/13$, this becomes:

$$p = \frac{1}{13(w-1)^2}\begin{bmatrix} 6 - 7w \\ 1 - 12w + 13w^2 \\ 6 - 7w \end{bmatrix}.$$

In this special case, one can check that the first critical w^* occurs when p_2 becomes zero; that is, when:

$$w^* = \frac{1}{13}(6 - \sqrt{23}),\ \beta^* = -\ln(w^*).$$

More generally, one can check that as long as $u \leq 3 - 2\sqrt{2}$, a critical w^* is determined from the formula:

$$w^* = \frac{1}{2}(1 - u - \sqrt{1 - 6u + u^2}).$$

In Figure 2, we plot for three different settings of $u = 1/13, 3 - 2\sqrt{2}, 1/5$, the values of the output distribution p_2 for the second state as a function of β, computed from the general solution (6). On the same plot, we place two circles on the x-axis where we find β representing peaks in the divergence rate in the case of $u = 1/13$. Notice they are located where the output probability of state two goes to zero or emerges from zero.

Figure 3 is obtained by plotting the divergence rate as a heatmap against u and β. The peak of the bottom convex shape corresponds to the value of $u = 3 - 2\sqrt{2}$ where a discontinuity appears in the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2024 doi:10.20944/preprints202411.0663.v1

number of critical β given a fixed u. Also shown in Figure 3 is a theoretically determined part of this curve, which matches precisely the estimated lower-right piece determined only using the divergence rate.

3.1.3. Criticality for Generic RD Theory

With these examples as evidence that the empirical divergence rate is able to detect critical β for RD theory, we explore its potential implications for the general case as we increase the number of states n.

We next consider random RD problems on n states in which the probabilities for q are chosen independently from a uniform distribution in the interval $[0, 1]$ (and then normalized to sum to one) and the distortion matrices are chosen to have diagonal zero and other entries d_{ij} drawn independently and uniformly in the interval $[0, 4]$. (Our experiments appear to be insensitive to the particular underlying distributions used to make q and d.)

In Figure 4, we vary n and plot the number of significant peaks in divergence rate for these random RD problems, averaged over 10 trials. The divergence rate is imperfect and thus there are error bars. The accuracy of the divergence rate depends highly on whether the range of β is adequately densely-sampled and that the range of samples includes all critical β. It should be noted that this method of finding critical points is an estimate and therefore may indicate large changes in the codebooks where there is no theoretical critical point.

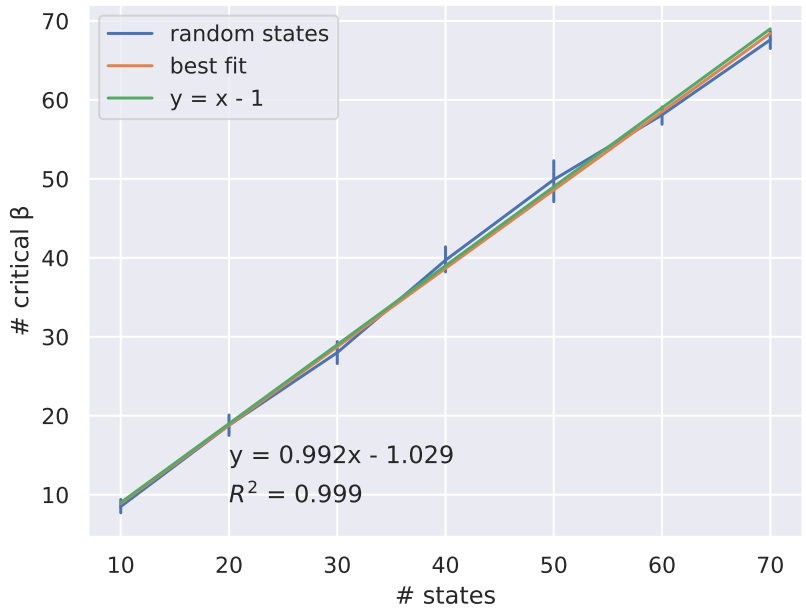

Figure 4. The number of peaks in the divergence rate across different values of β for varying numbers of states n is close to the line $n - 1$.

In particular, the numerical results shown in Figure 4 suggest the following new conjecture in the field of RD theory.

Conjecture 1. *Generically, there are $n - 1$ critical β^* along the rate-distortion curve as a function of the number of states n.*

Here, the word "generic" is used in the sense that an RD problem with data distribution q and distortion matrix d chosen at random will have this number of critical points. More precisely, if

 doi:10.20944/preprints202411.0663.v1

$d_{xy} = d(x, y)$ and q_x are $n^2 - 1 = (n^2 - n) + (n - 1)$ indeterminates (set one of the q_x to be 1 minus the sum of the others), then outside of a measure zero set in $\mathbb{R}^{n^2-1}$, the conjecture holds.

3.1.4. Weak Universality

Consider the case where all n states have the same probability $q_x = 1/n$ and two distortion matrices d_1, d_2 have entries drawn from the same distribution, say uniform in $[0, 1]$. Then a conjecture in RD theory, called Weak Universality, says that the RD curves for d_1 and d_2 lie on top of each other for large n.

Conjecture 2 (Weak Universality [43]). *Distortion matrices with entries drawn i.i.d from a fixed distribution ϕ define rate-distortion curves that are (universal) functions of ϕ in the limit of large numbers of states.*

To illustrate weak universality in Figure 5, we took two randomly drawn distortion matrices d_1 and d_2 and plotted their RD curves, computed using the BA algorithm. We then plotted the critical points found from determining peaks in divergence rate on each of the two RD curves. As predicted by Conjecture 2, the RD curves are close. However, notice that the two sets of critical points do not align with each other.

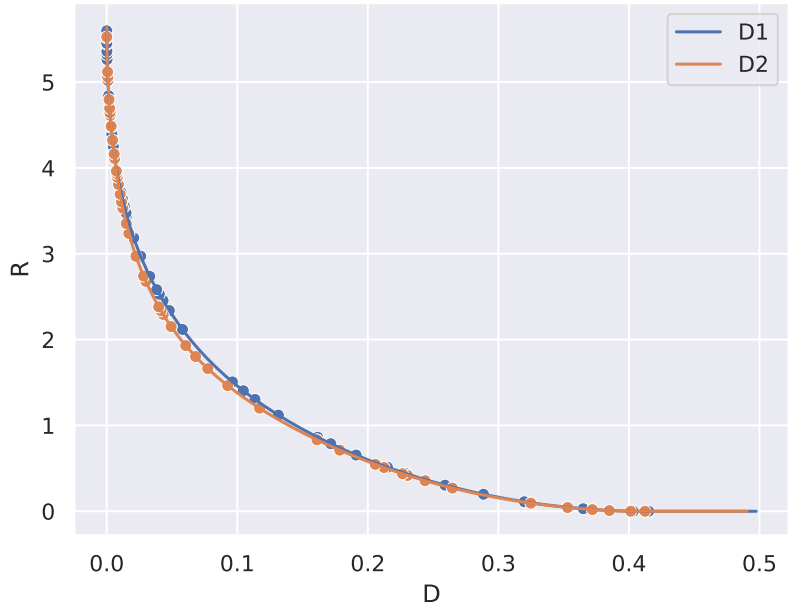

Figure 5. Taking two random RD distributions D1 and D2 gives two RD curves which look similar, but whose critical points differ (points indicated along the curves), estimated from divergence rate maxima.

This experiment suggests that we can rule out a line of attack for proving Weak Universality that attempts to show critical points on the RD curve also coincide with each other. In particular, this stronger form of the conjecture is likely not true.

3.2. Natural Signals

Discrete ON/OFF encodings of image patches have been shown to contain much perceptual information [77], and Hopfield Networks trained to store these binary vectors can be used for high quality lossy compression of natural images [78,79].

It was also observed in [80] that these ON/OFF distributions of natural patches exhibited critical β^* at a few special points and codebooks along the RD curve. Applying our noisy peak finder to the divergence rate on the codebooks at rate parameters β, we are able to discover all the critical codebooks

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2024 doi:10.20944/preprints202411.0663.v1

that were previously found by hand, as well as four more which have deviations in microstructure. We remark that it is also possible that some of these extra critical codebooks found may have arisen from imperfections in the tuning of the peak finder.

To obtain Figure 6, two-by-two pixel patches x are ternarized by first normalizing the pixel values within each patch to have variance one, and then each normalized value v is mapped to a ternary value b via the following rule:

$$b = \begin{cases} (0,1), & \text{if } v < -0.5, \\ (0,0), & \text{if } -0.5 \leq v \leq 0.5, \\ (1,0), & \text{if } v > 0.5. \end{cases}$$

We use as source q the probability distribution over this ternary representation of patches, and the Hamming distance between two binary vectors specifies the distortion matrix. Codebooks along the RD curve are then used as conditional distributions and peaks in the divergence rate are plotted as the 9 blue circles.

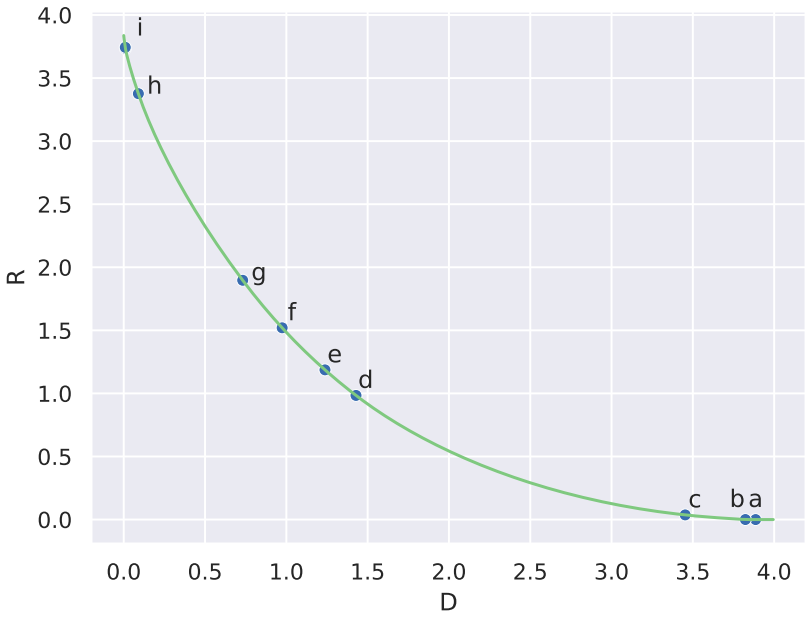

Figure 6. RD curve and locations of critical codebooks for 2×2 ON/OFF natural image patches (compare with [80, Fig. 9a]).

In Figure 7, we display the codebooks that arose from the critical β^* found using the divergence rate. These codebooks correspond to the lettered points in Figure 6.

 doi:10.20944/preprints202411.0663.v1

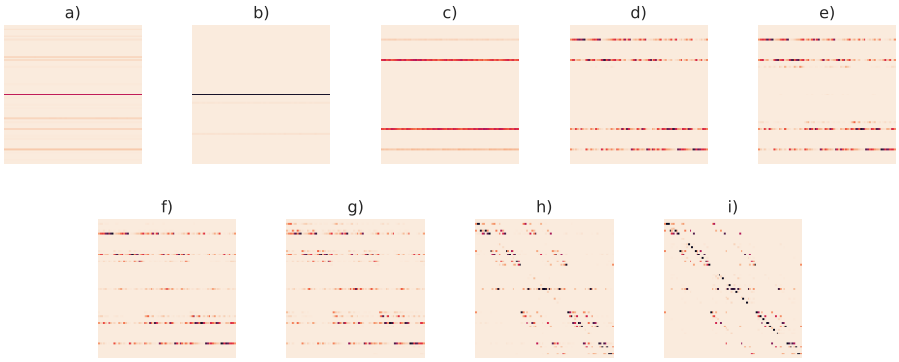

Figure 7. Natural image patch codebooks at critical β from Figure 6 estimated using divergence rate (compare with [80, Fig. 9c]). Each column represents the conditional distribution given an input state.

It is interesting to note that there are an order of magnitude fewer critical codebooks estimated than would be expected by Conjecture 1. This suggests that this set of natural signals does not represent a random or generic RD problem, indicating extra structure in the signal class.

3.3. Machine Learning

We consider a warm up problem on 18 states to test the divergence rate on the problem of clustering. The data contains three disjoint groups, each with six states, that have low intra-cluster distortion but high inter-class distortion.

We apply the noisy peak finder to the divergence rate of codebooks parameterizing the RD curve for this setup, and we plot a few of the interesting critical codebooks in Figure 8. We notice that critical codebooks code for when the cluster centers are found, and when the code progressively breaks away towards the identity codebook.

Figure 8. Critical codebooks for clusters corresponding to critical points determined by peaks in divergence rate. Each column is the conditional distribution given an input state.

Our next machine learning example comes from the theory of auto-encoding with Hopfield networks. First, we determine networks to store cliques as fixed-point attractors (memories) using Minimum Energy Flow (MEF) learning, as in [48,49]. Then, we look for peaks in divergence rate as networks are trained with different numbers of samples s. Our hypothesis is that as s is varied, there should be some critical s where the nature of network dynamics changes drastically.

For the clique example in Figure 9, the possible data consist of binary vectors representing absence or presence of an edge in a graph of size 16, with each sample being a clique of size 8.

 doi:10.20944/preprints202411.0663.v1

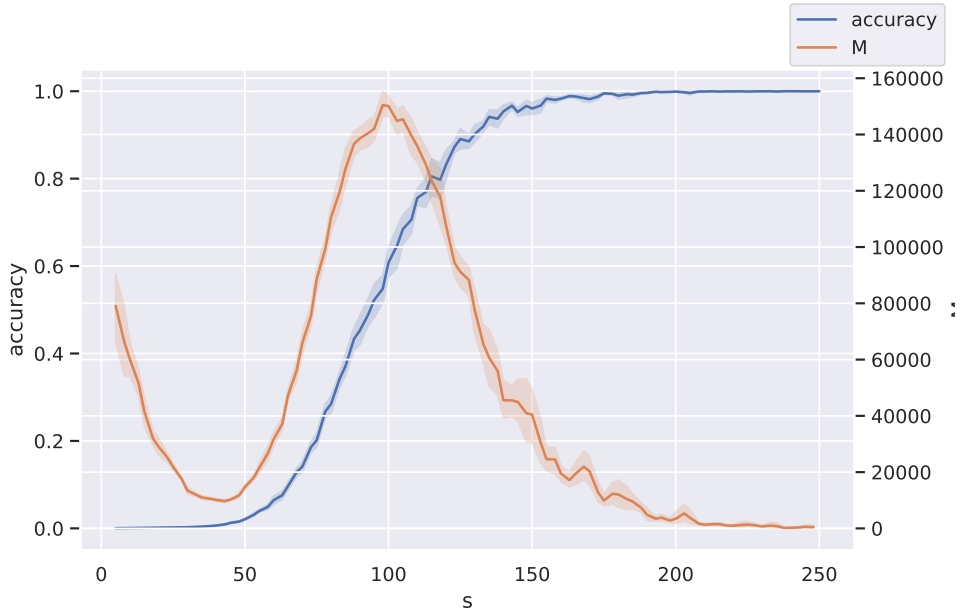

Figure 9. Peaks in divergence rate suggest networks exhibit phase transitions along MEF learning.

By randomly sampling a different set of training data for each trial, we obtain a codebook for each number of training samples as the deterministic map on binary vectors given by the dynamics of the Hopfield network. Specifically, the codebook $p_s(y|x)$ is the $0/1$-matrix of the output of the dynamics determined by the association $x \mapsto y$ of the sth network; that is, each map forms a conditional matrix entry $f_s(y|x)$ that is $\frac{1}{|X|+1}$ if the network dynamics takes state x to y, and zero, otherwise, where $|X|$ is the total number of cliques. We use one output state state to code for all non-cliques, leading to $|X|+1$ in the denominator.

We then plot the accuracy (proportion of cliques correctly stored by the network) in blue and the normalized divergence rate in orange against s. We averaged over 10 trials in the experiment.

A significant peak in the divergence rate as a function of s can be observed in Figure 9. This suggests that is a critical sample count for auto-encoding in Hopfield networks trained using MEF. Note that this peak appears to occur when half of the test data is accurately coded.

That signatures of criticality in self-organizing recurrent neural networks arise during training is also corroborated by [81]. It would be interesting to further explore the changes that happen near the critical sample count in Figure 9.

3.4. Language

In [55], the possibility of punctuated equilibrium and criticality in language evolution was studied using a similar approach based on the divergence rate. Naively using the divergence rate on raw word frequency distributions over time, we can also obtain a measure of language change over time. In this case, the divergence rate is over a single condition (fixed topic). We show that peaks are also observed in this measure of a language dataset; see Figure 10a, which indicates that critical changes are not only common, but a natural part of the way humans solve problems in communication.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2024 doi:10.20944/preprints202411.0663.v1

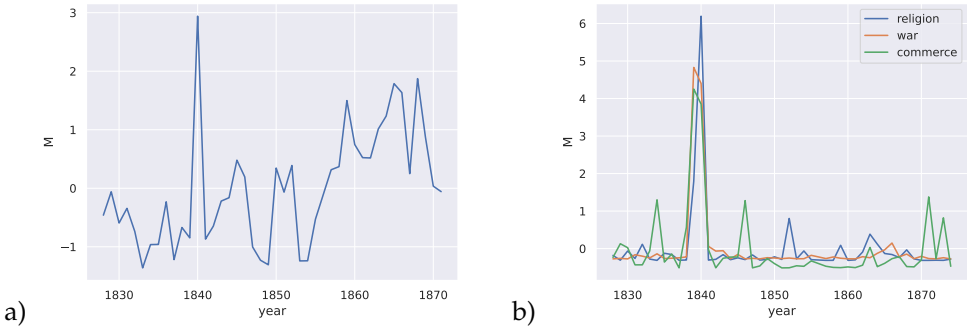

Figure 10. a) Normalized plot of divergence rate for a horticultural journal corpus against time. b) Peaks along time in the corpus associated to the indicated words.

For the language example, the codebook is conditioned on a single state, which is the topic of the dataset; namely, agriculture in America. This means we only consider the distribution of words of one topic over time as the independent variable t. The divergence rate is thus applied over a single state in the language concept space.

To discover what the underlying cause of this large change could be, we clustered the words in the corpus using Word2Vec [82] and find several interesting clusters. We then chose semantically similar words from those clusters, which represent concepts in war, religion and commerce. Again, by conditioning on words from these subtopics, we use the divergence rate to obtain the changes in the probability distributions of these clusters individually in Figure 10b. We notice that peaks occur exactly in the year of the most significant peak in Figure 10a.

Investigating possible reasons why this might be the case, we looked to the history of agriculture in the United States of America. In [83], Frolik mentioned that many innovations and technologies came to bear on the American agriculture industry (such as the cast iron plow, manufacturing of drills for sowing seeds, and horse-powered machines for harvesting grain) around the years leading up to and including the 1840s, which led to a booming agricultural industry. This possibly accounts for the peak in usage of words related to commerce.

4. Methods

The measure $M(f_t)$ is the sum of the distance over $x \in X$ between probability distributions f conditioned on x at different points t and $t+1$. It tracks the change in f over t. In our experiments, we used the Kullback-Leibler (KL) divergence for D.

$$M(f_t) = \frac{1}{\Delta t} \sum_{x \in X} D(f_t(y|x), f_{t+\Delta t}(y|x))$$

The value of t here could be replaced by some independent variable, such as β (in the case of the RD curve), size of training dataset, number of epochs (for training a model), or time. The codebook of a model f (for a discrete state space) is a distribution conditioned on each input state. For a set of models with random initializations, we normalize the distribution of the maps with respect to the inputs over all the maps. The measure over t is then obtained by taking the distance between codebooks between pairs of consecutive codebooks. Critical phase transitions are then where peaks occur along this 1-dimensional signal.

The measure $M(f_t)$ is potentially non-concave and the variable t is usually sampled at discrete intervals, thus producing a piecewise linear function. Thus, to detect local maxima in the measure which represent points in t where the measure changes significantly (a critical point in t), we need a noisy peak finder, as there may be spurious local maxima which do not correspond to any critical change in behavior. To this end, we develop a discrete version of Nesterov momentum algorithm [58] in Algorithm 1.

 doi:10.20944/preprints202411.0663.v1

Algorithm 1 Noisy local maxima finder

Require: List of 2D points p
Step size η
Momentum μ
Convergence threshold ϵ

Ensure: p is sorted by increasing x values
$\epsilon > 0$
$q \leftarrow$ Compute potential local maxima with **Find All Local Maxima** with input p
$u \leftarrow$ points in p with $p_i.y \leftarrow -p_i.y$
$r \leftarrow$ Compute local minima with **Find All Local Maxima** with input u
$i \leftarrow 1$
$\theta_0 \leftarrow r_i.x$
$o \leftarrow$ new empty set of 2D points
while $\theta \neq r_n.x$ **do**
  $\theta \leftarrow$ compute next local maxima with **Find Next Local Maxima** with inputs p, q, θ, η, μ and ϵ
  Add θ to o
  for $i = i + 1, i + 2, \ldots, r.length$ **do**
    if $r_j.x - \theta > 0$ **then**
      $i \leftarrow j$
      $\theta \leftarrow r_i.x$
      break
    end if
  end for
end while

Algorithm 2 Find All Local Maxima

Require:
List of 2D points p
$o \leftarrow$ new empty list of 2D points
for $i = 2, 3, \ldots, p.length - 1$ **do**
  if $p_{i-1}.y < p_i.y > p_{i+1}.y$ **then**
    Add p_i to o
  end if
end for
return o

 doi:10.20944/preprints202411.0663.v1

Algorithm 3 Find Next Local Maxima

Require:
List of 2D points p
List of 2D local maxima points q
Current x-value θ
Step size η
Momentum μ
Convergence threshold ϵ
$\theta_{prev} \leftarrow \infty$
$\phi_{prev} \leftarrow 0$
$\phi \leftarrow 0$
while $abs(\theta - \theta_{prev}) < \epsilon$ **do**
  $\theta_{prev} \leftarrow \theta$
  $\Delta_\theta \leftarrow$ compute gradient with **Piecewise gradient** with input p and θ
  $\phi_{prev} \leftarrow \phi$
  $\phi \leftarrow \theta + \eta\Delta_\theta$
  $\theta \leftarrow \phi + \mu(\phi - \phi_{prev})$
end while
$o \leftarrow q_1.x$
for $i = 1, 2, \ldots, q.length$ **do**
  if $abs(\theta - q_i.x) < abs(\theta - o)$ **then**
    $o \leftarrow q_i.x$
  end if
end for
return o

Algorithm 4 Piecewise gradient

Require:
List of 2D points p
Point to compute gradient at t
$o \leftarrow 1$
for $i = 2, 3, \ldots, p.length - 1$ **do**
  if $abs(t.x - p_i.x) < abs(t.x - p_o.x)$ **then**
    $o \leftarrow i$
  end if
end for
$g \leftarrow \frac{p_{o+1}.y - p_o.y}{p_{o+1}.x - p_o.x}$
return g

Given a distortion measure (non-negative matrix describing costs of encoding an input state with an output state), the rate-distortion curve describes the minimum amount of information required to encode a given distribution of symbols at a level of distortion specified by the gain parameter β. There is a corresponding codebook at any β, which describes a distribution over all symbols X for each symbol x with which to optimally code for x. This allows us to directly use the measure, using β as the independent variable t.

For the simple two-state rate-distortion example in Figure 1, we vary the distortion measure using a parameter a and find the critical value of $y = exp(-\beta)$ at which the measure peaks. At all values of a, the critical value of β found by the noisy peak-finder matches the analytic critical value of β. Note that there is no analytic critical value of β at $a = 1$, and thus there is a gap in the plot.

For the three-state case proposed by Berger in Figure 2 and 3, the distortion matrix is held constant and we vary the probability distribution q. As an example, Figure 2 shows the marginal probability of state 2 at $u = \frac{1}{13}$, and the points show where the peak-finder detects a critical change, which are visually exactly where the marginal probability goes to or emerges from 0. Figure 3 was made by plotting the measure as a heatmap against u, β.

 doi:10.20944/preprints202411.0663.v1

To arrive at Figure 4, we generate a random distortion matrix for a uniform distribution on k states for each trial. We vary k and plot the number of critical β against k for the RD curve on this setup. We perform 10 trials to obtain this plot.

To illustrate weak universality in Figure 5, we took two probability vectors P1 and P2 drawn from the same distribution with corresponding distortion matrices D1 and D2 also drawn from the same distribution. We then plotted the critical points found by the measure on the RD curve.

To test this measure on the coding of clusters, we consider a 18-state distribution, with disjoint groups of six states that have low intra-cluster distortion, but high inter-class distortion. We then apply the noisy peak finder to the measure on the RD curve for this setup and plot some of the interesting critical codebooks in Figure 8 and the corresponding critical points on the RD curve and measure in Figure 8.

For the clique example in Figure 9, the data is binary vectors representing absence or presence of an edge in a graph, with each sample being a clique of size $\frac{v}{2}$, where v is the number of nodes in the graph, taken to be 8 in our experiment. The distortion matrix is computed using Hamming distance. By randomly sampling a different set of training data for each trial, we obtain a codebook for each number of training samples by normalizing the . We then plot the accuracy (proportion of cliques correctly stored by the network) in green and the normalized measure in purple against the number of training samples.

In the natural image patch example in 6, the data is a ternarized form of two by two pixel patches. Each pixel is normalized and then represented by two bits. If the normalized value of the pixel is greater than a threshold value $\alpha = 0.5$, it is represented as $(0,1)$, if it is less than $-\alpha = -0.5$, it is represented by $(1,0)$, otherwise, it is represented by $(0,0)$. Thus in a two by two patch, there are $3^4 = 81$ states. Again, the distortion matrix is the Hamming distance between two bit vectors. We obtain codebooks along the RD curve and plot points where critical transitions are found.

For the language example, the codebook is conditioned on a single variable, which is the topic of the dataset; agriculture in America. This means we only consider the distribution of words of one topic over time as the independent variable t. The measure is thus over a single state in the language concept space.

5. Discussion

In this section, we discuss implications of our criticality measure. Our experiments show that criticality in system behavior appears to be relatively common, as predicted by RD theory. In this paper, we have shown that optimization procedures can also lead to critical transitions in function behavior and that in natural processes such as language evolution, such critical changes can also occur. It could be that these processes find encodings that are close to rate-optimal, as there may be a significant discontinuous change in the rate of change of encoding cost with respect to encoding error $\frac{dR}{dD}$ at these points. This means that approaching this point from a higher distortion would suddenly incur a higher cost of encoding for the same distortion at this point, and approaching it from a lower distortion would result in a sudden drop in encoding fidelity for the same reduction in encoding cost.

This analysis however, only accounts for the critical points on the RD curve. The critical points of a process would depend on the details of the process itself, and we can only expect the critical points to line up with the ones on the RD curve if the process is efficient enough to follow the RD curve near its critical points. It is possible to detect these discontinuities in $\frac{dR}{dD}$, though measuring it would require far more samples.

Another point that can be made is that criticality does not necessarily manifest as points, but manifolds. In Figure 3, a 1-dimensional critical manifold is observed in u. This is akin to how the boiling point of a substance is a function of pressure and temperature. Further work could involve studying the behavior of these manifolds.

Observing the codebooks at the critical points in order of increasing distortion, we find that the codebook breaks away further from encoding the distribution as the identity function. Further work

 doi:10.20944/preprints202411.0663.v1

could be to observe the microstructure of codebooks at the critical points to try to understand which states are chosen to reduce the cost of coding perfectly at each critical point.

There are several direct implications for critical behavior, some of which were outlined by Sims in [84]. One open question is how to apply these ideas to deep learning [85] which minimize a typically squared error loss. An insight is to choose models and normative principles that have critical signatures [65] (working memory). Given that optimal continuous codings are discrete [80,86–88], it is not that surprising for criticality to arise from near-optimal solutions to normative principles applied to information processing. Furthermore, it has been observed that generalization beyond overfitting a training dataset, such as that seen in grokking by large language models, is related to the double descent phenomena [28], which has been shown to be dependent on multiple factors [89]. These relationships could be further explored using the method described in this paper, by for example, estimating the D_{KL} of more complex distributions using [90].

Further work would involve continuous distributions which require a different treatment to obtain the RD curves and different probability distribution distance metrics. Other algorithms for mapping discrete spaces could also be analyzed with the techniques developed in this work.

6. Conclusion

Our initial hypothesis was that an information processing system which compresses and reinterprets information into a useful form usually has critical transitions in the way the information is transformed as some control parameter of the system is tweaked. This was confirmed with the use of divergence rate – the measure we developed to track the change in the maps along the control parameter – and a noisy peak finder which helped to identify the critical points. We believe this may provide fertile ground for further research into critical phenomena in other maps such as the behavior of learning algorithms.

Acknowledgments: We thank Sarah Marzen and Lav Varshney for their input on this work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RD	Rate-Distortion
D_{KL}	Kullback-Leibler Divergence
MEF	Minimum Energy Flow

References

1. Landau, L. The Theory of Phase Transitions. *Nature* **1936**, *138*, 840–841. doi:10.1038/138840a0.
2. Anisimov, M.A. Letter to the Editor: Fifty Years of Breakthrough Discoveries in Fluid Criticality. *International Journal of Thermophysics* **2011**, *32*, 2001–2009. doi:10.1007/s10765-011-1073-0.
3. Peierls, R. On Ising's model of ferromagnetism. *Mathematical Proceedings of the Cambridge Philosophical Society* **1936**, *32*, 477–481. doi:10.1017/S0305004100019174.
4. Bak, P.; Chen, K. The physics of fractals. *Physica D: Nonlinear Phenomena* **1989**, *38*, 5–12.
5. Paczuski, M.; Maslov, S.; Bak, P. Avalanche dynamics in evolution, growth, and depinning models. *Physical Review E* **1996**, *53*, 414–443. doi:10.1103/PhysRevE.53.414.
6. Watkins, N.W.; Pruessner, G.; Chapman, S.C.; Crosby, N.B.; Jensen, H.J. 25 years of self-organized criticality: concepts and controversies. *Space Science Reviews* **2016**, *198*, 3–44.
7. Langton, C.G. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: Nonlinear Phenomena* **1990**, *42*, 12–37. doi:10.1016/0167-2789(90)90064-V.
8. Prokopenko, M. Modelling complex systems and guided self-organisation. *Journal & Proceedings of the Royal Society of New South Wales* **2017**, *150*, 104–109.
9. Mora, T.; Bialek, W. Are Biological Systems Poised at Criticality? *Journal of Statistical Physics* **2011**, *144*, 268–302. doi:10.1007/s10955-011-0229-4.

10. Muñoz, M.A. Colloquium : Criticality and dynamical scaling in living systems. *Reviews of Modern Physics* **2018**, *90*, 031001. doi:10.1103/RevModPhys.90.031001.
11. Bertschinger, N.; Natschläger, T. Real-Time Computation at the Edge of Chaos in Recurrent Neural Networks. *Neural Computation* **2004**, *16*, 1413–1436. doi:10.1162/089976604323057443.
12. Kinouchi, O.; Copelli, M. Optimal dynamical range of excitable networks at criticality. *Nature Physics* **2006**, *2*, 348–351. doi:10.1038/nphys289.
13. Legenstein, R.; Maass, W. Edge of chaos and prediction of computational performance for neural circuit models. *Neural Networks* **2007**, *20*, 323–334. doi:10.1016/j.neunet.2007.04.017.
14. Boedecker, J.; Obst, O.; Lizier, J.T.; Mayer, N.M.; Asada, M. Information processing in echo state networks at the edge of chaos. *Theory in Biosciences* **2012**, *131*, 205–213. doi:10.1007/s12064-011-0146-8.
15. Hoffmann, H.; Payton, D.W. Optimization by Self-Organized Criticality. *Scientific Reports* **2018**, *8*, 2358. doi:10.1038/s41598-018-20275-7.
16. Wilting, J.; Priesemann, V. Inferring collective dynamical states from widely unobserved systems. *Nature Communications* **2018**, *9*, 2325. doi:10.1038/s41467-018-04725-4.
17. Avramiea, A.E.; Masood, A.; Mansvelder, H.D.; Linkenkaer-Hansen, K. Long-Range Amplitude Coupling Is Optimized for Brain Networks That Function at Criticality. *The Journal of Neuroscience* **2022**, *42*, 2221–2233. doi:10.1523/JNEUROSCI.1095-21.2022.
18. Kelso, J.A. Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **1984**, *246*, R1000–R1004. https://doi.org/10.1152/ajpregu.1984.246.6.R1000.
19. Shew, W.L.; Plenz, D. The Functional Benefits of Criticality in the Cortex. *The Neuroscientist* **2013**, *19*, 88–100.
20. Tkačik, G.; Bialek, W. Information Processing in Living Systems. *Annual Review of Condensed Matter Physics* **2016**, *7*, 89–117. doi:10.1146/annurev-conmatphys-031214-014803.
21. Erten, E.; Lizier, J.; Piraveenan, M.; Prokopenko, M. Criticality and Information Dynamics in Epidemiological Models. *Entropy* **2017**, *19*, 194. doi:10.3390/e19050194.
22. Cocchi, L.; Gollo, L.L.; Zalesky, A.; Breakspear, M. Criticality in the brain: A synthesis of neurobiology, models and cognition. *Progress in Neurobiology* **2017**, *158*, 132–152. doi:10.1016/j.pneurobio.2017.07.002.
23. Zimmern, V. Why Brain Criticality Is Clinically Relevant: A Scoping Review. *Frontiers in Neural Circuits* **2020**, *14*, 54. doi:10.3389/fncir.2020.00054.
24. Cramer, B.; Stöckel, D.; Kreft, M.; Wibral, M.; Schemmel, J.; Meier, K.; Priesemann, V. Control of criticality and computation in spiking neuromorphic networks with plasticity. *Nature Communications* **2020**, *11*, 2853. doi:10.1038/s41467-020-16548-3.
25. Heiney, K.; Huse Ramstad, O.; Fiskum, V.; Christiansen, N.; Sandvig, A.; Nichele, S.; Sandvig, I. Criticality, Connectivity, and Neural Disorder: A Multifaceted Approach to Neural Computation. *Frontiers in Computational Neuroscience* **2021**, *15*, 611183. doi:10.3389/fncom.2021.611183.
26. O'Byrne, J.; Jerbi, K. How critical is brain criticality? *Trends in Neurosciences* **2022**, *45*, 820–837. https://doi.org/10.1016/j.tins.2022.08.007.
27. Power, A.; Burda, Y.; Edwards, H.; Babuschkin, I.; Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177* **2022**.
28. Liu, Z.; Kitouni, O.; Nolte, N.S.; Michaud, E.; Tegmark, M.; Williams, M. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems* **2022**, *35*, 34651–34663.
29. Shannon, C.E. Probability of Error for Optimal Codes in a Gaussian Channel. *Bell System Technical Journal* **1959**, *38*, 611–656. doi:10.1002/j.1538-7305.1959.tb03905.x.
30. Berger, T. Rate Distortion Theory and Data Compression. In *Advances in Source Coding*; Springer Vienna: Vienna, 1975; pp. 1–39. doi:10.1007/978-3-7091-2928-9_1.
31. Sterling, P.; Laughlin, S. *Principles of Neural Design*; The MIT Press, 2015. https://doi.org/10.7551/mitpress/9780262028707.001.0001.
32. Barlow, H.B. Possible Principles Underlying the Transformations of Sensory Messages. In *Sensory Communication*; Rosenblith, W.A., Ed.; The MIT Press, 1961; pp. 216–234. doi:10.7551/mitpress/9780262518420.003.0013.
33. Rose, K. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE* **1998**, *86*, 2210–2239.

34. Wibisono, A.; Jog, V.; Loh, P.L. Information and estimation in Fokker-Planck channels. 2017 IEEE International Symposium on Information Theory (ISIT). IEEE, 2017, pp. 2673–2677.
35. Gould, S.J.; Eldredge, N. Punctuated equilibria: an alternative to phyletic gradualism. In *Models in paleobiology*; Freeman, Cooper, 1972.
36. Wallace, R.; Wallace, D. Punctuated Equilibrium in Statistical Models of Generalized Coevolutionary Resilience: How Sudden Ecosystem Transitions Can Entrain Both Phenotype Expression and Darwinian Selection. In *Transactions on Computational Systems Biology IX*; Istrail, S.; Pevzner, P.; Waterman, M.S.; Priami, C., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; Vol. 5121, pp. 23–85. Series Title: Lecture Notes in Computer Science, doi:10.1007/978-3-540-88765-2_2.
37. King, D.M. Classification of Phase Transition Behavior in a Model of Evolutionary Dynamics. PhD thesis, University of Missouri-St. Louis, 2012.
38. Wallace, R. Adaptation, Punctuation and Information: A Rate-Distortion Approach to Non-Cognitive 'Learning Plateaus' in evolutionary processes. *Acta Biotheoretica* **2002**, *50*, 101–116. https://doi.org/10.1023/A:1016381028734.
39. Tlusty, T. A model for the emergence of the genetic code as a transition in a noisy information channel. *Journal of Theoretical Biology* **2007**, *249*, 331–342. doi:10.1016/j.jtbi.2007.07.029.
40. Tlusty, T. A rate-distortion scenario for the emergence and evolution of noisy molecular codes. *Physical Review Letters* **2008**, *100*, 048101. arXiv: 1007.4149, doi:10.1103/PhysRevLett.100.048101.
41. Liuling Gong.; Bouaynaya, N.; Schonfeld, D. Information-Theoretic Model of Evolution over Protein Communication Channel. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2011**, *8*, 143–151. doi:10.1109/TCBB.2009.1.
42. Marzen, S.E.; DeDeo, S. The evolution of lossy compression. *Journal of The Royal Society Interface* **2017**, *14*, 20170166. doi:10.1098/rsif.2017.0166.
43. Marzen, S.; DeDeo, S. Weak universality in sensory tradeoffs. *Physical Review E* **2016**, *94*, 060101. doi:10.1103/PhysRevE.94.060101.
44. van der Schaaf, A.; van Hateren, J. Modelling the Power Spectra of Natural Images: Statistics and Information. *Vision Research* **1996**, *36*, 2759–2770. doi:10.1016/0042-6989(96)00002-8.
45. Ruderman, D.L.; Bialek, W. Statistics of natural images: Scaling in the woods. *Physical Review Letters* **1994**, *73*, 814–817. doi:10.1103/PhysRevLett.73.814.
46. Zhaoping, L. *Understanding Vision: Theory, Models, and Data*, 1 ed.; Oxford University PressOxford, 2014.
47. Sugar, C.A.; James, G.M. Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. *Journal of the American Statistical Association* **2003**, *98*, 750–763. doi:10.1198/016214503000000666.
48. Hillar, C.J.; Tran, N.M. Robust Exponential Memory in Hopfield Networks. *The Journal of Mathematical Neuroscience* **2018**, *8*, 1. doi:10.1186/s13408-017-0056-2.
49. Hillar, C.; Chan, T.; Taubman, R.; Rolnick, D. Hidden Hypergraphs, Error-Correcting Codes, and Critical Learning in Hopfield Networks. *Entropy* **2021**, *23*, 1494. doi:10.3390/e23111494.
50. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* **1982**, *79*, 2554–2558. doi:10.1073/pnas.79.8.2554.
51. Humayun, A.I.; Balestriero, R.; Baraniuk, R. Deep networks always grok and here is why. *arXiv preprint arXiv:2402.15555* **2024**.
52. Bowern, C. Punctuated Equilibrium and Language Change. In *Encyclopedia of Language & Linguistics*; Elsevier, 2006; pp. 286–289. doi:10.1016/B0-08-044854-2/01870-8.
53. Tadic, B.; Dankulov, M.M.; Melnik, R. The mechanisms of self-organised criticality in social processes of knowledge creation. *Physical Review E* **2017**, *96*, 032307. arXiv:1705.10982 [physics], https://doi.org/10.1103/PhysRevE.96.032307.
54. Barron, A.T.J.; Huang, J.; Spang, R.L.; DeDeo, S. Individuals, institutions, and innovation in the debates of the French Revolution. *Proceedings of the National Academy of Sciences* **2018**, *115*, 4607–4612. Publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.1717729115.
55. Nevalainen, T.; Säily, T.; Vartiainen, T.; Liimatta, A.; Lijffijt, J. History of English as punctuated equilibria? A meta-analysis of the rate of linguistic change in Middle English. *Journal of Historical Sociolinguistics* **2020**, *6*. Publisher: De Gruyter Mouton, doi:10.1515/jhsl-2019-0008.

56. Gupta, R.; Roy, S.; Meel, K.S. Phase Transition Behavior in Knowledge Compilation. In *Principles and Practice of Constraint Programming*; Simonis, H., Ed.; Springer International Publishing: Cham, 2020; Vol. 12333, pp. 358–374. Series Title: Lecture Notes in Computer Science, doi:10.1007/978-3-030-58475-7_21.
57. Seoane, L.F.; Solé, R. Criticality in Pareto Optimal Grammars? *Entropy* **2020**, *22*, 165. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, doi:10.3390/e22020165.
58. Nesterov, Y. A method for solving the convex programming problem with convergence rate O(1/k^2). *Proceedings of the USSR Academy of Sciences* **1983**, *269*, 543–547.
59. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed ed.; Wiley-Interscience: Hoboken, N.J, 2006. OCLC: ocm59879802.
60. Davisson, L. Rate Distortion Theory: A Mathematical Basis for Data Compression. *IEEE Transactions on Communications* **1972**, *20*, 1202–1202. Conference Name: IEEE Transactions on Communications, doi:10.1109/TCOM.1972.1091311.
61. Fang, H.C.; Huang, C.T.; Chang, Y.W.; Wang, T.C.; Tseng, P.C.; Lian, C.J.; Chen, L.G. 81MS/s JPEG2000 single-chip encoder with rate-distortion optimization. 2004 IEEE International Solid-State Circuits Conference (IEEE Cat. No.04CH37519), 2004, pp. 328–531 Vol.1. ISSN: 0193-6530, doi:10.1109/ISSCC.2004.1332727.
62. Choi, I.; Lee, J.; Jeon, B. Fast Coding Mode Selection With Rate-Distortion Optimization for MPEG-4 Part-10 AVC/H.264. *IEEE Transactions on Circuits and Systems for Video Technology* **2006**, *16*, 1557–1561. Conference Name: IEEE Transactions on Circuits and Systems for Video Technology, doi:10.1109/TCSVT.2006.883506.
63. Zarcone, R.V.; Engel, J.H.; Burc Eryilmaz, S.; Wan, W.; Kim, S.; BrightSky, M.; Lam, C.; Lung, H.L.; Olshausen, B.A.; Philip Wong, H.S. Analog Coding in Emerging Memory Systems. *Scientific Reports* **2020**, *10*, 6831. doi:10.1038/s41598-020-63723-z.
64. Sims, C.R. Rate–distortion theory and human perception. *Cognition* **2016**, *152*, 181–198. https://doi.org/10.1016/j.cognition.2016.03.020.
65. Jakob, A.M.; Gershman, S.J. Rate-distortion theory of neural coding and its implications for working memory. preprint, Neuroscience, 2022. https://doi.org/10.1101/2022.02.28.482269.
66. Wallace, R. A Rate Distortion approach to protein symmetry. *Nature Precedings* **2010**, p. 14. https://doi.org/10.1016/j.biosystems.2010.05.002.
67. Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory* **1972**, *18*, 460–473. Conference Name: IEEE Transactions on Information Theory, doi:10.1109/TIT.1972.1054855.
68. Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory* **1972**, *18*, 14–20. Conference Name: IEEE Transactions on Information Theory, doi:10.1109/TIT.1972.1054753.
69. Cox, D.A.; Little, J.; O'shea, D. *Using algebraic geometry*; Vol. 185, Springer Science & Business Media, 2005.
70. Browder, F.E. On continuity of fixed points under deformations of continuous mappings. *Summa Brasiliensis Mathematicae* **1960**, *4*, 183–191.
71. Solan, E.; Solan, O.N. Browder's theorem through brouwer's fixed point theorem. *The American Mathematical Monthly* **2023**, *130*, 370–374.
72. Sanyal, R.; Sturmfels, B.; Vinzant, C. The entropic discriminant. *Advances in Mathematics* **2013**, *244*, 678–707.
73. Hillar, C.; Wibisono, A. Maximum entropy distributions on graphs, 2013. Publisher: arXiv Version Number: 3, doi:10.48550/ARXIV.1301.3321.
74. Wang, X. A Simple Proof of Descartes's Rule of Signs. *The American Mathematical Monthly* **2004**, *111*, 525. doi:10.2307/4145072.
75. Haukkanen, P.; Tossavainen, T. A generalization of Descartes' rule of signs and fundamental theorem of algebra. *Applied Mathematics and Computation* **2011**, *218*, 1203–1207. doi:10.1016/j.amc.2011.05.107.
76. Berger, T. Rate-Distortion Theory. In *Wiley Encyclopedia of Telecommunications*; Proakis, J.G., Ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2003; p. eot142. doi:10.1002/0471219282.eot142.
77. Hillar, C.; Marzen, S. Revisiting Perceptual Distortion for Natural Images: Mean Discrete Structural Similarity Index. 2017 Data Compression Conference (DCC); IEEE: Snowbird, UT, USA, 2017; pp. 241–249. doi:10.1109/DCC.2017.84.
78. Hillar, C.; Mehta, R.; Koepsell, K. A Hopfield recurrent neural network trained on natural images performs state-of-the-art image compression. 2014 IEEE International Conference on Image Processing (ICIP); IEEE: Paris, France, 2014; pp. 4092–4096. doi:10.1109/ICIP.2014.7025831.

79. Mehta, R.; Marzen, S.; Hillar, C. Exploring discrete approaches to lossy compression schemes for natural image patches. 2015 23rd European Signal Processing Conference (EUSIPCO); IEEE: Nice, 2015; pp. 2236–2240. doi:10.1109/EUSIPCO.2015.7362782.
80. Hillar, C.J.; Marzen, S.E. Neural network coding of natural images with applications to pure mathematics. In *Algebraic and Geometric Methods in Discrete Mathematics*; American Mathematical Society, 2017; pp. 189–221.
81. Del Papa, B.; Priesemann, V.; Triesch, J. Criticality meets learning: Criticality signatures in a self-organizing recurrent neural network. *PLOS ONE* **2017**, *12*, e0178683. doi:10.1371/journal.pone.0178683.
82. Mikolov, T. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* **2013**, *3781*.
83. Frolik, E. The History of Agriculture in the United States Beginning With the Seventeenth Century. *Transactions of the Nebraska Academy of Sciences and Affiliated Societies* **1977**.
84. Jung, J.; Kim, J.H.J.; Matějka, F.; Sims, C.A. Discrete Actions in Information-Constrained Decision Problems. *The Review of Economic Studies* **2019**, *86*, 2643–2667. doi:10.1093/restud/rdz011.
85. Grohs, P.; Klotz, A.; Voigtlaender, F. Phase Transitions in Rate Distortion Theory and Deep Learning. *Foundations of Computational Mathematics* **2021**. doi:10.1007/s10208-021-09546-4.
86. Smith, J.G. The information capacity of amplitude- and variance-constrained scalar gaussian channels. *Information and Control* **1971**, *18*, 203–219. doi:10.1016/S0019-9958(71)90346-9.
87. Fix, S.L. Rate distortion functions for squared error distortion measures. Annual Allerton Conference on Communication, Control and Computing, 1978, pp. 704–711.
88. Rose, K. A mapping approach to rate-distortion computation and analysis. *IEEE Transactions on Information Theory* **1994**, *40*, 1939–1952. doi:10.1109/18.340468.
89. Schaeffer, R.; Khona, M.; Robertson, Z.; Boopathy, A.; Pistunova, K.; Rocks, J.W.; Fiete, I.R.; Koyejo, O. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle. *arXiv preprint arXiv:2303.14151* **2023**.
90. Borade, S.; Zheng, L. Euclidean information theory. 2008 IEEE International Zurich Seminar on Communications. IEEE, 2008, pp. 14–17.