Université de Montréal

Towards a Geometric Theory of Information

par

Jose Gallego

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Rapport pour la partie orale
de l'examen pré-doctoral

Août, 2020

Contents

1 Introduction

Shannon's seminal theory of information (1948) has been of paramount importance in the development of modern machine learning techniques. However, standard information measures deal with probability distributions over an alphabet considered as a mere set of symbols and disregard additional geometric structure, which might be available in the form of a metric or similarity function. As a consequence of this, information theory concepts derived from the Shannon entropy (such as cross entropy and the Kullback-Leibler divergence) are usually blind to the geometric structure in the domains over which the distributions are defined.

The development of machine learning and information theory as scientific disciplines has been strongly intertwined. Compiling an exhaustive collection of research at the intersection between these two fields might be as ambitious as reviewing those machine learning paper that make use of differential calculus. Among several important landmarks, both old and new, we find the use of information gain (in the sense of the Kullback-Leibler divergence) to measure the importance of attributes in decision trees (Quinlan, 1986); approximate second order optimization based on natural gradient descent (Amari, 1998); the information bottleneck framework for learning representations (Tishby and Zaslavsky, 2015; Tishby et al., 2000; Saxe et al., 2018); a rate-distortion analysis of variational autoencoders (Alemi et al., 2018); regularization based on entropy to model exploration in reinforcement learning (Haarnoja et al., 2017), the estimation of mutual information between high dimensional continuous random variables via optimizing neural networks Belghazi et al. (2018); and the measurement of information about a learning task stored in the weights of a neural network after training (Achille et al., 2019).

The blindness of Shannon's concepts to existent geometric structure limits their applicability. For example, the Kullback-Leibler divergence cannot be optimized for empirical measures with non-matching supports. Optimal transport distances, such as Wasserstein, have emerged as practical alternatives with theoretical grounding. These methods have been used to compute barycenters (Cuturi and Doucet, 2014) and train generative models (Genevay et al., 2018). However, optimal transport is computationally expensive as it generally lacks closed-form solutions and requires the solution of linear programs or the execution of matrix scaling algorithms, even when solved only in approximate form (Cuturi, 2013). Approaches based on kernel methods (Gretton et al., 2012; Li et al., 2017; Salimans et al., 2018), which take a functional analytic view on the problem, have also been widely applied. However, further exploration on the interplay between kernel methods and information theory is lacking.

In spite of all the abundant connections between the fields, there is scarce work on creating a geometric approach to information theory that resolves some of the mentioned difficulties. The main contributions of this work are as follows: we *i)* introduce to the machine learning community a similarity-sensitive definition of entropy developed by Leinster and Cobbold (2012). Based on this notion of entropy we *ii)* propose geometry-aware counterparts for several information theory concepts. We *iii)* present a novel notion of divergence which incorporates the geometry of the space when comparing probability distributions, as in optimal transport. However, while the former methods require the solution of an optimization problem or a relaxation thereof via matrix-scaling algorithms, our proposal enjoys a closed-form expression and can be computed efficiently.

We present this collection of ideas as a first step towards a geometric theory of information.

2 Background

In this section we provide a brief presentation of the main theoretical notions used in our work. We start by summarizing Shannon's theory of information. Then, we use the concept of convexity to define Bregman divergences and how this relates to Shannon's mutual information. Finally, we discuss the optimal transport framework to compare probability distributions and the formulation of generative models as a divergence minimization problem.

Notation. Calligraphic letters denote $\mathcal{S}$ets, bold letters represent $\mathbf{M}$atrices and $\mathbf{v}$ectors, and double-barred letters denote $\mathbb{P}$robability distributions and information-theoretic functionals. To emphasize certain computational aspects, we alternatively denote a distribution $\mathbb{P}$ over a finite space $\mathcal{X}$ as a vector of probabilities $\mathbf{p}$. $\mathbf{I}$, $\mathbf{1}$ and $\mathbf{J}$ denote the identity matrix, a vector of ones and matrix of ones, with context-dependent dimensions. For vectors $\mathbf{v}$, $\mathbf{u}$ and $\alpha \in \mathbb{R}$, $\frac{\mathbf{v}}{\mathbf{u}}$ and $\mathbf{v}^{\alpha}$ denote element-wise division and exponentiation. $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner-product between two vectors or matrices. $\mathbf{\Delta}_n \triangleq \{\mathbf{x} \in \mathbb{R}^n | \, \langle \mathbf{1}, \mathbf{x} \rangle = 1 \text{ and } x_i \geq 0\}$ denotes the probability simplex over n elements. δ_x denotes a Dirac distribution at point x. We adopt the conventions $0 \cdot \log(0) = 0$ and $x \log(0) = -\infty$ for $x > 0$. For a continuous map $f : \mathcal{X} \rightarrow \mathcal{Y}$ and a measure $\mathbb{P}$ on $\mathcal{X}$, $f \# \mathbb{P}$, denotes the push-forward measure of $\mathbb{P}$ induced by f over $\mathcal{Y}$, with samples obtained by applying f on $x \sim \mathbb{P}$.

2.1 Information Theory

This section introduces key notions of information theory that are relevant for our work, such as (conditional) entropy and mutual information. We do not aim to provide a comprehensive overview of this subject, and direct the interested reader to the excellent books by Cover and Thomas (2005) and MacKay (2003).

Complementary definitions and proofs for the theorems in this section can be found in the mentioned references as well as the foundational work of Shannon (1948).

Consider a random variable X over a discrete alphabet of symbols $\mathcal{X} = \{x_1, \ldots, x_{|\mathcal{X}|}\}$ characterized by a probability mass function $p(x) \triangleq \mathbb{P}(X = x)$. We are interested in defining a notion of the "information" gained when we observe a realization x of the random variable X, denoted by $I(x)$. Such a notion can be uniquely derived from the following set of axioms:

(I1) No information is gained from a "sure" event x for which $p(x) = 1$.

(I2) The observation of unlikely events provides more information.

(I3) The total amount of information learned from two independent events is the sum of the information gained from each of the individual events.

Theorem 1. (**Information Content**) *Up to a multiplicative constant, there exists only one function satisfying axioms (I1)-(I3). The information content of an event x is given by:*

$$I_X(x) \triangleq -\log(p(x)) \tag{2.1}$$

This point-wise concept can be extended to the random variable X itself yielding one of the central concepts in information theory:

Definition 1. *(Shannon, 1948)* (**Shannon Entropy**) *The entropy of a random variable X is given by its expected information content.*

$$\mathbb{H}[X] \triangleq -\sum_{x \in \mathcal{X}} p(x) \log(p(x)) = -\mathbb{E}_{x \sim X}[\log(p(x))] = \mathbb{E}_{x \sim X}[I_X(x)] \tag{2.2}$$

Note that the definitions in Thm. 1 and Def. 1 only depend on the outcome x via its probability mass. Therefore, these quantities are invariant with respect to injective transformations on the alphabet $\mathcal{X}$. Formally, let $\mathcal{Y}$ be a discrete alphabet and $f : \mathcal{X} \to \mathcal{Y}$ be an injective function. Then, the random variable $Y = f(X)$ satisfies $\mathbb{H}[Y] = \mathbb{H}[X]$.

Example. *The entropy of a Bernoulli process with parameter p is given by the* binary entropy function $\mathbb{H}_b(p) = -p\log(p) - (1-p)\log(1-p)$. *This is illustrated in Fig. 2.1.*

The Shannon entropy naturally encodes the uncertainty on the realizations of a random variable. Clearly, whenever there is a sure event $p(x^*) = 1$, $\mathbb{H}[X] = 0$, and it attains maximum uncertainty whenever X is uniformly distributed. Compare this to the behavior for the Bernoulli process in Fig. 2.1.

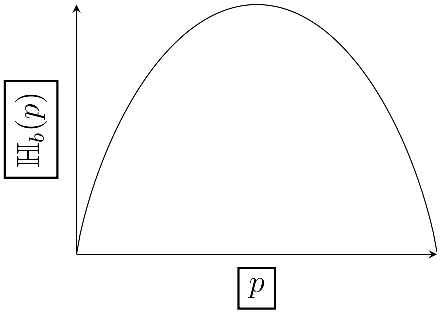

Figure 2.1 – Binary entropy function.

Remarkably, in a fashion analogous to the axiomatic characterization provided for the information content, the entropy of a random variable is the only function satisfying:

(E1) $\mathbb{H}$ should be continuous in the probabilities $p(x)$.

(E2) If X is uniformly distributed over an alphabet $\mathcal{X}$, then $\mathbb{H}$, should be a monotonic increasing function of $|\mathcal{X}|$.

(E3) If a choice is broken down into two successive choices, the original entropy should be the weighted sum of the individual entropies.

In fact, there exist many different characterizations of information measures. We refer the reader to the works of Csiszár (2008) and Aczél and Daróezy (1975) for a review, as well as Baez et al. (2011) for an elegant perspective in the language of category theory.

Def. 1 can accommodate for more general random variables, as well as standard operations between them such as conditioning.

Definition 2. (**Joint Entropy**) *The joint entropy of a pair of random variables X and Y taking values on the discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$ is the entropy of their joint distribution considered as a random variable over the alphabet $\mathcal{X} \times \mathcal{Y}$.*

$$\mathbb{H}[X, Y] \triangleq -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log(p(x, y)) \tag{2.3}$$

Definition 3. (**Conditional Entropy**) *Let (X, Y) have joint distribution $p(x, y)$. The conditional entropy of Y given X is defined as:*

$$\mathbb{H}[Y|X] \triangleq \sum_{x \in \mathcal{X}} p(x)\mathbb{H}[Y|X = x] = - \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log(p(y|x)) \tag{2.4}$$

The previous two definitions are tied together in what is often referred to as the "chain rule" of entropy. Thm. 2 relates the amount of information that is needed on average to describe the exact state of a system of two variables with the excess information unaccounted for after the observation of only one of the variables.

Theorem 2. (**Chain Rule**)

$$\mathbb{H}[X, Y] = \mathbb{H}[X] + \mathbb{H}[Y|X] = \mathbb{H}[Y] + \mathbb{H}[X|Y] \tag{2.5}$$

2.2 Convex Spaces

The entropy operator defined in the previous section acts naturally on the simplex comprising all possible categorical distributions over the discrete alphabet $\mathcal{X}$, which is a convex space. In this section we introduce some concepts of convex analysis. In particular, we highlight the fact that the Shannon entropy is a *concave* function on the simplex and the construction of Bregman divergences based on strictly convex functions. The books of Rockafellar (1970) and Boyd and Vandenberghe (2004) provide detailed information on convex analysis and optimization, respectively.

Definition 4. (**Convex Set**) *Let $\mathcal{V}$ be a vector space over some ordered field. A subset of $\mathcal{C}$ of $\mathcal{V}$ is called convex if for all u, $v \in \mathcal{C}$, and for all $\lambda \in [0, 1]$,*

$$(1 - \lambda)u + \lambda v \in \mathcal{C} \tag{2.6}$$

Example. *By construction, the simplex spanned by all possible convex combinations of elements of the standard basis of $\mathbb{R}^{|X|}$ is a convex set.*

Definition 5. (**Convex Function**) *Let $f : \mathcal{C} \subset \mathcal{V} \to \mathbb{R}$ be a function and define its epigraph by:*

$$\operatorname{epi}(f) \triangleq \{(v,t) \in \mathcal{C} \times \mathbb{R} \,|\, f(v) \leq t\} \tag{2.7}$$

We say that the function f is convex if $\operatorname{epi}(f)$ *is a convex set, with $\mathcal{C} \times \mathbb{R}$ equipped with the natural addition. A function f is called concave if $-f$ is convex.*

There exist many equivalent definitions of functional convexity. In particular, it is easy to see that Def. 5 is equivalent to requiring that for all u, $v \in \mathcal{C}$, and for all $\lambda \in [0,1]$,

$$f((1-\lambda)u + \lambda v) \leq (1-\lambda)f(u) + \lambda f(v). \tag{2.8}$$

The associated notions of strong convexity and strong concavity are obtained by requiring a strict inequality in Eq. (2.8) and suitably restricting the possible input and weighting parameter.

The convexity of once and twice differentiable functions can be characterized based solely on the local information provided by their derivatives.

Theorem 3. *Let f be a differentiable function and g twice differentiable defined on $\mathcal{C} \subset \mathbb{R}^d$. f is convex on $\mathcal{C}$ if and only if for all p, $q \in C$,*

$$f(p) - f(q) - \nabla_q f(q)^\top (p-q) \geq 0 \tag{2.9}$$

Moreover, g is convex on $\mathcal{C}$ if and only if the Hessian $\nabla_u^2 g(u)$ is positive semidefinite for all $u \in \mathcal{C}$.

Note that the first order condition in Eq. (2.9) stipulates that the *local* tangential approximation to f is a *global* lower bound. The calculus perspective on convexity allows us to provide a succinct proof of the concavity of the Shannon entropy.

Theorem 4. *The Shannon entropy $\mathbb{H}[\mathbf{p}]$ is a strictly concave function of the probability vector $\mathbf{p}$.*

Proof. The Hessian of the entropy with respect to $\mathbf{p}$ in the interior of the simplex is given by the negative definite matrix $\nabla_{\mathbf{p}}^2 \mathbb{H}[\mathbf{p}] = -\operatorname{diag}(\mathbf{p})^{-1}$. □

Another consequence of Eq. (2.9) is that it suggests a measurement of the separation between the points p and q based on how much the function f diverges from its tangential approximation at q.

Definition 6. *(Bregman, 1967)* (**Bregman Divergence**) *Let ψ be a real-valued, continuously-differentiable, strictly convex function on a closed convex set $\mathcal{C}$. The Bregman divergence from p to $q \in \mathcal{C}$ induced by the function ψ is given by:*

$$\mathfrak{D}_\psi[p \,||\, q] = \psi(p) - \psi(q) - \nabla_q \psi(q)^\top (p - q) \tag{2.10}$$

Bregman divergences are related to the notion of metric. However, the symmetry and triangle inequality conditions are not satisfied in general. The non-negativity and identifiability are consequences of the strict convexity of the inducing function ψ. Moreover, $\mathfrak{D}_\psi[p \,||\, q]$ is convex in p and linear in ψ.

Definition 7. *(Kullback and Leibler, 1951)* (**Kullback-Leibler Divergence**) *The Kullback-Leibler (KL) divergence between two distributions $\mathbb{P}$ and $\mathbb{Q}$ on an alphabet $\mathcal{X}$ is given by:*

$$\mathbb{KL}[\mathbb{P} \,||\, \mathbb{Q}] \triangleq \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) \tag{2.11}$$

Example. *Recall that the Shannon entropy is a strictly concave function. The KL divergence is the Bregman divergence induced by the* negative *Shannon entropy.*

Definition 8. (**Mutual Information**) *The mutual information between two random variables X and Y with joint distribution $p(x, y)$ is defined as the KL divergence between the joint distribution $p(x, y)$ and its independent factorization.*

$$\mathbb{I}[X; Y] \triangleq \mathbb{KL}[p(X, Y) \,||\, p(X) \otimes p(Y)] \tag{2.12}$$

The identifiability property of the Bregman divergences implies that the mutual information between X and Y is a measure of their statistical dependence. In particular, $I(X; Y)$ is zero precisely when X and Y are independent random variables. The mutual information has important connections to the definitions of entropy presented earlier:

Theorem 5.

$$\mathbb{I}[X; Y] = \mathbb{H}[X, Y] - \mathbb{H}[X|Y] - \mathbb{H}[Y|X] \tag{2.13}$$

Intuitively, the average reduction in the uncertainty about Y (X) that takes place by knowing the value of X (Y) is quantified by their mutual information.

An important theoretical result of this definition of mutual information is the data processing inequality. Informally, it states that "no clever manipulation of the data can improve the inferences that can be made from the data" (Cover and Thomas, 2005).

Theorem 6. (**Data Processing Inequality**) *Let $X \to Y \to Z$ form a Markov chain. In other words, Z is conditionally independent of Y given X.*

$$\mathbb{I}[X;Y] \geq \mathbb{I}[X;Z] \tag{2.14}$$

2.3 Generative Models

Consider the problem of approximating the distribution of a random variable X defined on a space $\mathcal{X}$ via a stochastic generation mechanism, $f : \mathcal{Z} \to \mathcal{X}$, that transforms samples from a base random variable Z into an *approximate* sample $\hat{X} = f(Z)$ of the distribution X. The most popular incarnations of this problem are related to families of explicit-probabilistic (Kingma and Welling, 2014) or implicit-adversarial (Goodfellow et al., 2014) generative models.

Although there is a vast literature regarding the implications of the adversarial formulation of this task in terms of the optimization dynamics, we center our attention to the general framework of generative modelling as a problem of divergence minimization. Formally, given a statistical distance of divergence $\mathfrak{D}$ that quantifies the separation between two distributions, we aim to find an adequate configuration of the parameters of a transformation f_θ such that $\mathfrak{D}[X, f_\theta \# Z]$ is small.

We have mentioned the Kullback-Leibler as an example of a statistical divergence. Arjovsky et al. (2017) provide a theoretical analysis of the topology induced by the KL divergence in the space of probability distributions, and use insights derived from such analysis to advocate for the use of distances based on optimal transport to train generative models. The theory of optimal transport, reviewed extensively in the work of Villani (2008), is based on the fundamental notion of a coupling between random variables:

Definition 9. (**Coupling**) *Let X and Y be two random variables on the spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ be the set of all joint distributions on $\mathcal{X} \times \mathcal{Y}$. An* *element $\pi \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ is called a coupling of X and Y if its marginals coincide with X and Y, respectively. The set of all couplings between X and Y is denoted by $\Pi(X, Y)$.*

Equivalently, $\pi \in \Pi(X, Y)$ if and only if:

$$\int_{\mathcal{X}\times\mathcal{Y}} \pi(x, y)dy = p(x) \quad \textit{and} \quad \int_{\mathcal{X}\times\mathcal{Y}} \pi(x, y)dx = p(y)$$

When $\mathcal{X}$ and $\mathcal{Y}$ are finite spaces of sizes n and m, the couplings correspond to matrices $\pi \in [0, 1]^{n\times m}$ such that $\pi\mathbf{1}_m = p(X)$ and $\mathbf{1}_n\pi = p(Y)$. Thus, $\Pi(X, Y)$ is a polytope in $\mathbb{R}^{n\times m}$.

The Wasserstein distance between two random variables X and Y corresponds to the solution of a (possibly infinitely dimensional) linear program on the space $\Pi(X, Y)$.

Definition 10. *(Wasserstein, 1969)* (**Wasserstein Distance**) *Let X and Y be two random variables on a metric space $(\mathcal{X}, d)$. For $p \geq 1$, the p-th Wasserstein distance between X and Y is defined as:*

$$\mathbb{W}_p(X, Y) = \inf_{\pi\in\Pi(X,Y)} \left(\mathbb{E}_{(x,y)\sim\pi}[d^p(x, y)]\right)^{1/p} = \inf_{\pi\in\Pi(X,Y)} \langle\pi, d^p\rangle^{1/p}, \tag{2.15}$$

where the $\langle\pi, d\rangle$ is the Frobenius product between the coupling π and a matrix representation of the distance d.

The definition of this statistical distance as a linear program presents a number of computational challenges. Arjovsky et al. (2017) provide a practical adversarial algorithm via the celebrated Kantorovich-Rubinstein duality (Kantorovich and Rubinstein, 1958). Cuturi (2013) presents an approach based on an entropic regularization of the linear objective in Eq. (2.15) and obtains a statistical distance which can be computed through Sinkhorn's matrix scaling iterations at a speed that is several orders of magnitude faster than that of transport solvers. Follow up works refine these ideas to the computation of barycenters between probability distributions (Cuturi and Doucet, 2014) and the training of generative models (Genevay et al., 2018).

3 Geometric Information Theory

Prologue

GAIT: A Geometric Approach to Information Theory. Jose Gallego, Ankit Vani, Max Schwarzer and Simon Lacoste-Julien. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), 2020.

Personal Contribution. JG proposed the central ideas of this work: import of a similarity sensitive notion of entropy to the machine learning community, definition of the Bregman-based divergence and other information theoretical concepts. JG was also in charge of the writing of the paper and proofs. AV and MS contributed on the design and execution of the experiments, as well as proofs for certain theorems. SLJ supervised this project and developed a proof for the conjecture in the 2-dimensional case.

Acknowledgments. This research was partially supported by the Canada CIFAR AI Chair Program and by a Google Focused Research award. Simon Lacoste-Julien is a CIFAR Associate Fellow in the Learning in Machines & Brains program. We thank Pablo Piantanida for the great tutorial on information theory which inspired this work.

Reproducibility. Our experimental results can be reproduced via: `https://github.com/jgalle29/gait`

3.1 A Geometry-Aware Approach

Suppose that we are given a finite space $\mathcal{X}$ with n elements along with a symmetric function that measures the similarity between elements, $\kappa : \mathcal{X} \times \mathcal{X} \to [0, 1]$. Let $\mathbf{K}$ be the Gram matrix induced by κ on $\mathcal{X}$; i.e, $\mathbf{K}_{x,y} \triangleq \kappa_{xy} \triangleq \kappa(x, y) = \kappa(y, x)$. $\mathbf{K}_{x,y} = 1$ indicates that the elements x and y are identical, while $\mathbf{K}_{x,y} = 0$ indicates full dissimilarity. We assume that $\kappa(x, x) = 1$ for all $x \in \mathcal{X}$. We call $(\mathcal{X}, \kappa)$ a (finite) similarity space. For brevity we denote $(\mathcal{X}, \kappa)$ by $\mathcal{X}$ whenever κ is clear from the context.

Of particular importance are the similarity spaces arising from metric spaces. Let $(\mathcal{X}, d)$ be a metric space and define $\kappa(x, y) \triangleq e^{-d(x,y)}$. Here, the symmetry and range conditions imposed on κ are trivially satisfied. The triangle inequality in $(\mathcal{X}, d)$ induces a multiplicative transitivity on $(\mathcal{X}, \kappa)$: for all $x, y, z \in \mathcal{X}$, $\kappa(x, y) \geq \kappa(x, z)\kappa(z, y)$. Moreover, for any (non-degenerate) metric space, the Gram matrix of its associated similarity space is positive definite (Reams, 1999, Lemma 2.5).

Note that there is an implicit choice of scale in the basis of the exponent in the definition of the similarity function. In fact, for each metric space we have a family of similarity spaces indexed by a scale parameter σ: define $\kappa_\sigma(x, y) \triangleq e^{-\frac{d(x,y)}{\sigma}}$. This is a central concept in the theory of the magnitude (a refined notion of size) of a metric space developed in (Leinster, 2013).

In this section, we present a theoretical framework which quantifies the "diversity" or "entropy" of a probability distribution defined on a similarity space, as well as a notion of divergence between such distributions.

3.1.1 Entropy and diversity

Let $\mathbb{P}$ be a probability distribution on $\mathcal{X}$. $\mathbb{P}$ induces a *similarity profile* $\mathbf{K}\mathbb{P} : \mathcal{X} \to [0, 1]$, given by $\mathbf{K}\mathbb{P}(x) \triangleq \mathbb{E}_{y\sim\mathbb{P}}[\kappa(x, y)] = (\mathbf{K}\mathbf{p})_x$.[1] $\mathbf{K}\mathbb{P}(x)$ represents the expected similarity between element x and a random element of the space sampled according to $\mathbb{P}$. Intuitively, it assesses how "satisfied" we would be by selecting x as a one-point summary of the space. In other words, it measures the ordinariness of x, and thus $\frac{1}{\mathbf{K}\mathbb{P}(x)}$ is the rarity or *distinctiveness* of x (Leinster and Cobbold, 2012). Note that the distinctiveness depends crucially on both the similarity structure of the space and the probability distribution at hand.

[1] This denotes the x-th entry of the result of the matrix-vector multiplication $\mathbf{K}\mathbf{p}$.

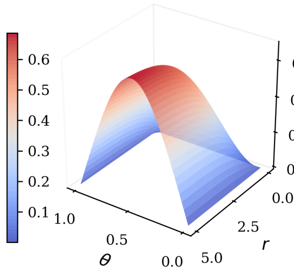

Figure 3.1 – $\mathbb{H}_1^{\mathbf{K}}$ interpolates towards the Shannon entropy as $r \to \infty$.

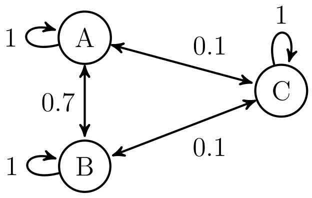

Figure 3.2 – A 3-point space with two highly similar elements.

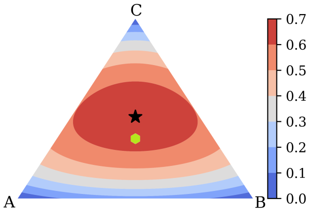

Figure 3.3 – $\mathbb{H}_1^{\mathbf{K}}$ for distributions over the space in Fig. 3.2.

Much like the interpretation of Shannon's entropy as the expected surprise of observing a random element of the space, we can define a notion of diversity as *expected distinctiveness*: $\sum_{x\in\mathcal{X}} \mathbb{P}(x)\frac{1}{\mathbf{K}\mathbb{P}(x)}$. This arithmetic weighted average is a particular instance of the family of power (or Hölder) means. Given $\mathbf{w} \in \mathbf{\Delta}_n$ and $\mathbf{x} \in \mathbb{R}^n_{\geq 0}$, the *weighted power mean of order* β is defined as $M_{\mathbf{w},\beta}(\mathbf{x}) \triangleq \left\langle \mathbf{w}, \mathbf{x}^\beta \right\rangle^{\frac{1}{\beta}}$. Motivated by this averaging scheme, Leinster and Cobbold (2012) proposed the following definition:

Definition 11. *(Leinster and Cobbold, 2012)* (**GAIT Entropy**) *The GAIT entropy of order $\alpha \geq 0$ of distribution $\mathbb{P}$ on finite similarity space $(\mathcal{X}, \kappa)$ is given by:*

$$\mathbb{H}_\alpha^{\mathbf{K}}[\mathbb{P}] \triangleq \log M_{\mathbf{p},1-\alpha}\left(\frac{1}{\mathbf{Kp}}\right) = \frac{1}{1-\alpha}\log\sum_{i=1}^{n}\mathbf{p}_i\frac{1}{(\mathbf{Kp})_i^{1-\alpha}}. \tag{3.1}$$

It is evident that whenever $\mathbf{K} = \mathbf{I}$, this definition reduces to the Rényi entropy (Rényi, 1961). Moreover, a continuous extension of Eq. (3.1) to $\alpha = 1$ via a L'Hôpital argument reveals a similarity-sensitive version of Shannon's entropy:

$$\mathbb{H}_1^{\mathbf{K}}[\mathbb{P}] = -\langle \mathbf{p}, \log(\mathbf{Kp})\rangle = -\mathbb{E}_{x\sim\mathbb{P}}[\log(\mathbf{K}\mathbb{P})_x]. \tag{3.2}$$

Let us dissect this definition via two simple examples. First, consider a distribution $\mathbf{p}_\theta = [\theta, 1-\theta]^\top$ over the points $\{x, y\}$ at distance $r \geq 0$, and define the similarity $\kappa_{xy} \triangleq e^{-r}$. As the points get further apart, the Gram matrix $\mathbf{K}_r$ transitions from $\mathbf{J}$ to $\mathbf{I}$. Fig. 3.1 displays the behavior of $\mathbb{H}_1^{\mathbf{K}_r}[\mathbf{p}_\theta]$. We observe that when r is large we recover the usual shape of Shannon entropy for a Bernoulli variable. In contrast, for low values of r, the curve approaches a constant zero function. In this case, we regard both elements of the space as identical: no matter how we distribute the probability among them, we have low uncertainty about the qualities of random samples. Moreover, the exponential of the maximum entropy, $\exp\left[\sup_\theta \mathbb{H}_1^{\mathbf{K}_r}[\mathbf{p}_\theta]\right] = 1 + \tanh(r) \in [1, 2]$, measures the *effective number of points* (Leinster and Meckes, 2016) at scale r.

Now, consider the space presented in Fig. 3.2, where the edge weights denote the similarity between elements. The maximum entropy distribution in this space following Shannon's view is the uniform distribution $\mathbf{u} = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^\top$. This is counterintuitive when we take into account the fact that points A and B are very similar. We argue that a reasonable expectation for a maximum entropy distribution is one which allocates roughly probability $\frac{1}{2}$ to point C and the remaining mass in equal proportions to points A and B. Fig. 3.3 displays the value of $\mathbb{H}_1^{\mathbf{K}}$ for all distributions on the 3-simplex. The green dot represents $\mathbf{u}$, while the black star corresponds to the maximum GAIT entropy with [A, B, C]-coordinates $\mathbf{p}^* \triangleq [0.273, 0.273, 0.454]^\top$. The induced similarity profile is $\mathbf{K}\mathbf{p}^* = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}]^\top$. Note how Shannon's probability-uniformity gets translated into a constant similarity profile.

Properties. We now list several important properties satisfied by the GAIT entropy, whose proofs and formal statements are contained in (Leinster and Cobbold, 2012) and (Leinster and Meckes, 2016):

- **Range**: $0 \leq \mathbb{H}_\alpha^{\mathbf{K}}[\mathbb{P}] \leq \log(|\mathcal{X}|)$.
- **K-monotonicity**: Increasing the similarity reduces the entropy. Formally, if $\kappa_{xy} \geq \kappa'_{xy}$ for all $x, y \in \mathcal{X}$, then $\mathbb{H}_\alpha^{\mathbf{J}}[\mathbb{P}] \leq \mathbb{H}_\alpha^{\mathbf{K}}[\mathbb{P}] \leq \mathbb{H}_\alpha^{\mathbf{K}'}[\mathbb{P}] \leq \mathbb{H}_\alpha^{\mathbf{I}}[\mathbb{P}]$.
- **Modularity**: If the space is partitioned into fully dissimilar groups, $(\mathcal{X}, \kappa) = \bigotimes_{c=1}^{C} (\mathcal{X}_c, \kappa_c)$, so that $\mathbf{K}$ is a block matrix ($x \in \mathcal{X}_c, y \in \mathcal{X}_{c'}, c \neq c' \Rightarrow \kappa_{xy} = 0$), then the entropy of a distribution on $\mathcal{X}$ is a weighted average of the block-wise entropies.
- **Symmetry**: Entropy is invariant to relabelings of the elements, provided that the rows of $\mathbf{K}$ are permuted accordingly.
- **Absence**: The entropy of a distribution $\mathbb{P}$ over $(\mathcal{X}, \kappa)$ remains unchanged when we restrict the similarity space to the support of $\mathbb{P}$.
- **Identical elements**: If two elements are identical (two equal rows in $\mathbf{K}$),

then combining them into one and adding their probabilities leaves the entropy unchanged.

- **Continuity**: $\mathbb{H}_{\alpha}^{\mathbf{K}}[\mathbb{P}]$ is continuous in $\alpha \in [0, \infty]$ for fixed $\mathbb{P}$, and continuous in $\mathbb{P}$ (w.r.t. standard topology on $\boldsymbol{\Delta}$) for fixed $\alpha \in (0, \infty)$.

- α-**Monotonicity**: $\mathbb{H}_{\alpha}^{\mathbf{K}}[\mathbb{P}]$ is non-increasing in α.

The role of α. Def. 11 establishes a family of entropies indexed by a non-negative parameter α, which determines the *relative importance of rare elements versus common ones*, where rarity is quantified by $\frac{1}{\mathbf{K}\mathbb{P}}$. In particular, $\mathbb{H}_{0}^{\mathbf{K}}[\mathbb{P}] = \log \left\langle \mathbf{p}, \frac{1}{\mathbf{Kp}} \right\rangle$. When $\mathbf{K} = \mathbf{I}$, $\mathbb{H}_{0}^{\mathbf{K}}[\mathbb{P}] = \log |\text{supp}(\mathbb{P})|$, which values rare and common species equally, while $\mathbb{H}_{\infty}^{\mathbf{K}}[\mathbb{P}] = -\log \max_{i \in \text{supp}(\mathbf{p})} (\mathbf{Kp})_i$ only considers the most common elements. Thus, in principle, the problem of finding a maximum entropy distribution depends on the choice of α.

Theorem 7. *(Leinster and Meckes, 2016) Let $(\mathcal{X}, \kappa)$ be a similarity space. There exists a probability distribution $\mathbb{P}_{\mathcal{X}}^{*}$ that maximizes $\mathbb{H}_{\alpha}^{\mathbf{K}}[\cdot]$ for all $\alpha \in \mathbb{R}_{\geq 0}$, simultaneously. Moreover, $\mathbb{H}_{\mathcal{X}}^{*} \triangleq \sup\limits_{\mathbb{P} \in \boldsymbol{\Delta}_{|\mathcal{X}|}} \mathbb{H}_{\alpha}^{\mathbf{K}}[\mathbb{P}]$ does not depend on α.*

Remarkably, Thm. 7 shows that the maximum entropy distribution is independent of α and thus, the maximum value of the GAIT entropy is an intrinsic property of the space: this quantity is a *geometric invariant*. In fact, if $\kappa(x, y) \triangleq e^{-d(x,y)}$ for a metric d on $\mathcal{X}$, there exist deep connections between $\mathbb{H}_{\mathcal{X}}^{*}$ and the magnitude of the metric space $(\mathcal{X}, d)$ (Leinster, 2013).

Theorem 8. *(Leinster and Meckes, 2016) Let $\mathbb{P}$ be a distribution on a similarity space $(\mathcal{X}, \kappa)$. $\mathbb{H}_{\alpha}^{\mathbf{K}}[\mathbb{P}]$ is independent of α if and only if $\mathbf{K}\mathbb{P}(x) = \mathbf{K}\mathbb{P}(y)$ for all $x, y \in supp(\mathbb{P})$.*

Recall the behavior of the similarity profile observed for $\mathbf{p}^*$ in Fig. 3.2. Thm. 8 indicates that this is not a coincidence: inducing a similarity profile which is constant over the support of a distribution $\mathbb{P}$ is a necessary condition for $\mathbb{P}$ being a maximum entropy distribution. In the setting $\alpha = 1$ and $\mathbf{K} = \mathbf{I}$, the condition $\mathbf{Kp} = \mathbf{p} = \lambda \mathbf{1}$ for some $\lambda \in \mathbb{R}_{\geq 0}$, is equivalent to the well known fact that the uniform distribution maximizes Shannon entropy.

3.1.2 Concavity of $\mathbb{H}_1^{\mathbf{K}}[\cdot]$

A common interpretation of the entropy of a probability distribution is that of the amount of *uncertainty* in the values/qualities of the associated random variable. From this point of view, the concavity of the entropy function is a rather intuitive and desirable property: "entropy should increase under averaging".

Consider the case $\mathbf{K} = \mathbf{I}$. $\mathbb{H}_\alpha^{\mathbf{I}}[\cdot]$ reduces to the the Rényi entropy of order α. For general values of α, this is not a concave function, but rather only Schur-concave (Ho and Verdú, 2015). However, $\mathbb{H}_1^{\mathbf{I}}[\cdot]$ coincides with the Shannon entropy, which is a strictly concave function. Since the subsequent theoretical developments make extensive use of the concavity of the entropy, we restrict our attention to the case $\alpha = 1$ for the rest of the paper.

To the best of our knowledge, whether the entropy $\mathbb{H}_1^{\mathbf{K}}[\mathbb{P}]$ is a (strictly) concave function of $\mathbb{P}$ for general similarity kernel $\mathbf{K}$ is currently an open problem. Although a proof of this result has remained elusive to us, we believe there are strong indicators, both empirical and theoretical, pointing towards a positive answer. We formalize these beliefs in the following conjecture:

Conjecture 1. *Let $(\mathcal{X}, \kappa)$ be a finite similarity space with Gram matrix $\mathbf{K}$. If $\mathbf{K}$ is positive definite and κ satisfies the multiplicative triangle inequality, then $\mathbb{H}_1^{\mathbf{K}}[\cdot]$ is strictly concave in the interior of $\boldsymbol{\Delta}_{|\mathcal{X}|}$.*

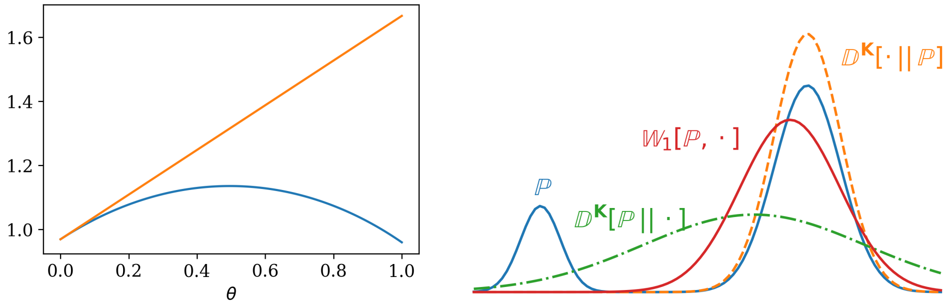

Figure 3.4 – **Left:** The entropy $\mathbb{H}_1^{\mathbf{K}}[(1-\theta)\mathbf{q} + \theta\mathbf{p}]$ is upper-bounded by the linear approximation at $\mathbf{q}$, given by $\mathbb{H}_1^{\mathbf{K}}[\mathbf{q}] + \theta \left\langle \nabla_{\mathbf{q}} \mathbb{H}_1^{\mathbf{K}}[\mathbf{q}], \mathbf{p} - \mathbf{q} \right\rangle$. **Right:** Optimal Gaussian model under various divergences on a simple mixture of Gaussians task under an RBF kernel. $\mathbb{W}_1$ denotes the 1-Wasserstein distance.

Fig. 3.4 shows the relationship between the linear approximation of the entropy and the value of the entropy over segment of the convex combinations between two measures. This behavior is consistent with our hypothesis on the concavity of $\mathbb{H}_1^{\mathbf{K}}[\cdot]$.

We emphasize the fact that the presence of the term $\log(\mathbf{Kp})$ complicates the analysis, as it incompatible with most linear algebra-based proof techniques, and it renders most information theory-based bounds too loose, as we explain in App A.3. Nevertheless, we provide extensive numerical experiments in App. A.3 which support our conjecture. In the remainder of this work, claims *dependent* on this conjecture are labelled ♣.

3.1.3 Comparing probability distributions

The previous conjecture implies that $-\mathbb{H}_1^{\mathbf{K}}[\cdot]$ is a strictly convex function. This naturally suggests considering the Bregman divergence induced by the negative GAIT entropy. This is analogous to the construction of the Kullback-Leibler divergence as the Bregman divergence induced by the negative Shannon entropy.

Straightfoward computation shows that the gap between the negative GAIT entropy at $\mathbf{p}$ and its linear approximation around $\mathbf{q}$ evaluated at $\mathbf{p}$ is:

$$-\mathbb{H}_1^{\mathbf{K}}[\mathbf{p}]-\left[-\mathbb{H}_1^{\mathbf{K}}[\mathbf{q}]+\left\langle-\nabla_{\mathbf{q}}\mathbb{H}_1^{\mathbf{K}}[\mathbf{q}],\,\mathbf{p}-\mathbf{q}\right\rangle\right]=1+\left\langle\mathbf{p},\log\frac{\mathbf{Kp}}{\mathbf{Kq}}\right\rangle-\left\langle\mathbf{q},\frac{\mathbf{Kp}}{\mathbf{Kq}}\right\rangle\overset{\text{(Conj. 1)}}{\geq}0.$$

Definition 12. (**GAIT Divergence**)♣ *The GAIT divergence between distributions $\mathbb{P}$ and $\mathbb{Q}$ on a finite similarity space $(\mathcal{X},\kappa)$ is given by:*

$$\mathbb{D}^{\mathbf{K}}[\mathbb{P}\,||\,\mathbb{Q}]\triangleq 1+\mathbb{E}_{\mathbb{P}}\left[\log\frac{\mathbf{K}\mathbb{P}}{\mathbf{K}\mathbb{Q}}\right]-\mathbb{E}_{\mathbb{Q}}\left[\frac{\mathbf{K}\mathbb{P}}{\mathbf{K}\mathbb{Q}}\right]. \tag{3.3}$$

When $\mathbf{K}=\mathbf{I}$, the GAIT divergence reduces to the Kullback-Leibler divergence. Compared to the family of f-divergences (Csiszár and Shields, 2004), this definition computes point-wise ratios between the similarity profiles $\mathbf{K}\mathbb{P}$ and $\mathbf{K}\mathbb{Q}$ rather than the probability masses (or more generally, Radon-Nikodym w.r.t. a reference measure). We highlight that $\mathbf{K}\mathbb{P}(x)$ provides a *global* view of the space via the Gram matrix from the perspective of $x\in\mathcal{X}$. Additionally, the GAIT divergence by definition inherits all the properties of Bregman divergences. In particular, $\mathbb{D}^{\mathbf{K}}[\mathbb{P}\,||\,\mathbb{Q}]$ is convex in $\mathbb{P}$.

Forward and backward GAIT divergence. Like the Kullback-Leibler divergence, the GAIT divergence is not symmetric and different orderings of the arguments induce different behaviors. Let $\mathcal{Q}$ be a family of distributions in which we would like to find an approximation $\mathbb{Q}$ to $\mathbb{P}\notin\mathcal{Q}$. $\arg\min_{\mathcal{Q}}\mathbb{D}^{\mathbf{K}}[\cdot\,||\,\mathbb{P}]$ concentrates around one of the modes of $\mathbb{P}$; this behavior is known as *mode seeking.* On the other hand, $\arg\min_{\mathbb{Q}} \mathbb{D}^{\mathbf{K}}[\mathbb{P} \,||\, \cdot]$ induces a *mass covering* behavior. Fig. 3.4 displays this phenomenon when finding the best (single) Gaussian approximation to a mixture of Gaussians.

Empirical distributions. Although we have developed our divergence in the setting of distributions over a finite similarity space, we can effectively compare two empirical distributions over a continuous space. Note that if an arbitrary $x \in \mathcal{X}$ (or more generally a measurable set E for a given choice of σ-algebra) has measure zero under both μ and ν, then such x (or E) is irrelevant in the computation of $\mathbb{D}^{\mathbf{K}}[\mathbb{P} \,||\, \mathbb{Q}]$. Therefore, when comparing empirical measures, the possibly continuous expectations involved in the extension of Eq. (12) to general measures reduce to finite sums over the corresponding supports.

Concretely, let $(\mathcal{X}, \kappa)$ be a (possibly continuous) similarity space and consider the empirical distributions $\hat{\mathbb{P}} = \sum_{i=1}^{n} \mathbf{p}_i \delta_{x_i}$ and $\hat{\mathbb{Q}} = \sum_{j=1}^{m} \mathbf{q}_j \delta_{y_i}$ with $\mathbf{p} \in \mathbf{\Delta}_n$ and $\mathbf{q} \in \mathbf{\Delta}_m$. The Gram matrix of the restriction of $(\mathcal{X}, \kappa)$ to $\mathcal{S} \triangleq \text{supp}(\mathbb{P}) \cup \text{supp}(\mathbb{Q})$ has the block structure $\mathbf{K}_{\mathcal{S}} \triangleq \begin{pmatrix} \mathbf{K}_{xx} & \mathbf{K}_{xy} \\ \mathbf{K}_{yx} & \mathbf{K}_{yy} \end{pmatrix}$, where $\mathbf{K}_{xx}$ is $n \times n$, $\mathbf{K}_{yy}$ is $m \times m$ and $\mathbf{K}_{xy} = \mathbf{K}_{yx}^{\top}$. It is easy to verify that

$$\mathbb{D}^{\mathbf{K}}[\hat{\mathbb{P}} \,||\, \hat{\mathbb{Q}}] = 1 + \left\langle \mathbf{p}, \log \frac{\mathbf{K}_{xx}\mathbf{p}}{\mathbf{K}_{xy}\mathbf{q}} \right\rangle - \left\langle \mathbf{q}, \frac{\mathbf{K}_{yx}\mathbf{p}}{\mathbf{K}_{yy}\mathbf{q}} \right\rangle. \tag{3.4}$$

Computational complexity. The computation of Eq. (3.4) requires $\mathcal{O}(|\kappa|(n+m)^2)$ operations, where $|\kappa|$ represents the cost of a kernel evaluation. This exhibits a quadratic behavior in the size of the union of the supports, typical of kernel-based approaches (Li et al., 2017). We highlight that Eqs. (12) and (3.4) provide a quantitative assessment of the dissimilarity between $\mathbb{P}$ and $\mathbb{Q}$ via a *closed form expression.* This is in sharp contrast to the multiple variants of optimal transport which require the solution of an optimization problem or the execution of several iterations of matrix scaling algorithms. Moreover, the proposals of Cuturi and Doucet (2014); Benamou et al. (2014) require at least $\Omega((|\kappa| + L)mn)$ operations, where L denotes the number of Sinkhorn iterations, which is an increasing function of the desired optimization tolerance. A quantitative comparison is presented in App. A.4.4.

Weak topology. The type of topology induced by a divergence on the space of probability measures plays important role in the context of training neural generative models. Several studies (Arjovsky et al., 2017; Genevay et al., 2018; Salimans et al., 2018) have exhibited how divergences which induce a weak topology constitute learning signals with useful gradients. In App. A.1, we provide an example in which the GAIT divergence can provide a smooth training signal despite being evaluated on distribution with disjoint supports.

Relation to Conj. 1 Thm. 9 displays the structure of the Hessian of the GAIT entropy. This is a straightfoward computation based on Def. 12.

Theorem 9. *The Hessian of the GAIT entropy with respect to* $\mathbf{p}$ *is given by:*

$$-\nabla^2_{\mathbf{p}}\mathbb{H}^{\mathbf{K}}_1[\mathbf{p}] = \mathbf{K}\,\mathrm{diag}\left(\frac{1}{\mathbf{Kp}}\right) - \mathbf{K}\,\mathrm{diag}\left(\frac{\mathbf{p}}{[\mathbf{Kp}]^2}\right)\mathbf{K} + \mathrm{diag}\left(\frac{1}{\mathbf{Kp}}\right)\mathbf{K} \tag{3.5}$$

Moreover, $-\nabla^2_{\mathbb{P}}[\mathbb{H}^{\mathbf{K}}_1[\mathbb{P}]]$ *is positive definite in the* 2*-dimensional case.*

A characterization of the definiteness of this Hessian is a promising direction towards the verification of Conj. 1. Furthermore, since we are interested in the behavior of the GAIT entropy operating on probability distributions, it is even sufficient to only consider the action of this matrix as a quadratic form the set of mass-preserving vectors with entries adding up to zero.

Careful analysis of this Hessian suggests we must strengthen the constraints on the kernel for Conj. 1 to hold for $|\mathcal{X}| > 2$. We found an instance of a 3x3 kernel satisfying all the conditions in the conjecture and distribution $\mathbf{p}$ for which the Hessian has a negative eigenvalue. It turns out that this specific $\mathbf{K}$ violates a form of triangle inequality. The condition, $\mathbf{K}_{ij} \geq \mathbf{K}_{il}\mathbf{K}_{lj}$ for all l, translates to the triangle inequality for an exponential kernel $e^{-d(x,y)}$. Furthermore, the "triangle inequality" for $\mathbf{K}$ implies $(\mathbf{Kp})_i >= \mathbf{K}_{ij}(\mathbf{Kp})_j$. Intuitively, this means that element i must be at least as popular (with respect to $\mathbf{p}$) as element j, times how close i and j are to each other, $\mathbf{K}_{ij}$.

3.1.4 Mutual Information

We now use the GAIT entropy to define similarity-sensitive generalization of standard concepts related to mutual information. As before, we restrict our attention to $\alpha = 1$. This is required to get the chain rule of conditional probability for the Rényi entropy and to use Conj. 1. Finally, we note that although one could use the GAIT divergence to define a mutual information, in a fashion analogous to how traditional mutual information is defined via the KL divergence, the resulting object is challenging to study theoretically. Instead, we use a definition based on entropy, which is equivalent in spaces without similarity structure.

Definition 13. *Let X, Y, Z be random variables taking values on the similarity spaces $(\mathcal{X}, \kappa)$, $(\mathcal{Y}, \lambda)$, $(\mathcal{Z}, \theta)$ with corresponding Gram matrices $\mathbf{K}$, $\mathbf{\Lambda}$, $\mathbf{\Theta}$. Let $[\kappa \otimes \lambda]((x,y),(x',y')) \triangleq \kappa(x,x')\lambda(y,y')$, and $(\mathbf{K}\mathbb{Q})_x \triangleq \mathbb{E}_{x'\sim\mathbb{Q}}[\kappa(x,x')]$ denotes the expected similarity between object x and a random $\mathbb{Q}$-distributed object. Let $\mathbb{P}$ be the joint distribution of X and Y. Then the joint entropy, conditional entropy, mutual information and conditional mutual information are defined following the formulas in Table. 3.1.*

Table 3.1 – Definitions of GAIT mutual information and joint entropy.

Joint Entropy	$\mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda}}[X,Y] \triangleq -\mathbb{E}_{x,y\sim\mathbb{P}}[\log([\mathbf{K}\otimes\mathbf{\Lambda}]\mathbb{P})_{x,y}]$
Conditional Entropy	$\mathbb{H}^{\mathbf{K},\mathbf{\Lambda}}[X\vert Y] \triangleq \mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda}}[X,Y] - \mathbb{H}^{\mathbf{\Lambda}}[Y]$
Mutual Information	$\mathbb{I}^{\mathbf{K},\mathbf{\Lambda}}[X;Y] \triangleq \mathbb{H}^{\mathbf{K}}[X] + \mathbb{H}^{\mathbf{\Lambda}}[Y] - \mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda}}[X,Y]$
Conditional M.I.	$\mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y\vert Z] \triangleq \mathbb{H}^{\mathbf{K},\mathbf{\Theta}}[X\vert Z] + \mathbb{H}^{\mathbf{\Lambda},\mathbf{\Theta}}[Y\vert Z] - \mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda},\mathbf{\Theta}}[X,Y\vert Z]$

Note that the GAIT joint entropy is simply the entropy of the joint distribution with respect to the tensor product kernel. This immediately implies monotonicity in the kernels $\mathbf{K}$ and $\mathbf{\Lambda}$. Note also that the chain rule of conditional probability holds by definition.

Subject to these definitions, similarity-sensitive versions of a number theorems analogous to standard results of information theory follow:

Theorem 10. *Let X, Y be independent, then $\mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda}}[X,Y] = \mathbb{H}^{\mathbf{K}}[X] + \mathbb{H}^{\mathbf{\Lambda}}[Y]$.*

When the conditioning variables are perfectly identifiable ($\mathbf{\Lambda} = \mathbf{I}$), we recover a simple expression for the conditional entropy:

Theorem 11. *For any kernel κ, $\mathbb{H}^{\mathbf{K},\mathbf{I}}[X|Y] = \mathbb{E}_{y\sim\mathbb{P}_y}[\mathbb{H}^{\mathbf{K}}[X|Y=y]]$.*

Using Conj. 1, we are also able to prove that conditioning on additional information cannot increase entropy, as intuitively expected.

Theorem 12. ♣ *For any similarity kernel κ, $\mathbb{H}^{\mathbf{K},\mathbf{I}}[X|Y] \leq \mathbb{H}^{\mathbf{K}}[X]$.*

Theorem 12 is equivalent to Conj. 1 when considering a categorical Y mixing over distributions $\{X_y\}_{y \in \mathcal{Y}}$.

Finally, a form of the data processing inequality (DPI), a fundamental result in information theory governing the mutual information of variables in a Markov chain structure, follows from Conj. 1.

Theorem 13. (**Data Processing Inequality**)$^{\clubsuit}$.
If $X \to Y \to Z$ is a Markov chain, then $\mathbb{I}^{\mathbf{K},\mathbf{\Theta}}[X;Z] \leq \mathbb{I}^{\mathbf{K},\mathbf{\Lambda}}[X;Y] + \mathbb{I}^{\mathbf{K},\mathbf{\Theta},\mathbf{\Lambda}}[X;Z|Y]$.

Note the presence of the additional term $\mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Z|Y]$ relative to the non-similarity-sensitive DPI given by $\mathbb{I}[X;Z] \leq \mathbb{I}[X;Y]$. Intuitively, this can be understood as reflecting that conditioning on Y does not convey all of its usual "benefit", as some information is lost due to the imperfect identifiability of elements in Y. When $\mathbf{\Lambda} = \mathbf{I}$ this term is 0, and the original DPI is recovered.

3.2 Related work

Theories of Information. Information theory is ubiquitous in modern machine learning: from variable selection via information gain in decision trees (Ben-David and Shalev-Shwartz, 2014), to using entropy as a regularizer in reinforcement learning (Fox et al., 2016), to rate-distortion theory for training generative models (Alemi et al., 2018). To the best of our knowledge, the work of Leinster and Cobbold (2012); Leinster and Meckes (2016) is the first formal treatment of information-theoretic concepts in spaces with non-trivial geometry, albeit in the context of ecology.

Comparing distributions. The ability to compare probability distributions is at the core of statistics and machine learning. Although traditionally dominated by maximum likelihood estimation, a significant portion of research on parameter estimation has shifted towards methods based on optimal transport, such as the Wasserstein distance (Villani, 2008). Two main reasons for this transition are (i) the need to deal with degenerate distributions (which might have density only over a low dimensional manifold) as is the case in the training of generative models (Goodfellow et al., 2014; Arjovsky et al., 2017; Salimans et al., 2018); and (ii) the development of alternative formulations and relaxations of the original optimal transport objective which make it feasible to approximately compute in practice (Cuturi and Doucet, 2014; Genevay et al., 2018).

Relation to kernel theory. The theory we have presented in this paper revolves around a notion of similarity on $\mathcal{X}$. The operator $\mathbf{K}\mathbb{P}$ corresponds to the embedding of the space of distributions on $\mathcal{X}$ into a reproducing kernel Hilbert space used for comparing distributions without the need for density estimation (Smola et al., 2007). In particular, a key concept in this work is that of a characteristic kernel, i.e., a kernel for which the embedding is injective. Note that this condition is equivalent to the positive definiteness of the Gram matrix $\mathbf{K}$ imposed above. Under these circumstances, the metric structure present in the Hilbert space can be imported to define the Maximum Mean Discrepancy distance between distributions (Gretton et al., 2012). Our definition of divergence also makes use of the object $\mathbf{K}\mathbb{P}$, but has motivations rooted in information theory rather than functional analysis. We believe that the framework proposed in this paper has the potential to foster connections between both fields.

3.3 Experiments

3.3.1 Comparison to Optimal Transport

Image barycenters. Given a collection of measures $\mathcal{P} = \{\mathbb{P}_i\}_{i=1}^n$ on a similarity space, we define the barycenter of $\mathcal{P}$ with respect to the GAIT divergence as $\arg\min_{\mathbb{Q}} \frac{1}{n} \sum_{i=1}^n \mathbb{D}^{\mathbf{K}}[\mathbb{P}_i \,||\, \mathbb{Q}]$. This is inspired by the work of Cuturi and Doucet (2014) on Wasserstein barycenters. Let the space $\mathcal{X} = [1:28]^2$ denote the pixel grid of an image of size 28×28. We consider each image in the MNIST dataset as an empirical measure over this grid in which the probability of location (x, y) is proportional to the intensity at the corresponding pixel. In other words, image i is considered as a measure $\mathbb{P}_i \in \mathbf{\Delta}_{|\mathcal{X}|}$. Note that in this case the kernel is a function of the distance between two pixels in the grid (two elements of $\mathcal{X}$), rather than the distance between two different images. We use a Gaussian kernel, and compute $\mathbf{K}\mathbb{P}_i$ by convolving the image $\mathbb{P}_i$ with an adequate filter, as proposed by Solomon et al. (2015).

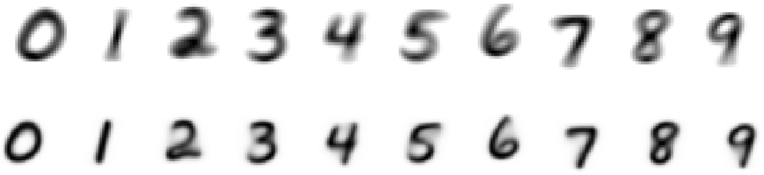

Figure 3.5 – Barycenters for each class of MNIST with our divergence (top) and the method of Cuturi and Doucet (2014) (bottom).

Fig. 3.5 shows the result of gradient-based optimization to find barycenters for each of the classes in MNIST (LeCun et al., 1998) along with the corresponding results using the method of Cuturi and Doucet (2014). We note that our method achieves results of comparable quality. Remarkably, the time for computing the barycenter for each class on a single CPU is reduced from 90 seconds using the efficient method proposed by Cuturi and Doucet (2014); Benamou et al. (2014) (implemented using a convolutional kernel (Solomon et al., 2015)) to less than 5 seconds using our divergence. Further experiments can be found in App. A.4.1.

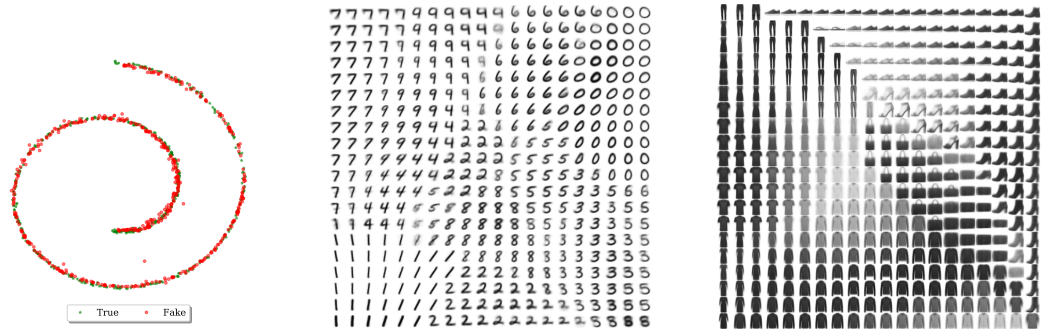

Figure 3.6 – **Left:** Generated Swiss roll data. **Center and Right:** Manifolds for MNIST and Fashion MNIST.

Generative models. The GAIT divergence can also be used as an objective for training generative models. We illustrate the results of using our divergence with a RBF kernel to learn generative models in Fig. 3.6 on a toy Swiss roll dataset, in addition to the MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017) datasets. For all three datasets, we consider a 2D latent space and replicate the experimental setup used by Genevay et al. (2018) for MNIST. We were able to use the same 2-layer multilayer perceptron architecture and optimization hyperparameters for all three datasets, requiring only the tuning of the kernel variance for Swiss roll data's scale.

Moreover, we do not need large batch sizes to get good quality generations from our models. The quality of our samples obtained using batch sizes as small as 50 are comparable to the ones requiring batch size of 200 by Genevay et al. (2018). We include additional experimental details and results in App. A.4.3, along with comparisons to variational auto-encoders (Kingma and Welling, 2014).

3.3.2 Approximating measures

Our method allows us to find a finitely-supported approximation $\mathbb{Q} = \sum_{j=1}^{m} \mathbf{q}_j \delta_{y_i}$ to a (discrete or continuous) target distribution $\mathbb{P}$. This is achieved by minimizing the divergence $\mathbb{D}^{\mathbf{K}}[\mathbb{P}||\mathbb{Q}]$ between them with respect to the locations $\{y_i\}_{i=1}^{m}$ and/or the masses of the atoms $\mathbf{q} \in \boldsymbol{\Delta}_m$ in the approximating measure. In this section, we consider situations where $\mathbb{P}$ is not a subset of the support of $\mathbb{Q}$. As a result, the Kullback-Leibler divergence (the case $\mathbf{K} = \mathbf{I}$) would be infinite and could not be minimized via gradent-based methods. However, the GAIT divergence can be minimized even in the case of non-overlapping supports since it takes into account similarities between items.

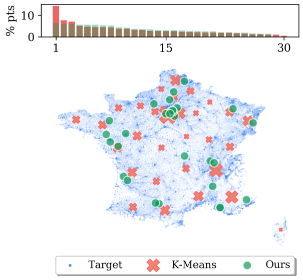

Figure 3.7 – Approximating a discrete measure with a uniform empirical measure.

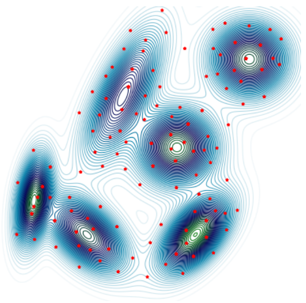

Figure 3.8 – Approximating a continuous density with a finitely-supported measure.

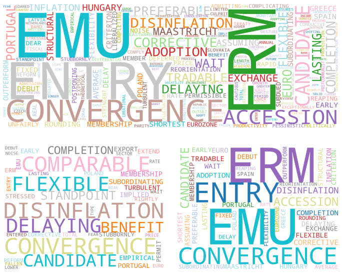

Figure 3.9 – **Top:** Original word cloud. **Left:** Sparse approximation with support size 43. **Right:** Top 43 original TF-IDF words.

In Fig. 3.7, we show the results of such an approximation on data for the population of France in 2010 consisting of 36,318 datapoints (Charpentier, 2012), similar to the setting of Cuturi and Doucet (2014). The weight of each atom in the blue measure is proportional to the population it represents. We use an RBF kernel and an approximating measure consisting of 50 points with uniform weights, and use gradient-based optimization to minimize $\mathbb{D}^{\mathbf{K}}$ with respect to the location of the atoms of the approximating measure. We compare with K-means (Pedregosa et al.,

2011) using identical initialization. Note that when using K-means, the resulting allocation of mass from points in the target measure to the nearest centroid can result in a highly unbalanced distribution, shown in the bar plot in orange. In contrast, our objective allows a uniformity constraint on the weight of the centroids, inducing a more homogeneous allocation. This is important in applications where an imbalanced allocation is undesirable, such as the placement of hospitals or schools.

Fig. 3.8 shows the approximation of the density of a mixture of Gaussians $\mathbb{P}$ by a uniform distribution $\mathbb{Q} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$ over $N = 200$ atoms with a polynomial kernel of degree 1.5, similar to the approximate super-samples (Chen et al., 2010) task presented by Claici et al. (2018) using the Wasserstein distance. We minimize $\mathbb{D}^{\mathbf{K}}[\mathbb{P} \,||\, \mathbb{Q}]$ with respect to the locations $\{x_i\}_{i=1}^{n}$. We estimate the continuous expectations with respect to $\mathbb{P}$ by repeatedly sampling minibatches to construct an empirical measure $\hat{\mathbb{P}}$. Note how the solution is a “uniformly spaced” allocation of the atoms through the space, with the number of points in a given region being proportional to mass of the region. See App. A.4.1 for a comparison to Claici et al. (2018).

Finally, one can approximate a measure when the locations of the atoms are fixed. As an example, we take an article from the News Commentary Parallel Corpus (Tiedemann, 2012), using as a measure $\mathbb{P}$ the normalized TF-IDF weights of each non-stopword in the article. Here, $\mathbf{K}$ is given by an RBF kernel applied to the 300-dimensional GLoVe (Pennington et al., 2014) embeddings of each word. We optimize $\mathbb{Q}$ applying a penalty to encourage sparsity. We show the result of this summarization in word-cloud format in Fig. 3.9. Note that compared to TF-IDF, which places most mass on a few unusual words, our method produces a summary that is more representative of the original text. This behavior can be modified by varying the bandwidth σ of the kernel, producing approximately the same result as TF-IDF when σ is very small; details are presented in App. A.4.1.

3.3.3 Measuring diversity and counting modes

As mentioned earlier, the exponential of the entropy $\exp(\mathbb{H}_1^{\mathbf{K}}[\mathbb{P}])$ provides a measure of the effective number of points in the space (Leinster, 2013). In Fig. 3.10, we use an empirical distribution to estimate the number of modes of a mixture of C Gaussians. As the kernel bandwidth σ increases, $\exp(\mathbb{H}_1^{\mathbf{K}}[\hat{\mathbb{P}}])$ decreases, with a marked plateau around C. We highlight that the lack of direct consideration

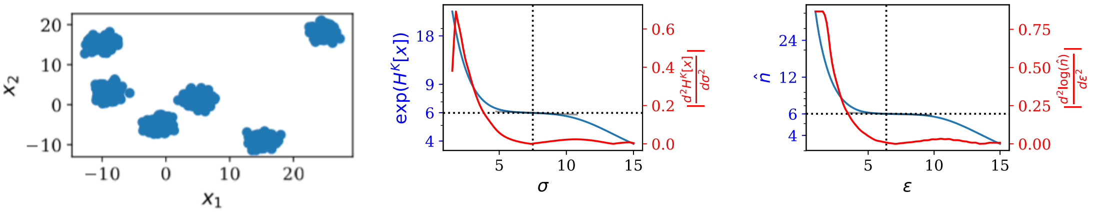

Figure 3.10 – **Left:** 1,000 samples from a mixture of 6 Gaussians. **Center:** Modes detected by varying σ in our method. **Right:** Modes detected by varying collision threshold ϵ in the birthday paradox-based method.

of geometry of the space in the Shannon entropy renders it useless here: at any (non-trivial) scale, $\exp(\mathbb{H}[\hat{\mathbb{P}}])$ equals the number of samples, and not the number of classes. Our approach obtains similar results as (a form of) the birthday paradox-based method of Arora et al. (2018), while avoiding the need for human evaluation of possible duplicates. Details and tests on MNIST can be found in App. A.4.2.

4 Future Work

Four central questions arise from the presented discussion:

- *Is the GAIT entropy a concave function?* The theoretical verification of this property plays a crucial role in the proposed definition for the divergence.
- *What are the desired axioms for geometry-aware information concepts?* Although the definition of Leinster and Cobbold (2012) enjoys certain desirable properties, it originates from heuristic rather than axiomatic motivations. Morevover, at the moment, there is no definite description of the properties that we would like such a definition to satisfy.
- *Is there an operational representation of the proposed entropy?* The Shannon entropy can be seen as the theoretical compression limit achievable for a specific distribution. Are the any analogous results for the proposed geometric entropy?
- *Which central information theoretic results can be recovered in this framework?* We presented a version of the data processing inequality which accounts for ”lost information” due to imperfect identifiability. Results such as Fano’s inequality and theorems in the area of lossy compression are natural candidates for the presented notions.

Parallel to this, the evident connections of our proposed approach with the kernel methods literature are worthy of further exploration. Our definitions make use of the *probability embedding* $\mathbf{K}\mathbb{P}$, but their motivations are rooted in information theory rather than functional analysis. We believe that our framework proposed has the potential to foster connections between both fields.

We hope the presented methods can prove fruitful in extending frameworks such as similarity-sensitive cross entropy objectives in the spirit of loss-calibrated decision theory (Lacoste-Julien et al., 2011), or the use of entropic regularization of policies in reinforcement learning (Fox et al., 2016), as well as information bottleneck for representation learning (Tishby and Zaslavsky, 2015). In particular, Tschannen et al. (2020) call for alternative measures of information in the context of representation learning. They state that “a new notion of information should account for both the amount of information stored in a representation and the geometry of the induced space necessary for good performance on downstream tasks”. Moreover, the work of Xu et al. (2020) indirectly re-evaluates the geometry insensitivity of Shannon’s information by accounting for the modeling power and computational constraints of the observer. We believe our proposed research is well suited to answering some of these questions and providing the machine learning community with useful and much needed theoretical tools.

Finally, we present a tentative timeline regarding the organization of future research and other PhD landmarks:

- Summer 2020: Completion of internship at Qualcomm Research.
- Fall 2020: Exploration of applicability of geometry based approaches for representation learning. Supervision of research executed by undergraduate student at EAFIT University.
- Winter - Summer 2021: Desiderata and axiomatic characterization of geometric information theory concepts.
- Fall 2021: Application of developed notions to data compression or entropy-regularized reinforcement learning.
- Winter 2021 - Summer 2022: Further research and thesis writing period.
- Summer 2022: Estimated graduation date.

5 Conclusion

In this work, we advocated the use of geometry-aware information theory concepts in machine learning. We presented the similarity-sensitive entropy of Leinster and Cobbold (2012) along with several important properties that connect it to fundamental notions in geometry. We then proposed a divergence induced by this entropy, which compares probability distributions by taking into account the similarities among the objects on which they are defined. Our proposal shares the empirical performance properties of distances based on optimal transport theory, such as the Wasserstein distance (Villani, 2008), but enjoys a closed-form expression. This obviates the need to solve a linear program or use matrix scaling algorithms (Cuturi, 2013), reducing computation significantly. Finally, we also proposed a similarity-sensitive version of mutual information based on the GAIT entropy.

The pervasiveness of information theoretic ideas and the potential benefits of a geometric perspective presented in our work portray the proposed research direction as a promising an impactful agenda.

A Appendix

A.1 Revisiting parallel lines

Let $Z \sim \mathcal{U}([0,1])$, $\phi \in \mathbb{R}$, and let $\mathbb{P}_\phi$ be the distribution of $(\phi, Z) \in \mathbb{R}^2$. This is a uniform distribution on the segment $\{\phi\} \times [0,1] \subset \mathbb{R}^2$, illustrated in Fig. A.1.

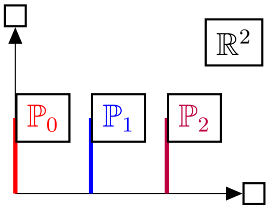

Figure A.1 – Distribution $\mathbb{P}_\phi$ with support on the 1-dim segment $\{\phi\} \times [0,1]$ for different values of ϕ.

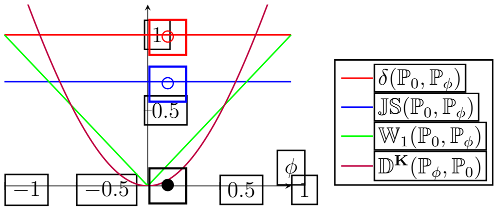

Figure A.2 – Values of the divergences as functions of ϕ. KL divergence values are ∞ except at $\phi = 0$.

Our goal is to find the *right* value of ϕ for a model distribution $\mathbb{P}_\phi$ using the dissimilarity with $\mathbb{P}_0$ as a learning signal. The behavior of common divergences on this type of problem was presented by Arjovsky et al. (2017) as a motivating example for the introduction of OT distances in the context of GANs.

$$\delta(\mathbb{P}_0, \mathbb{P}_\phi) = \begin{cases} 0 & \text{if } \phi = 0 \\ 1 & \text{else} \end{cases} \qquad \mathbb{KL}(\mathbb{P}_0, \mathbb{P}_\phi) = \mathbb{KL}(\mathbb{P}_\phi, \mathbb{P}_0) = \begin{cases} 0 & \text{if } \phi = 0 \\ \infty & \text{else} \end{cases}$$

$$\mathbb{W}_1(\mathbb{P}_0, \mathbb{P}_\phi) = |\phi| \qquad \mathbb{JS}(\mathbb{P}_0, \mathbb{P}_\phi) = \log(2)\, \delta(\mathbb{P}_0, \mathbb{P}_\phi)$$

Note that among all these divergences, illustrated in Fig. A.2, only the Wasserstein distance provides a continuous (even a.e. differentiable) objective on ϕ. We will now study the behavior of the GAIT divergence in this setting.

Recall that the action of the kernel on a given probability measure corresponds to the mean map $\mathbf{K}\mu : \mathcal{X} \to \mathbb{R}$, defined by $\mathbf{K}\mu(x) \triangleq \mathbb{E}_{x' \sim \mu}\left[\kappa(x, x')\right] = \int \kappa(x, x')\, \mathrm{d}\mu(x')$. In particular, for $\mathbb{P}_\phi$:

$$\mathbf{K}\mathbb{P}_\phi(x, y) = \int_{\mathbb{R}^2} \kappa((x, y), (x', y'))\, \mathrm{d}\mathbb{P}_\phi(x', y') = \int_0^1 \kappa((x, y), (\phi, y'))\, \mathrm{d}y'.$$

Let us endow $\mathbb{R}^2$ with the Euclidean norm $\|\cdot\|_2$, and define the kernel $\kappa((x, y), (x', y')) \triangleq \exp(-\|(x, y) - (x', y')\|_2^2)$. Note that this choice is made only for its mathematically convenience in the following algebraic manipulation, but other choices of kernel are possible. In this case, the mean map reduces to:

$$\mathbf{K}\mathbb{P}_\phi(x, y) = \int_0^1 \exp\left[-\left((x-\phi)^2 + (y-y')^2\right)\right]\, \mathrm{d}y' = \exp\left[-(x-\phi)^2\right] \underbrace{\int_0^1 \exp\left[-(y-y')^2\right]\, \mathrm{d}y'.}_{\triangleq\, I_y,\ \text{independent of } \phi.}$$

We obtain the following expressions for the terms appearing in the divergence:

$$\mathbb{E}_{(x,y)\sim\mathbb{P}_\phi} \log\left[\frac{\mathbf{K}\mathbb{P}_\phi(x, y)}{\mathbf{K}\mathbb{P}_0(x, y)}\right] = \mathbb{E}_{(x,y)\sim\mathbb{P}_\phi} \log\left[\frac{\exp\left[-(x-\phi)^2\right] \cancel{I_y}}{\exp\left[-x^2\right] \cancel{I_y}}\right] \phi^2.$$

$$\mathbb{E}_{(x,y)\sim\mathbb{P}_0} \left[\frac{\mathbf{K}\mathbb{P}_\phi(x, y)}{\mathbf{K}\mathbb{P}_0(x, y)}\right] = \mathbb{E}_{(x,y)\sim\mathbb{P}_0} \exp\left[x^2 - (x-\phi)^2\right] = \exp\{-\phi^2\}.$$

Finally, we replace the previous expressions in the definition of the GAIT divergence. Remarkably, the result is a smooth function of the parameter ϕ with a global optimum at $\phi = 0$. See Fig. A.2.

$$\mathbb{D}^{\mathbf{K}}(\mathbb{P}_\phi, \mathbb{P}_0) = 1 + \mathbb{E}_{(x,y)\sim\mathbb{P}_\phi} \log\left[\frac{\mathbf{K}\mathbb{P}_\phi(x, y)}{\mathbf{K}\mathbb{P}_0(x, y)}\right] - \mathbb{E}_{(x,y)\sim\mathbb{P}_0} \left[\frac{\mathbf{K}\mathbb{P}_\phi(x, y)}{\mathbf{K}\mathbb{P}_0(x, y)}\right] = \phi^2 + 1 - e^{-\phi^2} \geq 0.$$

A.2 Proofs

Theorem 9. $-\nabla^2_{\mathbb{P}}[\mathbb{H}_1^{\mathbf{K}}[\mathbb{P}]]$ *is positive definite in the 2-dimensional case.*

Proof. In this case, we can express the distribution $\mathbb{P}$ in terms of a single degree of freedom p as $[p, 1-p]$. Moreover, denote the off-diagonal kernel entry as $0 \leq a < 1$. Thus, the GAIT entropy can be written as:

$$-\mathbb{H}_1[p] = p \log[p + (1-p)a] + (1-p) \log[pa + (1-p)].$$

The first and second derivatives of $-\mathbb{H}_1[p]$ with respect to p are given by:

$$-\nabla_p[\mathbb{H}_1[p]] = \frac{a}{1+(a-1)p} - \frac{a}{a+(1-a)p} - \log[1+(a-1)p] + \log[a+(1-a)p]$$

$$\begin{aligned}-\nabla_p^2[\mathbb{H}_1[p]] &= -\frac{(a-1)^2(1-p)}{(1+(a-1)p)^2} - \frac{(1-a)^2 p}{(a+(1-a)p)^2} + \frac{2(1-a)}{1+(a-1)p} + \frac{2(1-a)}{a+(1-a)p} \\ &= \frac{a(2-a-a^3)+(a-1)^4 p(1-p)}{(1+(a-1)p)^2(a+(1-a)p)^2}.\end{aligned}$$

This is clearly greater than zero whenever $p \in (0,1)$, since $a \in [0,1)$. □

Theorem 10. *Let X, Y be independent, then $\mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda}}[X,Y] = \mathbb{H}^{\mathbf{K}}[X] + \mathbb{H}^{\mathbf{\Lambda}}[Y]$.*

Proof.

$$\begin{aligned}\mathbb{H}^{\mathbf{K}\otimes\mathbf{J}}[X,Y] &= \mathbb{E}_{x,y} \log\left[\mathbb{E}_{x',y'} \kappa(x,x')\lambda(y,y')\right] \\ &= \mathbb{E}_{x,y} \log\left[\mathbb{E}_{x'}\left[\mathbb{E}_{y'} \kappa(x,x')\lambda(y,y')\right]\right] \\ &= \mathbb{E}_{x,y} \log\left[\mathbb{E}_{x'}\left[\kappa(x,x')\right] \mathbb{E}_{y'}\left[\lambda(y,y')\right]\right] \\ &= \mathbb{E}_{x,y} \log\left[\mathbb{E}_{x'}\left[\kappa(x,x')\right]\right] + \log\left[\mathbb{E}_{y'}\left[\lambda(y,y')\right]\right] \\ &= \mathbb{E}_{x} \log\left[\mathbb{E}_{x'}\left[\kappa(x,x')\right]\right] + \mathbb{E}_{y} \log\left[\mathbb{E}_{y'}\left[\lambda(y,y')\right]\right] \\ &= \mathbb{H}^{\mathbf{K}}[X] + \mathbb{H}^{\mathbf{\Lambda}}[Y].\end{aligned}$$

□

Theorem 11. *For any kernel κ, $\mathbb{H}^{\mathbf{K},\mathbf{I}}[X|Y] = \mathbb{E}_{y\sim\mathbb{P}_y}[\mathbb{H}^{\mathbf{K}}[X|Y=y]]$.*

Proof.

$$\begin{aligned}
\mathbb{H}^{\mathbf{K},\mathbf{I}}[X|Y] &= \mathbb{H}^{\mathbf{K},\mathbf{I}}[X,Y] - \mathbb{H}^{\mathbf{I}}[Y] \\
&= \mathbb{E}_{x,y} \log \left[\mathbb{E}_{x',y'} \kappa(x,x') \mathbf{1}(y,y')\right] - \mathbb{H}^{\mathbf{I}}[Y] \\
&= \mathbb{E}_{x,y} \log \left[\int_{x'} \int_{y'} p(x',y') \kappa(x,x') \mathbf{1}(y,y')\right] - \mathbb{H}^{\mathbf{I}}[Y] \\
&= \mathbb{E}_{x,y} \log \left[\int_{x'} p(x'|y) p(y) \kappa(x,x')\right] - \mathbb{H}^{\mathbf{I}}[Y] \\
&= \mathbb{E}_{x,y} \log \left[p(y) \mathbb{E}_{x'|y} \kappa(x,x')\right] - \mathbb{H}^{\mathbf{I}}[Y] \\
&= \mathbb{E}_{x,y} \log \left[\mathbb{E}_{x'|y} \kappa(x,x')\right] = \mathbb{E}_{y} \left[\mathbb{E}_{x|y} \log \left[\mathbb{E}_{x'|y} \kappa(x,x')\right]\right] \\
&= \mathbb{E}_{y} \left[\mathbb{H}^{\mathbf{K}}[X|y]\right] .
\end{aligned}$$

□

Theorem 12. ♣ *For any similarity kernel* κ, $\mathbb{H}^{\mathbf{K},\mathbf{I}}[X|Y] \leq \mathbb{H}^{\mathbf{K}}[X]$.

Proof. $\mathbb{H}^{\mathbf{K},\mathbf{I}}[X|Y] = \mathbb{E}_{y\sim\mathbb{P}_y}[\mathbb{H}^{\mathbf{K}}[X|Y=y]] = \mathbb{E}_{y\sim\mathbb{P}_y}[\mathbb{H}^{\mathbf{K}}[X|Y=y]] \overset{\text{(Jensen)}}{\leq} \mathbb{H}^{\mathbf{K}}[\mathbb{E}_{y\sim\mathbb{P}_y}[X|Y=y]] = \mathbb{H}^{\mathbf{K}}[X]$. □

Lemma 1. (Chain Rule of Mutual Information)♣. $\mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y,Z] = \mathbb{I}^{\mathbf{K},\mathbf{\Lambda}}[X;Y] + \mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y|Z]$

Proof. By definition:

$$\mathbb{I}^{\mathbf{K},\mathbf{\Theta}}[X;Z] = \mathbb{H}^{\mathbf{K}}[X] + \mathbb{H}^{\mathbf{\Theta}}[Z] - \mathbb{H}^{\mathbf{K}\otimes\mathbf{\Theta}}[X,Z].$$

$$\begin{aligned}
\mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y|Z] &= \mathbb{H}^{\mathbf{K},\mathbf{\Theta}}[X|Z] + \mathbb{H}^{\mathbf{\Lambda},\mathbf{\Theta}}[Y|Z] - \mathbb{H}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X,Y|Z] \\
&= \mathbb{H}^{\mathbf{K},\mathbf{\Theta}}[X,Z] - \mathbb{H}^{\mathbf{\Theta}}[Z] + \mathbb{H}^{\mathbf{\Lambda},\mathbf{\Theta}}[Y,Z] - \mathbb{H}^{\mathbf{\Theta}}[Z] - \mathbb{H}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X,Y,Z] + \mathbb{H}^{\mathbf{\Theta}}[Z]
\end{aligned}$$

Thus, $\mathbb{I}^{\mathbf{K},\mathbf{\Theta}}[X;Z] + \mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y|Z] = \mathbb{H}^{\mathbf{K}}[X] + \mathbb{H}^{\mathbf{\Lambda},\mathbf{\Theta}}[Y,Z] - \mathbb{H}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X,Y,Z] = \mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y,Z]$. □

Theorem 13. (Data Processing Inequality)♣.
If $X \to Y \to Z$ *is a Markov chain,* $\mathbb{I}^{\mathbf{K},\mathbf{\Theta}}[X;Z] \leq \mathbb{I}^{\mathbf{K},\mathbf{\Lambda}}[X;Y] + \mathbb{I}^{\mathbf{K},\mathbf{\Theta},\mathbf{\Lambda}}[X;Z|Y]$.

Proof.

$$\mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y,Z] \overset{\text{(Thm. 1)}}{=} \mathbb{I}^{\mathbf{K},\mathbf{\Lambda}}[X;Y] + \mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y|Z] \overset{\text{(Thm. 1)}}{=} \mathbb{I}^{\mathbf{K},\mathbf{\Theta}}[X;Z] + \mathbb{I}^{\mathbf{K},\mathbf{\Theta},\mathbf{\Lambda}}[X;Z|Y].$$

Therefore $\mathbb{I}^{\mathbf{K},\mathbf{\Lambda}}[X;Z] + \mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y|Z] = \mathbb{I}^{\mathbf{K},\mathbf{\Theta}}[X;Z] + \mathbb{I}^{\mathbf{K},\mathbf{\Theta},\mathbf{\Lambda}}[X;Z|Y]$. Finally, we have that $\mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y|Z] \geq 0$, which in turn implies that $\mathbb{I}^{\mathbf{K},\mathbf{\Lambda}}[X;Z] \leq \mathbb{I}^{\mathbf{K},\mathbf{\Lambda}}[X;Y] + \mathbb{I}^{\mathbf{K},\mathbf{\Theta},\mathbf{\Lambda}}[X;Z|Y]$. □

Additionally, we are able to prove a series of inequalities illuminating the influence of the similarity matrix on joint entropy in extreme cases:

Theorem 7. *For any similarity kernels κ and λ, $\mathbb{H}^{\mathbf{K}}[X] = \mathbb{H}^{\mathbf{K}\otimes\mathbf{J}}[X,Y] \leq \mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda}}[X,Y] \leq \mathbb{H}^{\mathbf{K}\otimes\mathbf{I}}[X,Y] = \mathbb{H}^{\mathbf{I}}[Y] + \mathbb{H}^{\mathbf{K},\mathbf{I}}[X|Y]$*

Proof. The first result, $\mathbb{H}^{\mathbf{K}}[X] = \mathbb{H}^{\mathbf{K}\otimes\mathbf{J}}[X,Y]$ follows by noting that $\lambda(y,y') = 1$ for all y, y':

$$\mathbb{H}^{\mathbf{K}\otimes\mathbf{J}}[X,Y] = \mathbb{E}_{x,y}\log\left[\mathbb{E}_{x',y'}\kappa(x,x')\lambda(y,y')\right] = \mathbb{E}_{x}\log\left[\mathbb{E}_{x'}\kappa(x,x')\right] = \mathbb{H}^{\mathbf{K}}[X]$$

$\mathbb{H}^{\mathbf{K}\otimes\mathbf{J}}[X,Y] \leq \mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda}}[X,Y] \leq \mathbb{H}^{\mathbf{K}\otimes\mathbf{I}}[X,Y]$ follows by monotonicity of the entropy in the similarity matrices.

$\mathbb{H}^{\mathbf{K}\otimes\mathbf{I}}[X,Y] = \mathbb{H}^{\mathbf{I}}[Y] + \mathbb{H}^{\mathbf{K},\mathbf{I}}[X|Y]$ follows by the chain rule of conditional entropy. □

A.3 Verifying the concavity of $\mathbb{H}_1^{\mathbf{K}}[\cdot]$

A.3.1 Proof attempts

We have made several attempts to show that the GAIT entropy is a concave function at $\alpha = 1$. As this is a critical component in our theoretical developments, we provide a list of our previously unsuccessful approaches, in the hopes of facilitating the participation of interested researchers in answering this question.

- Jensen's inequality for the $\log(\mathbf{Kp})$ or $\log(\mathbf{Kq})$ terms is too loose.

- The bound $\log b \leq \frac{b}{a} + \log(a) - 1$ applied to the ratio $\frac{\mathbf{Kp}}{\mathbf{Kq}}$ results in a loose bound.

- $-p\log(p)$ is known to be a concave function. However, the action of the similarity matrix on the distribution inside the logarithmic factor in $-\mathbf{p}^T \log(\mathbf{Kp})$ complicates the analysis.

- The Donsker-Varadhan representation of the Kullbach-Leibler divergence goes in the wrong direction and adds extra terms.

- Bounding a Taylor series expansion of the gap between the linear approximation of an interpolation and the value of the entropy along the interpolation. The analysis is promising but becomes unwieldy due to the presence of $\frac{\mathbf{Kq}}{\mathbf{Kp}}$ terms.

A.3.2 Numerical experiments

Random search on $\mathbb{D}^{\mathbf{K}}[\mathbf{p}||\mathbf{q}] \geq 0$. We perform a search over vectors $\mathbf{p}$ and $\mathbf{q}$ drawn randomly from the simplex, and over random positive definite similarity Gram matrices $\mathbf{K}$. We have tried restricting our searches to $\mathbf{p}$ and $\mathbf{q}$ near the center of the simplex and away from the center, and to $\mathbf{K}$ closer to $\mathbf{I}$ or $\mathbf{J}$. In every experiment, we find that $\mathbb{D}^{\mathbf{K}}[\mathbf{p}||\mathbf{q}] \geq 0$.

Quantity	Sampling process
$n \in \mathbb{Z}$	$n \sim \text{Uniform}(\{2, \ldots, 11\})$
$\boldsymbol{\gamma} \in \mathbb{Z}^{n\times n}$	$\gamma_{i,j} \sim \text{Uniform}(\{0, \ldots, 9\})$
$\mathbf{L} \in \mathbb{R}^{n\times n}$	$L_{i,j} \sim \text{Uniform}(0,1)^{\gamma_{i,j}}$
$\mathbf{K} \in \mathbb{R}^{n\times n}$	$\mathbf{K} = \min(1, \mathbf{I} + \mathbf{L}\mathbf{L}^T/n)$
$\boldsymbol{\alpha} \in \mathbb{R}^n$	$\alpha_i \sim \text{Uniform}(0,10)$
$\boldsymbol{\beta} \in \mathbb{R}^n$	$\beta_i \sim \text{Uniform}(0,10)$
$\mathbf{p} \in \boldsymbol{\Delta}_n$	$\mathbf{p} \sim \text{Dirichlet}(\boldsymbol{\alpha})$
$\mathbf{q} \in \boldsymbol{\Delta}_n$	$\mathbf{q} \sim \text{Dirichlet}(\boldsymbol{\beta})$

Table A.1 – Experimental setup for random search.

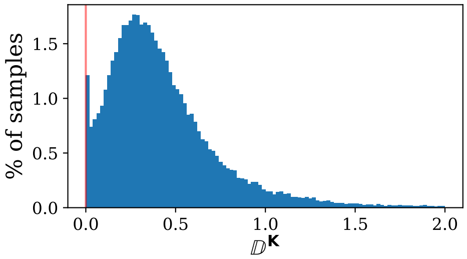

Figure A.3 – Histogram of GAIT entropies from settings in Tab. A.1.

Consider the wide experimental setup for search defined in Tab. A.1. Fig. A.3 shows the histogram of $\mathbb{D}^{\mathbf{K}}$ over this search, empirically showing the non-negativity of the divergence, and, thus the concavity of the GAIT entropy.

Random search on $-\nabla^2_{\mathbb{P}}[\mathbb{H}^{\mathbf{K}}_1[\mathbb{P}]]$. We empirically study the positive definiteness of this matrix via its spectrum. For this, we sample a set of n points in $\mathbb{R}^d$ as well as a (discrete) distribution $\mathbb{P}$ over those points. Then we construct the Gram matrix induced by the kernel $\kappa(x, y) = \exp(-||x - y||_p)$. The location of the points, $\mathbb{P}$, n, d and $p >= 1$ are sampled randomly.

We performed extensive experiments under this setting and never encountered an instance such that $-\nabla^2_{\mathbb{P}}[\mathbb{H}^{\mathbf{K}}_1[\mathbb{P}]]$ would have any negative eigenvalues. We believe this experimental setting is more holistic than the above experiments since it considers the whole spectrum of the (negative) Hessian rather than a "directional derivative" towards another sampled distribution $\mathbb{Q}$.

Optimization. As an alternative to random search, we also use gradient-based optimization on $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{K}$ to minimize $\mathbb{D}^{\mathbf{K}}[\mathbf{p}||\mathbf{q}]$. Starting from random initializations, our objective function always converges to values very close to (yet above) zero.

Furthermore, freezing $\mathbf{K}$ and optimizing over either $\mathbf{p}$ or $\mathbf{q}$ while holding the other fixed, results in $\mathbf{p} = \mathbf{q}$ at convergence. On the other hand, if $\mathbf{p}$ and $\mathbf{q}$ are fixed such that $\mathbf{p} \neq \mathbf{q}$, optimization over $\mathbf{K}$ converges to $\mathbf{K} = \mathbf{J}$. We note from the definition of the GAIT divergence that when $\mathbf{p} = \mathbf{q}$ or $\mathbf{K} = \mathbf{J}$, $\mathbb{D}^{\mathbf{K}}[\mathbf{p}||\mathbf{q}] = 0$, which matches the value we obtain at convergence when trying to minimize this quantity.

Recall that the experiments presented in Sec. 3.3 involve the minimization of some GAIT divergence. We never encountered a negative value for the GAIT divergence during any of these experiments.

A.3.3 Finding maximum entropy distributions with gradient ascent

An algorithm with an exponential run-time to find *exact* maximizers of the entropy $\mathbb{H}^{\mathbf{K}}_\alpha[\cdot]$ is presented in Leinster and Meckes (2016). We exploit the fact that the objective is amenable to gradient-based optimization techniques and conduct experiments in spaces with thousands of elements. This also serves as an empirical test for the conjecture about the concavity of the function: there must be a unique maximizer for $\mathbb{H}^{\mathbf{K}}_1[\cdot]$ if it is concave.

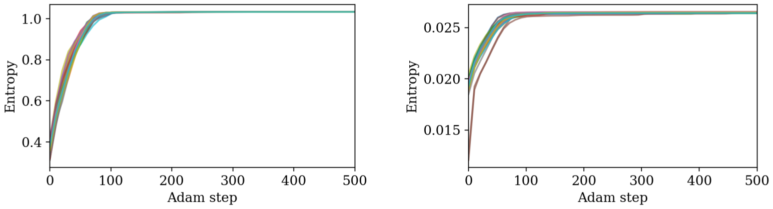

Figure A.4 – Optimization curves for measures with support 1000 in dimension 5 (left) and 100 (right).

We test our ability to find distributions with maximum GAIT entropy via gradient descent. We sample 1000 points in dimensions 5 and 10, and construct a similarity space using a RBF kernel with $\sigma = 1$. Then we perform 100 trials by setting the logits of the initialization using a Gaussian distribution with variance 4 for each of the 1000 logits that describe our distribution. We use Adam with learning rate 0.1 and $\alpha = 1$. The optimization results are shown in Fig. A.4. We reliably obtain negligible variance in the objective value at convergence across random initializations, thus providing an efficient alternative for finding approximate maximum-entropy distributions.

A.4 Experimental details

A.4.1 Interpolation and Approximation

In all experiments for Figs. 3.7-3.5, we minimize the GAIT divergence using AMSGrad (Reddi et al., 2018) in PyTorch (Paszke et al., 2019). We parameterize the weights of empirical distributions using a softmax function on a vector of temperature-scaled logits. All experiments in the section are run on a single CPU.

Approximating measures with finite samples

In Fig. 3.7 we optimize our approximating measure using Adam for 3000 steps with a learning rate of 10^{-3} and minibatches constructed by sampling 50 examples at each step. We use a Gaussian kernel with $\sigma = 0.02$.

In Fig. 3.8, we approximate a continuous measure with an empirical measure supported on 200 atoms. We execute Adam for 500 steps using a learning rate of

0.05 and minibatches of 100 samples from the continuous measure to estimate the discrepancy. The similarity function is given by a polynomial kernel with exponent 1.5: $\kappa(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{1+\|\mathbf{x}-\mathbf{y}\|^{1.5}}$. Fig. A.5 shows that we achieve results of comparable quality to those of Claici et al. (2018).

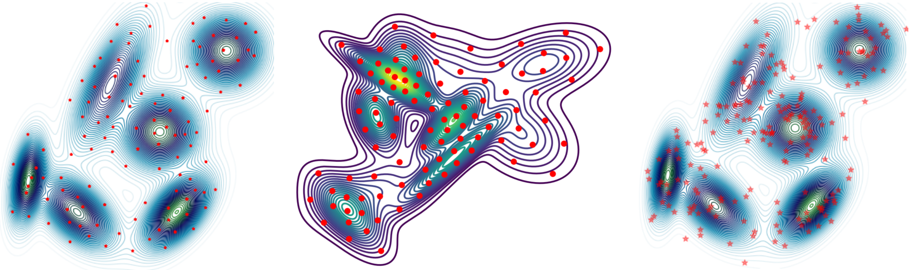

Figure A.5 – **Left and Center:** Approximation of a mixture of Gaussians density using our method and the proposal of Claici et al. (2018) (taken from paper). **Right:** i.i.d samples from the real data distribution.

Image barycenters

We compute barycenters for each class of MNIST and Fashion-MNIST. We perform gradient descent with Adam using a learning rate of 0.01 with minibatches of size 32 for 500 optimization steps. We use a Gaussian kernel with $\sigma = 0.04$. The geometry of the grid on which images are defined is given by the Euclidean distance between the coordinates of the pixels. In Fig. A.6, we provide barycenters for the each of classes of Fashion MNIST computed via a combination of the methods of Benamou et al. (2014) and Cuturi and Doucet (2014).

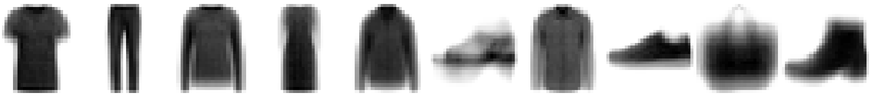

Figure A.6 – Barycenters for Fashion MNIST computed using our method.

Text summarization

For our text example, we use the article from the STAT-MT parallel news corpus titled "Why Wait for the Euro?", by Leszek Balcerowicz. The full text of the article can be found at `https://pastebin.com/CnBgbpsJ`. We use the 300-dimensional GLoVe vectors found at `http://nlp.stanford.edu/data/glove.6B.zip` as word embeddings. TF-IDF is calculated over the entire English portion of the parallel news corpus using the implementation in Scikit-Learn (Pedregosa et al., 2011). We filter stopwords based on the list provided by the Natural Language Toolkit (Bird et al., 2009). To encourage sparsity in the approximating measure $\mathbf{q}$, we add the 0.75-norm of $\mathbf{q}$ to the divergence loss, weighted by a factor of 0.01. We optimize the loss with gradient descent using Adam optimizer, with hyperparameters $\beta_1 = 0, \beta_2 = 0.9,$ learning rate $= 0.001$, for 25,000 iterations. Since a truly sparse $\mathbf{q}$ is not reachable using the softmax function and gradient descent, we set all entries $\mathbf{q}_i < 0.01$ to be 0 and renormalize after the end of training. $\mathbf{q}$ is represented by the softmax function, and is initialized uniformly.

We examine the influence of varying σ in Fig. A.7. Decreasing σ leads to $\mathbf{K}$ approaching $\mathbf{I}$, and the resulting similarity more closely approximates the original measure. As σ approaches 0.01, the two measures become almost identical. See Fig. A.7, bottom-left and bottom-right.

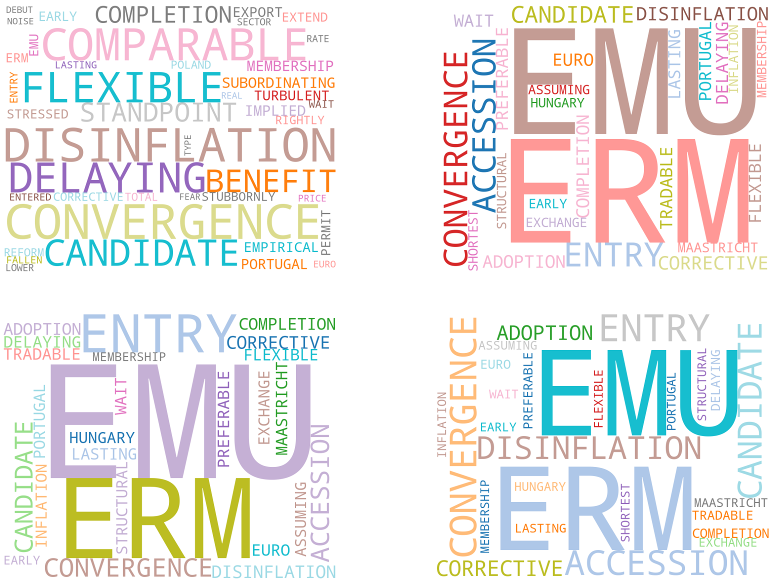

Figure A.7 – **Top-left:** Word cloud generated by our method at $\sigma = 0.5$. **Top-right:** Word cloud generated by our method at $\sigma = 0.1$. **Bottom-left:** Word cloud generated by our method at $\sigma = 0.01$. **Bottom-right:** Top 43 original TF-IDF words.

A.4.2 GAN evaluation and mode counting

When the data available takes the form of many i.i.d. samples from a continuous distribution, a natural choice is to generate a Gram matrix $\mathbf{K}$ using a similarity measure such as an RBF kernel $\kappa_\sigma(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}\right)$.

For comparison, we adapt the birthday paradox-based approach of Arora et al. (2018). Strictly speaking, their method requires human evaluation of possible duplicates, and is thus not comparable to our approach. As such, we propose an automated version using the same assumptions. We define $\mathbf{x}$ and $\mathbf{y}$ as colliding when $d(\mathbf{x}, \mathbf{y}) < \epsilon$, and note that the expected number of collisions for a distribution with support n in a sample of size m is $c = \frac{m(m-1)}{n}$. We can thus estimate $\hat{n} = \frac{m(m-1)}{c}$. When varying ϵ, we observe behavior very similar to that of our entropy measure, with a plateau at $\hat{n} = C$ in our example of a mixture of C Gaussians. The results of this comparison are presented in Fig. 3.10.

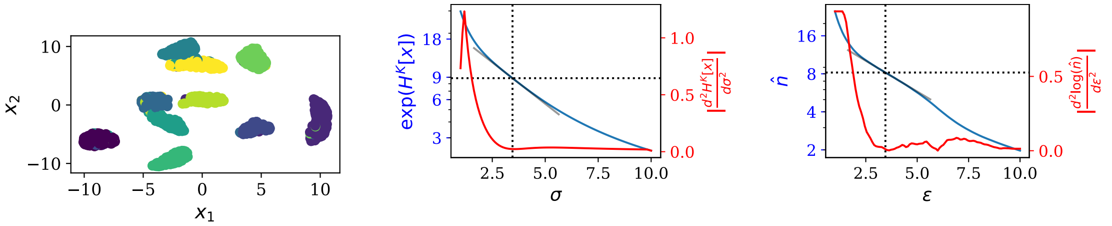

Figure A.8 – **Left:** A 2000-image subset of MNIST reduced to 2 dimensions by UMAP. **Center:** Our mode estimation. **Right:** The birthday paradox method estimate. Note that the left axis is logarithmic.

To test this on a more challenging dataset, we use a 2-dimensional representation for MNIST obtained using UMAP (McInnes et al., 2018), shown in Fig. A.8. Although our method no longer shows a clear plateau at $\mathbb{H}_1^{\mathbf{K}}[\mathbb{P}] \approx \log 10 \approx 2.3$, it does transition from exponential to linear decay at approximately this point, which coincides with the point of minimum curvature with respect to σ, $\mathbb{H}_1^{\mathbf{K}}[\mathbb{P}] \approx \log 10$. Similar behavior is observed in the case with birthday-inspired estimate; here the point of minimum curvature has $\hat{n} \approx 8$.

Finally, we also apply this method to evaluating the diversity of GAN samples. We train a simple WGAN (Arjovsky et al., 2017) on MNIST, and find that the assessed entropy increases steadily as training progresses and the generator masters more modes (see Fig. A.9). Note that the entropy estimate stabilizes once the generator begins to produce all 10 digits, but long before sample quality ceases improving.

In all of the experiments corresponding to mode counting, we use $\alpha = 1$ and the standard RBF kernel $\kappa_\sigma(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}\right)$. Note that this differs from the kernel given in Section 3.1 by using squared Euclidean distance rather than

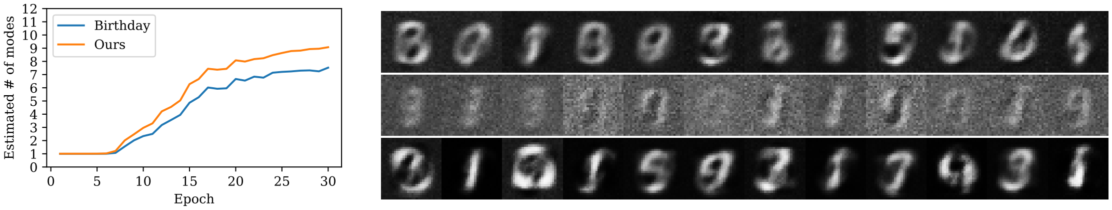

Figure A.9 – **Left:** The estimated numbers of modes in the output of a WGAN trained on MNIST. **Right:** Samples from the same WGAN after 5, 15 and 25 epochs.

Euclidean distance. To estimate the point with minimum curvature, we find the value of $\log \hat{n}$ or $\mathbb{H}_1^{\mathbf{K}}[\mathbf{p}]$ at 100 values of σ or ϵ evenly spaced between 0.1 and 25, and empirically estimate the second derivative with respect to σ or ϵ. In the case of the birthday estimate, which is not continuous on finite sample sizes, we use a Savitzky-Golay filter (Savitzky, 1964) of degree 3 and window size 11 to smooth the derivatives. We estimate the point of minimum curvature to be the first point when the absolute second derivative passes below 0.01.

To evaluate GANs, we train a simple WGAN-GP (Gulrajani et al., 2017) with a 3-hidden-layer fully-connected generator, using the ReLU nonlinearity and 256 units in each hidden layer, on a TITAN Xp GPU. Our latent space has 32 dimensions sampled i.i.d. from $\mathcal{N}(0, 1)$ and the discriminator is trained for four iterations for each generator update. We use the Adam with learning rate 10^{-4} and $\beta_1 = 0$, $\beta_2 = 0.9$. The weight of the gradient penalty in the WGAN-GP objective is $\lambda = 10$.

To count the number of modes in the output of the generator, we use an instance of UMAP fitted to the entire training set of MNIST to embed all input in $\mathbb{R}^2$. We use 1,000 samples of true MNIST data to estimate values of σ (for our entropy method) and ϵ for the birthday paradox-based method that minimize curvature and yield estimates of $\exp \mathbb{H}_1^{\mathbf{K}}[\mathbf{p}] \approx 10$ and $\hat{n} \approx 10$. We then apply these to the generated outputs after each of the first 30 epochs, and report $\hat{n}$ or $\exp \mathbb{H}_1^{\mathbf{K}}[\mathbf{p}]$.

A.4.3 Generative models

For all the generative models in Section 3.3.1, we employ an experimental setup similar to the setup used by Genevay et al. (2018) for learning generative models on MNIST. Thus, our generative model is a 2-layer multilayer perceptron with one hidden layer of 500 dimensions with ReLU non-linearities, using a 2D latent space, trained using mini-batches of size 200. Note that their method requires a batch size of 200 to get reasonable generations, but we also obtain comparable results with a significantly smaller batch size of 50. Since Genevay et al. (2018) sample latent codes from a unit square, we do the same for MNIST here for easy comparison but sample from a standard Gaussian for Swiss roll and Fashion-MNIST datasets. We train our models by minimizing $\mathbb{D}^{\mathbf{K}}[\hat{\mathbb{P}}\,||\,\hat{\mathbb{Q}}]$, where $\hat{\mathbb{P}}$ is the target empirical measure and $\hat{\mathbb{Q}}$ is the model. $\mathbf{K}$ is the Gram matrix corresponding to a RBF kernel with $\sigma = 0.2$ for Swiss roll data, and $\sigma = 1.6$ for MNIST and Fashion-MNIST. We use Adam with a learning rate of 5×10^{-4} to train our models. Fig. A.10 compares the manifolds learned by minimizing our divergence with batch sizes 200 and 50 with that learned by minimizing the Sinkhorn loss (Genevay et al., 2018) for MNIST.

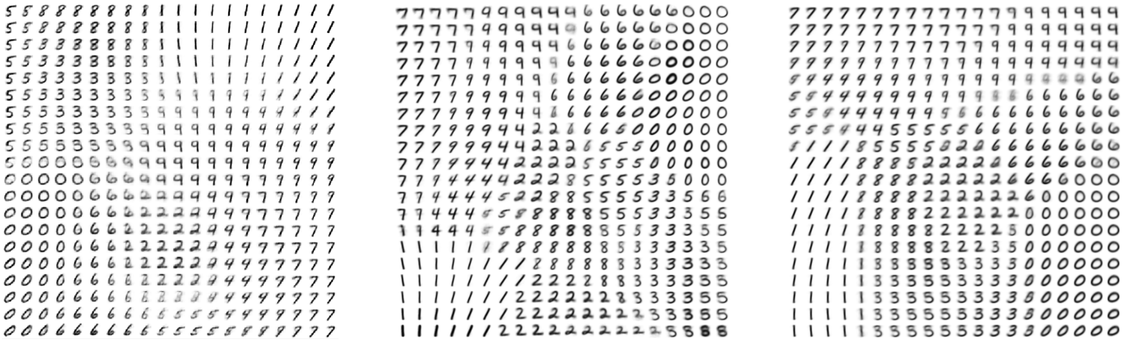

Figure A.10 – **Left:** Manifold learned by minimizing Sinkhorn loss, taken from Genevay et al. (2018). **Center:** Manifold learned by minimizing GAIT divergence using their experimental setup. **Right:** Manifold learned by minimizing GAIT divergence with batch size 50.

We further compare our generations with those done by variational auto-encoders (Kingma and Welling, 2014). Following their setup, we use tanh as the non-linearity in the 2-layer multilayer perceptron and a lower batch size of 100, along with the latent codes sampled from a standard Gaussian distribution. We compare our results with theirs in Fig. A.11. Both figures are generated using latent codes obtained by taking the inverse c.d.f. of the Gaussian distribution at the corresponding grid locations, similar to the work of Kingma and Welling (2014).

Finally, in Fig. A.12, we illustrate Fashion-MNIST and MNIST samples generated by our generative model with a 20D latent space. The quality of our generations with a 20D latent space is comparable to the samples generated by the variational auto-encoder with the same latent dimensions in Kingma and Welling (2014).

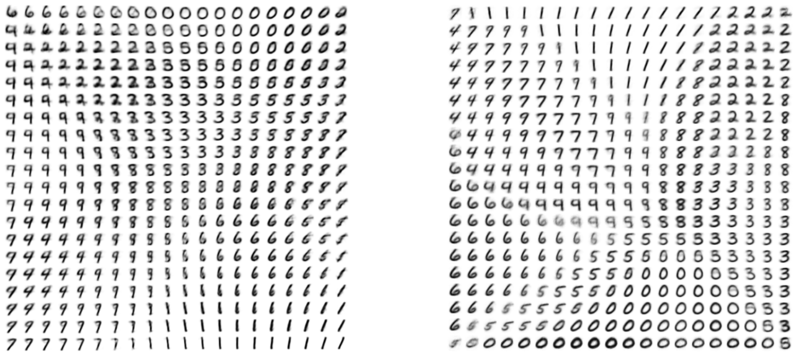

Figure A.11 – **Left:** Manifold learned by Variational Autoencoder, taken from Kingma and Welling (2014). **Right:** Manifold learned by minimizing GAIT divergence using their experimental setup.

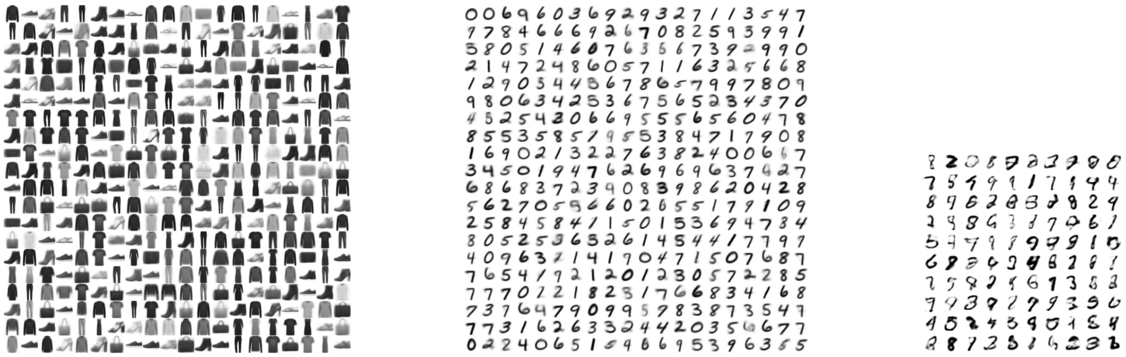

Figure A.12 – **Left:** Fashion-MNIST samples from our model with 20D latent space. **Center:** MNIST samples from our model with 20D latent space. **Right:** MNIST samples from Variational Autoencoder with 20D latent space, picture taken from (Kingma and Welling, 2014).

A.4.4 Computational complexity

Solomon et al. (2015) shows how the computation of $\mathbf{K}\mathbb{P}$ can be efficiently performed using convolutions in the case of image-like data. For $d \times d$ images, this takes time $\mathcal{O}(d^3)$, instead of $\mathcal{O}(d^4)$ using a naive approach. Sinkhorn regularized optimal transport requires performing this computation this computation L, which highlights the value of the work of Solomon et al. (2015) for applications with large d. The complexity for computing the close-form GAIT divergence is thus $\mathcal{O}(d^3)$, and the cost for approximating solving the optimal transport problem via Sinkhorn iterations is $\mathcal{O}(Ld^3)$. We draw the attention of the reader to the distinction between the width d of the image, and the size of the support of the measures, $n = d^2$.

Fig. A.13 shows compares the time required by the convolutional approaches of

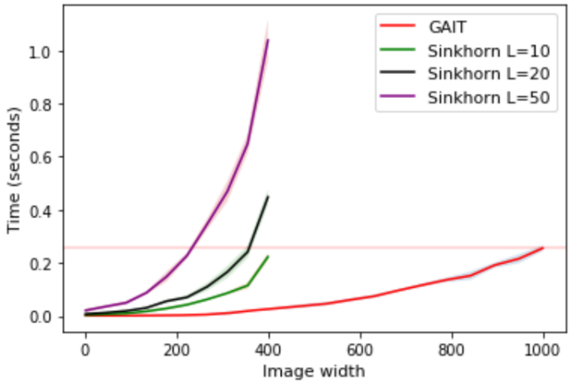

Figure A.13 – Time comparison between the computation of the GAIT and Sinkhorn divergences between randomly generated images of varying size. Error bars correspond to one standard deviation over a sample of size 30.

the GAIT divergence computation and the Sinkhorn algorithm approximating the Sinkhorn divergence, between two images of size $d \times d$. Genevay et al. (2018) found $L = 100$ necessary to perform well on generative modeling. Even for the comparatively low values of L presented in Fig. A.13, we observe that the computation of the GAIT divergence is significantly faster than that of the approximate Sinkhorn divergence. It is possible to compute the GAIT divergence between two images of one megapixel in a quarter of a second (horizontal line).

Bibliography

A. Achille, G. Paolini, and S. Soatto. Where is the Information in a Deep Neural Network? Technical report, 2019.

J. Aczél and Z. Daróezy. *On Measures of Information and Their Characterizations*, volume 115. Academic Press, New York, 1975.

A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy. Fixing a Broken ELBO. In *ICML*, 2018.

S.-I. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10 (2):251–276, 1998.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

S. Arora, A. Risteski, and Y. Zhang. Do GANs Learn the Distribution? Some Theory and Empirics. In *ICLR*, 2018.

J. C. Baez, T. Fritz, and T. Leinster. A Characterization of Entropy in Terms of Information Loss. *Entropy*, 13(11):1945–1957, jun 2011.

M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm. MINE: Mutual Information Neural Estimation. In *ICML*, 2018.

S. Ben-David and S. Shalev-Shwartz. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman Projections for Regularized Transportation Problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2014.

S. Bird, E. Loper, and E. Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

S. Boyd and L. Vandenberghe. *Convex Optimization*. 2004.

L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967.

A. Charpentier. French dataset: population and GPS coordinates, 2012.

Y. Chen, M. Welling, and A. Smola. Super-samples from Kernel Herding. In *UAI*, 2010.

S. Claici, E. Chien, and J. Solomon. Stochastic Wasserstein Barycenters. In *ICML*, 2018.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 2005.

I. Csiszár. Axiomatic characterizations of information measures. *Entropy*, 2008.

I. Csiszár and P. C. Shields. Information Theory and Statistics: A Tutorial. *Foundations and Trends^{TM} in Communications and Information Theory*, 2004.

M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *NeurIPS*. 2013.

M. Cuturi and A. Doucet. Fast Computation of Wasserstein Barycenters. In *ICML*, 2014.

R. Fox, A. Pakman, and N. Tishby. Taming the Noise in Reinforcement Learning via Soft Updates. In *UAI*, 2016.

A. Genevay, G. Peyré, and M. Cuturi. Learning Generative Models with Sinkhorn Divergences. In *AISTATS*, 2018.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *NeurIPS*, 2014.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved Training of Wasserstein GANs. In *NeurIPS*, 2017.

T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement Learning with Deep Energy-Based Policies. In *ICML*, 2017.

S. W. Ho and S. Verdú. Convexity/concavity of Rényi entropy and α-mutual information. In *Proceedings of the IEEE International Symposium on Information Theory*, 2015.

L. V. Kantorovich and G. S. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

S. Lacoste-Julien, F. Huszár, and Z. Ghahramani. Approximate inference for the loss-calibrated Bayesian. In *AISTATS*, 2011.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

T. Leinster. The magnitude of metric spaces. *Documenta Mathematica*, 18:857–905, 2013.

T. Leinster and C. A. Cobbold. Measuring diversity: The importance of species similarity. *Ecology*, 93(3):477–489, 2012.

T. Leinster and M. W. Meckes. Maximizing diversity in biology and beyond. *Entropy*, 18(3):88, 3 2016.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos. MMD GAN: Towards Deeper Understanding of Moment Matching Network. In *NeurIPS*, 2017.

D. J. MacKay. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, 2003.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*, 2018.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Pennington, R. Socher, and C. Manning. Glove: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

J. R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1986.

R. Reams. Hadamard inverses, square roots and products of almost semidefinite matrices. *Linear Algebra and its Applications*, 288:35–43, 2 1999.

S. J. Reddi, S. Kale, and S. Kumar. On the Convergence of Adam and Beyond. In *ICLR*, 2018.

A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1961.

R. Rockafellar. *Convex Analysis*. Princeton Univ. Press, Princeton, N. J, 1970.

T. Salimans, H. Zhang, A. Radford OpenAI, and D. Metaxas. Improving GANs Using Optimal Transport. In *ICLR*, 2018.

M. J. E. Savitzky, Abraham; Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox. On the information bottleneck theory of deep learning. In *ICLR*, 2018.

C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert Space Embedding for Distributions. pages 13–31, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

J. Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *LREC*, 2012.

N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*, 2015.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. 2000.

M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On Mutual Information Maximization for Representation Learning. In *ICLR*, 2020.

C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

L. Wasserstein. Markov processes on the countable product of spaces describing large systems of automata. *Problems Inform. Transmission*, 5(3):64 – 72, 1969. In Russian.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Y. Xu, S. Zhao, J. Song, R. Stewart, and S. Ermon. A Theory of Usable Information Under Computational Constraints. In *ICLR*, 2020.