GAIT: A Geometric Approach to Information Theory

Jose Gallego **Ankit Vani** **Max Schwarzer** **Simon Lacoste-Julien**[†]
Mila and DIRO, Université de Montréal

Abstract

We advocate the use of a notion of entropy that reflects the relative abundances of the symbols in an alphabet, as well as the similarities between them. This concept was originally introduced in theoretical ecology to study the diversity of ecosystems. Based on this notion of entropy, we introduce geometry-aware counterparts for several concepts and theorems in information theory. Notably, our proposed divergence exhibits performance on par with state-of-the-art methods based on the Wasserstein distance, but enjoys a closed-form expression that can be computed efficiently. We demonstrate the versatility of our method via experiments on a broad range of domains: training generative models, computing image barycenters, approximating empirical measures and counting modes.

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.

This blindness limits the applicability of these concepts. 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.

Contributions. 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 refer to this collection of concepts as Geometry-Aware Information Theory: *GAIT*.

Paper structure. We introduce the theory behind the GAIT entropy and provide motivating examples justifying its use. We then introduce and characterize a divergence as well as a definition of mutual information derived from the GAIT entropy. Finally, we demonstrate applications of our methods including training generative models, approximating measures and finding barycenters. We also show that the GAIT entropy can be used to estimate the number of modes of a probability distribution.

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 cer-

[†]Canada CIFAR AI Chair.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

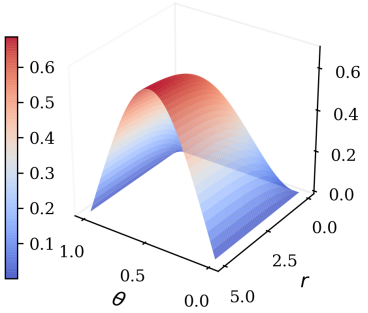

Figure 1: $\mathbb{H}_1^{\mathbf{K}}$ interpolates towards the Shannon entropy as $r \to \infty$.

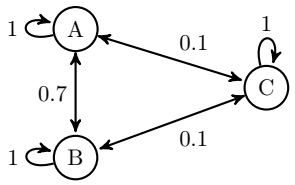

Figure 2: A 3-point space with two highly similar elements.

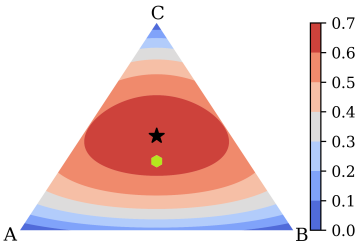

Figure 3: $\mathbb{H}_1^{\mathbf{K}}$ for distributions over the space in Fig. 2.

tain 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 $x > 0$.

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

2 Geometry-Aware Information Theory

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).

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.

2.1 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.

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 $M_{\mathbf{w},\beta}(\mathbf{x}) \triangleq \langle \mathbf{w}, \mathbf{x}^\beta \rangle^{\frac{1}{\beta}}$. Motivated by this averaging scheme, Leinster and Cobbold (2012) proposed the following definition:

Definition 1. *(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:*

$$\mathbb{H}_\alpha^{\mathbf{K}}[\mathbb{P}] \triangleq \log M_{\mathbf{p}, 1-\alpha}\left(\frac{1}{\mathbf{K}\mathbf{p}}\right) \tag{1}$$

$$= \frac{1}{1-\alpha} \log \sum_{i=1}^{n} \mathbf{p}_i \frac{1}{(\mathbf{K}\mathbf{p})_i^{1-\alpha}}. \tag{2}$$

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

$$\mathbb{H}_1^{\mathbf{K}}[\mathbb{P}] = -\langle \mathbf{p}, \log(\mathbf{K}\mathbf{p}) \rangle = -\mathbb{E}_{x \sim \mathbb{P}}[\log(\mathbf{K}\mathbb{P})_x]. \tag{3}$$

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

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. 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. 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 counter-intuitive 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 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 $\mathbf{\Delta}$) 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. 1 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, $\mathbb{H}_0^{\mathbf{K}}[\mathbb{P}] = \log\left\langle \mathbf{p}, \frac{1}{\mathbf{K}\mathbf{p}} \right\rangle$. When $\mathbf{K} = \mathbf{I}$, $\mathbb{H}_0^{\mathbf{K}}[\mathbb{P}] = \log|\mathrm{supp}(\mathbb{P})|$, which values rare and common species equally, while $\mathbb{H}_\infty^{\mathbf{K}}[\mathbb{P}] = -\log\max_{i \in \mathrm{supp}(\mathbf{p})}(\mathbf{K}\mathbf{p})_i$ only considers the most common elements. Thus, in principle, the problem of finding a maximum entropy distribution depends on the choice of α.

Theorem 1. *(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, $\mathbb{H}_\mathcal{X}^* \triangleq \sup_{\mathbb{P} \in \mathbf{\Delta}_{|\mathcal{X}|}} \mathbb{H}_\alpha^{\mathbf{K}}[\mathbb{P}]$ does not depend on α.*

Remarkably, Thm. 1 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 2. *(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. 2. Thm. 2 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 $\mathbf{K}\mathbf{p} = \mathbf{p} = \lambda\mathbf{1}$ for some $\lambda \in \mathbb{R}_{\geq 0}$, is equivalent to the well known fact that the uniform distribution maximizes Shannon entropy.

2.2 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 $\mathbf{K} = \mathbf{I}$. $\mathbb{H}^{\mathbf{I}}_{\alpha}[\cdot]$ 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, $\mathbb{H}^{\mathbf{I}}_{1}[\cdot]$ 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 $\mathbb{H}^{\mathbf{K}}_{1}[\mathbb{P}]$ 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 $\mathbb{H}^{\mathbf{K}}_{1}[\cdot]$ is strictly concave in the interior of $\mathbf{\Delta}_{|\mathcal{X}|}$.*

Fig. 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 $\mathbb{H}^{\mathbf{K}}_{1}[\cdot]$.

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 C. Nevertheless, we provide extensive numerical experiments in App. C which support our conjecture. In the remainder of this work, claims *dependent* on this conjecture are labelled ♣.

2.3 Comparing probability distributions

The previous conjecture implies that $-\mathbb{H}^{\mathbf{K}}_{1}[\cdot]$ 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:

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

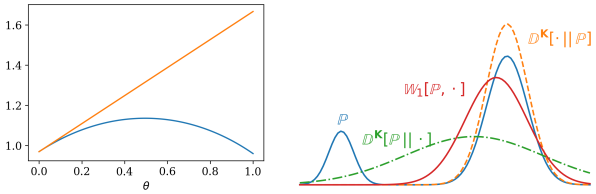

Figure 4: **Left:** The entropy $\mathbb{H}^{\mathbf{K}}_{1}[(1-\theta)\mathbf{q}+\theta\mathbf{p}]$ is upper-bounded by the linear approximation at $\mathbf{q}$, given by $\mathbb{H}^{\mathbf{K}}_{1}[\mathbf{q}] + \theta \left\langle \nabla_{\mathbf{q}} \mathbb{H}^{\mathbf{K}}_{1}[\mathbf{q}],\, \mathbf{p} - \mathbf{q} \right\rangle$. **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.

Definition 2. **(GAIT Divergence)**♣ *The GAIT divergence between distributions $\mathbb{P}$ and $\mathbb{Q}$ on a finite similarity space $(\mathcal{X}, \kappa)$ is given by:*

$$
\mathbb{D}^{\mathbf{K}}[\mathbb{P} \,||\, \mathbb{Q}] \triangleq 1 + \mathbb{E}_{\mathbb{P}}\left[\log \frac{\mathbf{K}\mathbb{P}}{\mathbf{K}\mathbb{Q}}\right] - \mathbb{E}_{\mathbb{Q}}\left[\frac{\mathbf{K}\mathbb{P}}{\mathbf{K}\mathbb{Q}}\right]. \tag{4}
$$

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 $\mathbb{P} \notin \mathcal{Q}$. $\arg\min_{\mathbb{Q}} \mathbb{D}^{\mathbf{K}}[\cdot \,||\, \mathbb{P}]$ 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. 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. (2) to general measures reduce to finite sums over the corresponding supports.

Table 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	$\mathbb{H}^{\mathbf{K},\mathbf{\Lambda}}[X\|Y] \triangleq \mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda}}[X,Y] - \mathbb{H}^{\mathbf{\Lambda}}[Y]$
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.	$\mathbb{I}^{\mathbf{K},\mathbf{\Lambda},\mathbf{\Theta}}[X;Y\|Z] \triangleq \mathbb{H}^{\mathbf{K},\mathbf{\Theta}}[X\|Z] + \mathbb{H}^{\mathbf{\Lambda},\mathbf{\Theta}}[Y\|Z] - \mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda},\mathbf{\Theta}}[X,Y\|Z]$

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 $\mathcal{S}\triangleq \operatorname{supp}(\mathbb{P})\cup\operatorname{supp}(\mathbb{Q})$ 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

$$\mathbb{D}^{\mathbf{K}}[\hat{\mathbb{P}}\,||\,\hat{\mathbb{Q}}] = 1 + \left\langle \mathbf{p}, \log\frac{\mathbf{K}_{xx}\mathbf{p}}{\mathbf{K}_{xy}\mathbf{q}}\right\rangle - \left\langle \mathbf{q}, \frac{\mathbf{K}_{yx}\mathbf{p}}{\mathbf{K}_{yy}\mathbf{q}}\right\rangle. \quad (5)$$

Computational complexity. The computation of Eq. (5) 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. (2) and (5) 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. G.

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, we provide an example in which the GAIT divergence can provide a smooth training signal despite being evaluated on distribution with disjoint supports.

2.4 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 3. *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. 1.*

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 3. *Let X, Y be independent, then:*

$$\mathbb{H}^{\mathbf{K}\otimes\mathbf{\Lambda}}[X,Y] = \mathbb{H}^{\mathbf{K}}[X] + \mathbb{H}^{\mathbf{\Lambda}}[Y]. \quad (6)$$

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

Theorem 4. *For any kernel κ,*

$$\mathbb{H}^{\mathbf{K},\mathbf{I}}[X|Y] = \mathbb{E}_{y\sim\mathbb{P}_y}[\mathbb{H}^{\mathbf{K}}[X|Y=y]]. \quad (7)$$

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

Theorem 5. ♣ *For any similarity kernel κ,*

$$\mathbb{H}^{\mathbf{K},\mathbf{I}}[X|Y] \leq \mathbb{H}^{\mathbf{K}}[X]. \quad (8)$$

Theorem 5 is equivalent to Conj. 1 when considering a categorical Y mixing over distributions $\{X_y\}_{y\in\mathcal{Y}}$.

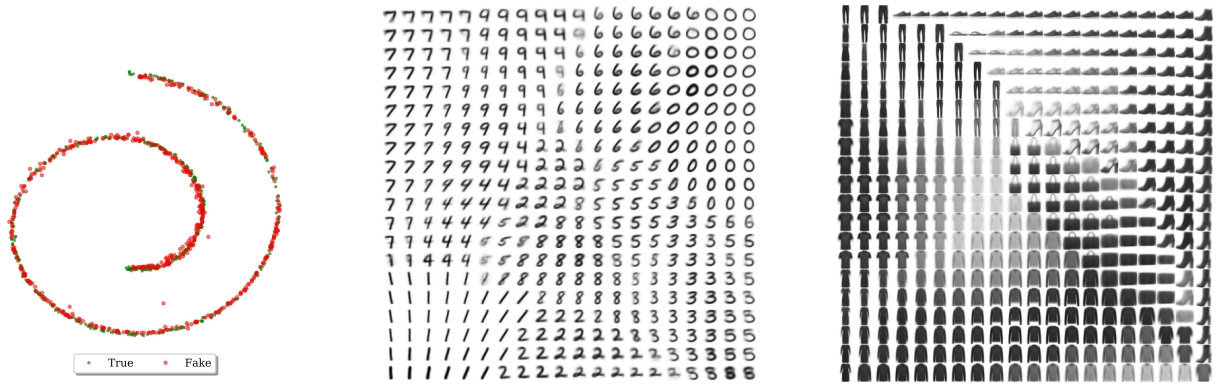

Figure 5: **Left:** Generated Swiss roll data. **Center and Right:** Manifolds for MNIST and Fashion MNIST.

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 6. (**Data Processing Inequality**)♣. *If $X \to Y \to Z$ is a Markov chain, then*

$$\mathbb{I}^{\mathbf{K},\boldsymbol{\Theta}}[X;Z] \leq \mathbb{I}^{\mathbf{K},\boldsymbol{\Lambda}}[X;Y] + \mathbb{I}^{\mathbf{K},\boldsymbol{\Theta},\boldsymbol{\Lambda}}[X;Z|Y]. \quad (9)$$

Note the presence of the additional term $\mathbb{I}^{\mathbf{K},\boldsymbol{\Lambda},\boldsymbol{\Theta}}[X;Z|Y]$ 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 $\boldsymbol{\Lambda} = \mathbf{I}$ this term is 0, and the original DPI is recovered.

3 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.

4 Experiments

4.1 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)

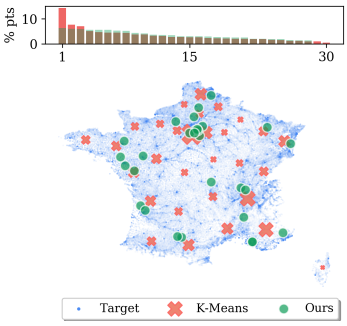

Figure 6: Approximating a discrete measure with a uniform empirical measure.

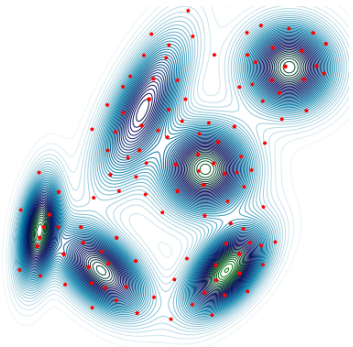

Figure 7: Approximating a continuous density with a finitely-supported measure.

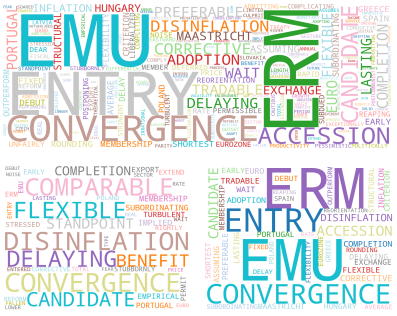

Figure 8: **Top:** Original word cloud. **Left:** Sparse approximation with support size 43. **Right:** Top 43 original TF-IDF words.

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).

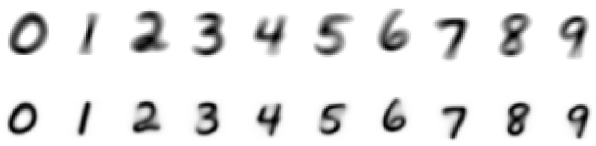

Figure 9: Barycenters for each class of MNIST with our divergence (top) and the method of Cuturi and Doucet (2014) (bottom).

Fig. 9 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. D.

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. 5 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. F, along with comparisons to variational auto-encoders (Kingma and Welling, 2014).

4.2 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 $\mathbf{q} \in \mathbf{\Delta}_m$ 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 gradient-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.

In Fig. 6, 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

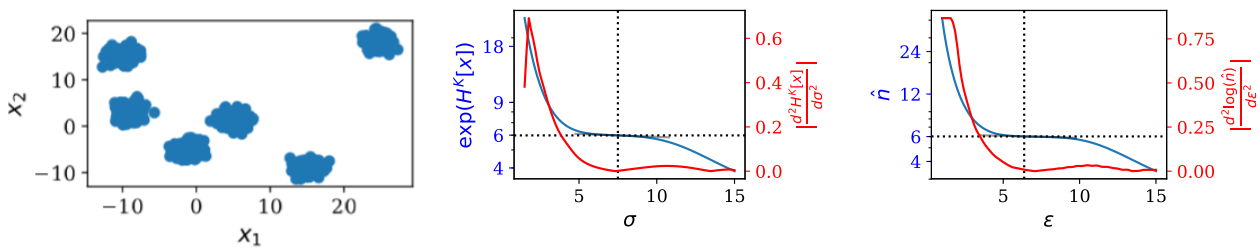

Figure 10: **Left:** 1,000 samples from a mixture of 6 Gaussians. **Center:** Modes detected by varying σ in our method. **Right:** Modes detected by varying collision threshold ϵ in the birthday paradox-based method.

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. 7 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. D 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. 8. 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. D.3.

4.3 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. 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 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. E.

5 Conclusions

In this paper, we advocate the use of geometry-aware information theory concepts in machine learning. We present the similarity-sensitive entropy of Leinster and Cobbold (2012) along with several important properties that connect it to fundamental notions in geometry. We then propose 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 propose a similarity-sensitive version of mutual information based on the GAIT entropy. We hope these methods can prove fruitful in extending frameworks such as the information bottleneck for representation learning (Tishby and Zaslavsky, 2015), 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).

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.

References

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

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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

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