Geometric Losses for Distributional Learning

Arthur Mensch [1 2] **Mathieu Blondel** [3] **Gabriel Peyré** [1 2]

Abstract

Building upon recent advances in entropy-regularized optimal transport, and upon Fenchel duality between measures and continuous functions, we propose a generalization of the logistic loss that incorporates a metric or cost between classes. Unlike previous attempts to use optimal transport distances for learning, our loss results in unconstrained convex objective functions, supports infinite (or very large) class spaces, and naturally defines a geometric generalization of the softmax operator. The geometric properties of this loss make it suitable for predicting sparse and singular distributions, for instance supported on curves or hyper-surfaces. We study the theoretical properties of our loss and showcase its effectiveness on two applications: ordinal regression and drawing generation.

1. Introduction

For probabilistic classification, the most popular loss is arguably the (multinomial) logistic loss. It is smooth, enabling fast convergence rates, and the softmax operator provides a consistent mapping to probability distributions. In many applications, different costs are associated to misclassification errors between classes. While a cost-aware generalization of the logistic loss exists (Gimpel & Smith, 2010), it does not provide a cost-aware counterpart of the softmax. The softmax is pointwise by nature: it is oblivious to misclassification costs or to the geometry of classes.

Optimal transport (Wasserstein) losses have recently gained popularity in machine learning, for their ability to compare probability distributions in a geometrically faithful manner, with applications such as classification (Kusner et al., 2015), clustering (Cuturi & Doucet, 2014), domain adaptation (Courty et al., 2017), dictionary learning (Rolet et al., 2016) and generative models training (Montavon et al., 2016; Arjovsky et al., 2017). For probabilistic classification, Frogner et al. (2015) proposes to use entropy-regularized optimal transport (Cuturi, 2013) in the multi-label setting. Although this approach successfully leverages a cost between classes, it results in a non-convex loss, when combined with a softmax. A similar regularized Wasserstein loss is used by Luise et al. (2018) in conjunction with a kernel ridge regression procedure (Ciliberto et al., 2016) in order to obtain a consistency result.

The relation between the logistic loss and the maximum entropy principle is well-known. Building upon a generalization of the Shannon entropy originating from entropy regularized optimal transport (Feydy et al., 2019) and Fenchel duality between measures and continuous functions, we propose a generalization of the logistic loss that takes into account a metric or cost between classes. Unlike previous attempts to use optimal transport distances for learning, our loss is convex, and naturally defines a geometric generalization of the softmax operator. Besides providing novel insights in the logistic loss, our loss is theoretically sound, even when learning and predicting *continuous* probability distributions over a potentially infinite number of classes. To sum up, our contributions are as follows.

Organization and contributions.

- We introduce the distribution learning setting, review existings losses leveraging a cost between classes and point out their shortcomings (§2).
- Building upon entropy-regularized optimal transport, we present a novel cost-sensitive distributional learning loss and its corresponding softmax operator. Our proposal is theoretically sound even in continuous measure spaces (§3).
- We study the theoretical properties of our loss, such as its Fisher consistency (§4). We derive tractable methods to compute and minimize it in the discrete distribution setting. We propose an abstract Frank-Wolfe scheme for computations in the continuous setting.
- Finally, we demonstrate its effectiveness on two discrete prediction tasks involving a geometric cost: ordinal regression and drawing generation using VAEs (§5).

[1]École Normale Supérieure, DMA, Paris, France [2]CNRS, France [3]NTT Communication Science Laboratories, Kyoto, Japan. Correspondence to: AM <arthur.mensch@m4x.org>, MB <mathieu@mblondel.org>, GP <gabriel.peyre@ens.fr>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Notation. We denote $\mathcal{X}$ a finite or infinite input space, and $\mathcal{Y}$ a compact potentially infinite output space. When $\mathcal{Y}$ is a finite set of d classes, we write $\mathcal{Y} = [d] \triangleq \{1, \dots, d\}$. We denote $\mathcal{C}(\mathcal{Y})$, $\mathcal{M}(\mathcal{Y})$, $\mathcal{M}^+(\mathcal{Y})$ and $\mathcal{M}_1^+(\mathcal{Y})$ the sets of continuous (bounded) functions, Radon (positive) measures and probability measures on $\mathcal{Y}$. Note that in finite dimensions, $\mathcal{M}_1^+([d]) = \triangle^d$ is the probability simplex and $\mathcal{C}([d]) = \mathbb{R}^d$. We write vectors in $\mathbb{R}^d$ and continuous functions in $\mathcal{C}(\mathcal{Y})$ with small normal letters, e.g., f, g. In the finite setting, where $\mathcal{Y} = \{y_1, \dots, y_d\}$, we define $f_i \triangleq f(y_i)$. We write elements of $\triangle^d$ and measures in $\mathcal{M}(\mathcal{Y})$ with greek letters α, β. We write matrices and operators with capital letters, e.g., C. We denote by $\otimes$ and $\oplus$ the tensor product and sum, and $\langle \cdot, \cdot \rangle$ the scalar product.

2. Background

In this section, after introducing distributional learning in a discrete setting, we review two lines of work for taking into account a cost C between classes: cost-augmented losses, and geometric losses based on Wasserstein and energy distances. Their shortcomings motivate the introduction of a new geometric loss in §3.

2.1. Discrete distribution prediction and learning

We consider a general predictive setting in which an input vector $x \in \mathcal{X}$ is fed to a parametrized model $g_\theta : \mathcal{X} \to \mathbb{R}^d$ (e.g., a neural network), that predicts a score vector $f = g_\theta(x) \in \mathbb{R}^d$. At test time, that vector is used to predict the most likely class $\hat{y} = \mathrm{argmax}_{y \in [d]} f_y$. In order to predict a probability distribution $\alpha \in \triangle^d$, it is common to compose g_θ with a link function $\psi(f)$, where $\psi \colon \mathbb{R}^d \to \triangle^d$. A typical example of link function is the softmax.

To learn the model parameters θ, it is necessary to define a loss $\ell(\alpha, f)$ between a ground-truth $\alpha \in \triangle^d$ and the score vector $f \in \mathbb{R}^d$. Composite losses (Reid & Williamson, 2010; Williamson et al., 2016) decompose that loss into a loss $\ell_\triangle(\alpha, \beta)$, where $\ell_\triangle \colon \triangle^d \times \triangle^d \to \mathbb{R}$ and ψ: $\ell(\alpha, f) \triangleq \ell_\triangle(\alpha, \psi(f))$. Note that depending on $\ell_\triangle$ and ψ, ℓ is not necessarily convex in f. More recently, Blondel et al. (2018; 2019) introduced Fenchel-Young losses, a generic way to directly construct a loss ℓ and a corresponding link ψ. We will revisit and generalize that framework to the continuous output setting in the sequel of this paper. Given a loss ℓ and a training set of input-distribution pairs, (x_i, α_i), where $x_i \in \mathcal{X}$ and $\alpha_i \in \triangle^d$, we then minimize $\sum_i \ell(\alpha_i, g_\theta(x_i))$, potentially with regularization on θ.

2.2. Cost-augmented losses

Before introducing a new geometric cost-sensitive loss in §3, let us now review classical existing cost-sensitive loss functions. Let C be a $d \times d$ matrix, such that $c_{y,y'} \geq 0$ is the cost of misclassifying class $y \in [d]$ as class $y' \in [d]$. We assume $c_{y,y} = 0$ for all $y \in [d]$. To take into account the cost C, in the single label setting, it is natural to define a loss $L \colon [d] \times \mathbb{R}^d \to \mathbb{R}$ as follows

$$L(y, f) = c_{y,y'} \quad \text{where} \quad y' \in \operatorname*{argmax}_{i \in [d]} f_i. \tag{1}$$

To obtain a loss $\ell \colon \triangle^d \times \mathbb{R}^d \to \mathbb{R}$, we simply define $\ell(\delta_y, f) \triangleq L(y, f)$, where δ_y is the one-hot representation of $y \in [d]$. Note that choosing $c_{y,y'} = 1$ when $y \neq y'$ and $c_{y,y'} = 0$ otherwise (i.e., $C = 1 - I_{d \times d}$) reduces to the zero-one loss. To obtain a convex upper-bound, (1) is typically replaced with a cost-augmented hinge loss (Crammer & Singer, 2001; Tsochantaridis et al., 2005):

$$L(y, f) = \max_{i \in [d]} c_{y,i} + f_i - f_y.$$

Replacing the max above with a log-sum-exp leads to a cost-augmented version of the logistic (or conditional random field) loss (Gimpel & Smith, 2010). Another convex relaxation is the cost-sensitive *pairwise* hinge loss (Weston & Watkins, 1999; Duchi et al., 2018). Remarkably, all these losses use only one row of C, the one corresponding to the ground truth y. Because of this dependency on y, it is not clear how to define a probabilistic mapping at *test* time. In this paper, we propose a loss which comes with a geometric generalization of the softmax operator. That operator uses the entire cost matrix C.

2.3. Wasserstein and energy distance losses

Wasserstein or optimal transport distances recently gained popularity as a loss in machine learning for their ability to compare probability distributions in a geometrically faithful manner. As a representative application, Frogner et al. (2015) proposed to use entropy-regularized optimal transport (Cuturi, 2013) for cost-sensitive multi-label classification. Effectively, optimal transport lifts a distance or cost $C \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ to a distance between probability distributions over $\mathcal{Y}$. Following Genevay et al. (2016), given a ground-truth probability distribution $\alpha \in \triangle^d$ and a predicted probability distribution $\beta \in \triangle^d$, we define

$$\mathrm{OT}_{C,\epsilon}(\alpha, \beta) \triangleq \min_{\pi \in \mathcal{U}(\alpha, \beta)} \langle \pi, C \rangle + \varepsilon \mathrm{KL}(\pi | \alpha \otimes \beta), \tag{2}$$

where $\mathcal{U}$ is the transportation polytope, a subset of $\triangle^{d \times d}$ whose elements π have constrained marginals: $\pi 1 = \alpha$ and $\pi^\top 1 = \beta$. KL is the Kullback–Leibler divergence (a.k.a. relative entropy). Because β needs to be a valid probability distribution, Frogner et al. (2015) propose to use $\beta = \psi(f) = \mathrm{softmax}(f)$, where $f \in \mathbb{R}^d$ is a vector of prediction scores. Unfortunately, the resulting composite loss, $\ell(\alpha, f) = \mathrm{OT}_{C,\epsilon}(\alpha, \mathrm{softmax}(f))$, is not convex w.r.t. f. Another class of divergences between measures α and β

stems from energy distances (Székely & Rizzo, 2013) and maximum mean discrepancies. However, composing these divergences with a softmax again breaks convexity in f. In contrast, our proposal is convex in f and defines a natural geometric softmax.

3. Continuous and cost-sensitive distributional learning and prediction

In this section, we construct a loss between probability measures and score functions, canonically associated with a link function. Our construction takes into account a cost function $C : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ between classes. Unlike existing methods reviewed in §2.2, our loss is well defined and convex on compact, possibly infinite spaces $\mathcal{Y}$. We start by extending the setting of §2.1 to predicting *arbitrary probabilities*, for instance having continuous densities with respect to the Lebesgue measure or singular distributions supported on curves or surfaces.

3.1. Continuous probabilities and score functions

We consider a compact metric space of outputs $\mathcal{Y}$, endowed with a *symmetric* cost function $C : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. We wish to predict probabilities over $\mathcal{Y}$, that is, learn to predict distributions $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$. The space of probability measures forms a closed subset of the space of Radon measures $\mathcal{M}(\mathcal{Y})$, i.e., $\mathcal{M}_1^+(\mathcal{Y}) \subseteq \mathcal{M}(\mathcal{Y})$. From the Riesz representation theorem, $\mathcal{M}(\mathcal{Y})$ is the topological dual of the space of continuous measures $\mathcal{C}(\mathcal{Y})$, endowed with the uniform convergence norm $\|\cdot\|_\infty$. The topological duality between the primal $\mathcal{M}(\mathcal{Y})$ and the dual $\mathcal{C}(\mathcal{Y})$ defines a pairing, similar to a "scalar product", between these spaces:

$$\langle \alpha, f \rangle \triangleq \int_{\mathcal{Y}} f(y) \mathrm{d}\alpha(y) = \mathbb{E}[f(Y)],$$

for all $\alpha \in \mathcal{M}(\mathcal{Y})$ and $f \in \mathcal{C}(\mathcal{Y})$, where Y is a random variable with law α. This pairing also defines the natural topology to compare measures and to differentiate functionals depending on measures. This is the so-called weak* topology, which corresponds to the convergence in law of random variables. A sequence α_n is said to converge weak* to some α if for all functions $f \in \mathcal{C}(\mathcal{Y})$, $\langle \alpha_n, f \rangle \to \langle \alpha, f \rangle$. Note that when endowing $\mathcal{M}(\mathcal{Y})$ with this weak* topology, the dual of $\mathcal{M}(\mathcal{Y})$ is $\mathcal{C}(\mathcal{Y})$, which is the key to be able to use duality (and in particular Legendre-Fenchel transform) from convex optimization. Using this topology is fundamental to define geometric losses that can cope with arbitrary, possibly highly localized or even singular distributions (for instance sparse sums of Diracs or measures concentrated on thin sets such as 2-D curves or 3-D surfaces).

Similarly to the discrete setting reviewed in §2.1, in the continuous setting, we now wish to predict a distribution $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$ by setting $\alpha = \psi(f)$, where $f = g_\theta(x) \in \mathcal{C}(\mathcal{Y})$, $g_\theta \colon \mathcal{X} \to \mathcal{C}(\mathcal{Y})$ (i.e., g_θ is unconstrained), and $\psi \colon \mathcal{C}(\mathcal{Y}) \to \mathcal{M}_1^+(\mathcal{Y})$ is a link function. We propose to use maps between the primal $\mathcal{M}_1^+(\mathcal{Y})$ and the *dual* score space $\mathcal{C}(\mathcal{Y})$ as link functions. As we shall see, such *mirror* maps are naturally defined by continuous convex function on the primal space, through Fenchel-Legendre duality. Our framework recovers the discrete case $\mathcal{Y} = [d]$ as a particular case, with $\triangle^d$ corresponding to $\mathcal{M}_1^+([d])$ and $\mathbb{R}^d$ to $\mathcal{C}([d])$, though the isomorphisms $\alpha \to \sum_{i=1}^d \alpha_i \delta_i$ and for all $i \in [d]$, $f(i) = f_i$.

Regularization of optimal transport is our key tool to construct entropy functions which are continuous with respect to the weak* topology, and that can be conjugated to define a $\mathcal{C}(\mathcal{Y}) \to \mathcal{M}_1^+(\mathcal{Y})$ link function. It allows us to naturally leverage a cost $C : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ between classes.

3.2. An entropy function for continuous probabilities

The regularized optimal transport cost (2) remains well defined when α and β belong to a continuous measure space $\mathcal{M}_1^+(\mathcal{Y})$, with $\mathcal{U}$ now being a subset of $\mathcal{M}_1^+(\mathcal{Y} \times \mathcal{Y})$ with marginal constraints. It induces the self-transport functional (Feydy et al., 2019), that we reuse for our purpose:

$$\Omega_C(\alpha) \triangleq \begin{cases} -\frac{1}{2}\mathrm{OT}_{C,\varepsilon=2}(\alpha, \alpha) & \text{for } \alpha \in \mathcal{M}_1^+(\mathcal{Y}) \\ +\infty & \text{otherwise.} \end{cases} \quad (3)$$

We will omit the dependency of Ω on C when clear from context. It is shown by Feydy et al. (2019) that Ω is continuous and convex on $\mathcal{M}(\mathcal{Y})$, and strictly convex on $\mathcal{M}_1^+(\mathcal{Y})$, where continuity is taken w.r.t. the weak* topology. We call Ω, the *Sinkhorn negentropy*. As a negative entropy function, it can be used to measure the uncertainty in a probability distribution (lower is more uncertain), as illustrated in Figure 2. It will prove crucial in our loss construction. In the above, we have set w.l.o.g. $\varepsilon = 2$ to recover simple asymptotical behavior of Ω, as will be clear in Prop. 1.

We first recall some known results from Feydy et al. (2019). Using Fenchel-Rockafellar duality theorem (Rockafellar, 1966), the function Ω rewrites as the solution to a Kantorovich-type dual problem (see e.g., Villani, 2008). For all $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$, we have that

$$-\Omega_C(\alpha) = \max_{f \in \mathcal{C}(\mathcal{Y})} \langle \alpha, f \rangle - \log \langle \alpha \otimes \alpha, e^{\frac{f \oplus f - C}{2}} \rangle, \quad (4)$$

where we use the *homogeneous* dual (i.e. with a log in the maximization), as explained in Cuturi & Peyré (2018).

Gradient and extrapolation. Ω is differentiable in the sense of measures (Santambrogio, 2015), meaning that there exists a continuous function $\nabla\Omega(\alpha)$ such that, for all

$\xi_1, \xi_2 \in \mathcal{M}_1^+(\mathcal{Y})$, $t > 0$,

$$\Omega(\alpha + t(\xi_2 - \xi_1)) = \Omega(\alpha) + t\langle \xi_2 - \xi_1, \nabla\Omega(\alpha)\rangle + o(t). \quad (5)$$

As shown in Feydy et al. (2019), this function $f = \nabla\Omega(\alpha)$, that we call the *symmetric Sinkhorn potential*, is a particular solution of the dual problem. It is the only function in $\mathcal{C}(\mathcal{Y})$ such that $-f = T(-f, \alpha)$, where the soft C-transform operator (Cuturi & Peyré, 2018) is defined as

$$T(f, \alpha)(y) \triangleq -2 \log\langle \alpha, e^{\frac{f - C(y,\cdot)}{2}}\rangle.$$

This operator can be understood as the log-convolution of the measure $\alpha e^{\frac{f}{2}}$ with the Sinkhorn kernel $e^{-\frac{C}{2}}$. The Sinkhorn potential f has the remarkable property of being defined on all $\mathcal{Y}$, even though the support of α may be smaller. Given any dual solution g to (4), which is defined α-almost everywhere, we have $f = -T(-g, \alpha)$, i.e. f *extrapolates* the values of g on the support of α, using the Sinkhorn kernel.

Special cases. The following proposition, which is an original contribution, shows that the Sinkhorn negentropy asymptotically recovers the negative Shannon entropy and Gini index (Gini, 1912) when rescaling the cost. The Sinkhorn negentropy therefore defines a parametric family of negentropies, recovering these important special cases. Note however that on continuous spaces $\mathcal{Y}$, the Shannon entropy is not weak* continuous and thus cannot be used to define geometric loss and link functions, the softmax link function being geometry-oblivious. Similarly, the Gini index is not defined on $\mathcal{M}_1^+(\mathcal{Y})$, as it involves the *squared values* of α in a discrete setting.

Proposition 1 (Asymptotics of Sinkhorn negentropies)**.** *For $\mathcal{Y}$ compact, the rescaled Sinkhorn negentropy converges to a kernel norm for high regularization ε. Namely, for all $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$, we have*

$$\varepsilon\Omega_{C/\varepsilon}(\alpha) \overset{\varepsilon \to +\infty}{\longrightarrow} \frac{1}{2}\langle \alpha \otimes \alpha, -C\rangle.$$

Let $\mathcal{Y} = [d]$ be discrete and choose $C = 1 - I_{d\times d}$. The Sinkhorn negentropy converges to the Shannon negentropy for low-regularization, and into the negative Gini index for high regularization:

$$\Omega_{C/\varepsilon}(\alpha) \overset{\varepsilon \to 0}{\longrightarrow} \langle \alpha, \log \alpha\rangle,\ \varepsilon\Omega_{C/\varepsilon}(\alpha) \overset{\varepsilon \to +\infty}{\longrightarrow} \frac{1}{2}(\|\alpha\|_2^2 - 1).$$

Proof is provided in §A.1. The first part of the proposition shows that the Sinkhorn negentropies converge to a kernel norm (see e.g., Sriperumbudur et al., 2011). This is similar to the regularized Sinkhorn divergences converging to an Energy Distance (Székely & Rizzo, 2013) for $\varepsilon \to \infty$ (Genevay et al., 2018; Feydy et al., 2019).

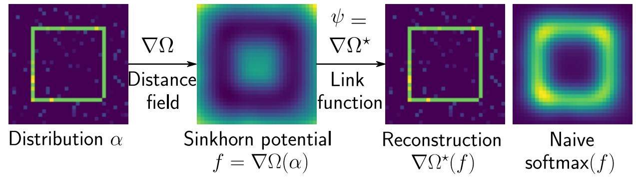

Figure 1. The symmetric Sinkhorn potentials form a distance field to a weighted measure. The link function $\psi = \nabla\Omega^\star(f)$ allows to go back from this field in $\mathcal{C}(\mathcal{Y})$ to a measure $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$.

From probabilities to potentials. The symmetric Sinkhorn potential $f = \nabla\Omega(\alpha)$ is a continuous function, or a vector in the discrete setting. It can be interpreted as a distance field to the distribution α. We visualize this field on a 2D space in Figure 1, where $\mathcal{Y}$ is the set of $h \times w$ pixels of an image, and we wish to predict a 2-dimensional probability distribution in $\mathcal{M}_1^+(\mathcal{Y}) = \triangle^{h\times w}$. Predicting a distance field $f \in \mathcal{C}(\mathcal{Y})$ to a measure is more convenient than predicting a distribution directly, as it has unconstrained values and is therefore easier to optimize against. For this reason, we propose to learn parametric models that predict a "distance field" $f = g_\theta(x)$ given an input $x \in \mathcal{X}$. In the following section, we construct a link function $\psi : \mathcal{C}(\mathcal{Y}) \to \mathcal{M}_1^+(\mathcal{Y})$, for general probability measure and function spaces $\mathcal{M}_1^+(\mathcal{Y})$ and $\mathcal{C}(\mathcal{Y})$, so to obtain a distributional estimator $\alpha_\theta = \psi \circ g_\theta : \mathcal{X} \to \mathcal{M}_1^+(\mathcal{Y})$.

3.3. Fenchel-Young losses in continuous setting

To that end, we generalize in this section the recently-proposed Fenchel-Young (FY) loss framework (Blondel et al., 2018; 2019), originally limited to discrete cost-oblivious measure spaces, to infinite measure spaces. Inspired by that line of work, we use Legendre-Fenchel duality to define loss and link functions from Sinkhorn negative entropies, in a principled manner. We define the Legendre-Fenchel conjugate $\Omega^\star : \mathcal{C}(\mathcal{Y}) \to \mathbb{R}$ of Ω as

$$\Omega^\star(f) \triangleq \max_{\alpha \in \mathcal{M}_1^+(\mathcal{Y})} \langle \alpha, f\rangle - \Omega(\alpha).$$

Rigorously, $\Omega^\star(f)$ is a pre-conjugate, as Ω is defined on $\mathcal{M}(\mathcal{Y})$, the topological dual of continuous functions $\mathcal{C}(\mathcal{Y})$. For a comprehensive and rigorous treatment of the theory of conjugation in infinite spaces, and in particular Banach spaces as is the case of $\mathcal{C}(\mathcal{Y})$, see Mitter (2008).

As Ω is strictly convex, $\Omega^\star$ is differentiable everywhere and we have, from a Danskin theorem (Danskin, 1966) with left Banach space and right compact space (Bernhard & Rapaport, 1995, Theorem C.1):

$$\nabla\Omega^\star(f) \triangleq \underset{\alpha \in \mathcal{M}_1^+(\mathcal{Y})}{\operatorname{argmax}} \langle \alpha, f\rangle - \Omega(\alpha) \in \mathcal{C}(\mathcal{Y}).$$

That gradient can be used as a link ψ from $f \in \mathcal{C}(\mathcal{Y})$ to $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$. It can also be interpreted as a regularized prediction function (Blondel et al., 2018; Mensch & Blondel, 2018). Following the FY loss framework, we define the loss associated with $\nabla\Omega^\star$ by

$$\ell_\Omega(\alpha, f) \triangleq \Omega^\star(f) + \Omega(\alpha) - \langle \alpha, f \rangle. \tag{6}$$

In the discrete single-label setting, that loss is also related to the construction of Duchi et al. (2018, Proposition 3). From the Fenchel-Young theorem (Rockafellar, 1970), $\ell_\Omega(\alpha, f) \geq 0$, with equality if and only if $\alpha = \nabla\Omega^\star(f)$. The loss ℓ_Ω is thus positive, convex and differentiable in its second argument, and minimizing it amounts to find the pre-image $f^\star$ of the target distribution α with respect to the link (mapping) $\nabla\Omega^\star$.

Our construction is a generalization of the Fenchel-Young loss framework (Blondel et al., 2018; 2019), in the sense that it relies on topological duality between $\mathcal{C}(\mathcal{Y})$ and $\mathcal{M}_1^+(\mathcal{Y})$, instead of the Hilbertian structure of $\mathbb{R}^d$ and $\triangle^d$, to construct the loss ℓ_Ω and link function $\nabla\Omega^\star$. We now instantiate the Fenchel-Young loss (6) with Sinkhorn negentropies in order to obtain a novel cost-sensitive loss.

3.4. A new geometrical loss and softmax

The key ingredients to derive a Fenchel-Young loss ℓ_Ω and a link $\nabla\Omega^\star$ are the conjugate $\Omega^\star$ and its gradient. Remarkably, they enjoy a simple form with Sinkhorn negentropies, as shown in the following proposition.

Proposition 2 (Conjugate of the Sinkhorn negentropy)**.** *For all $f \in \mathcal{C}(\mathcal{Y})$, the Legendre-Fenchel conjugate $\Omega^\star$ of Ω defined in* (3) *and its gradient read*

$$\text{g-LSE}(f) \triangleq \Omega^\star(f) = -\log \min_{\alpha \in \mathcal{M}_1^+(\mathcal{Y})} \Phi(\alpha, f)$$

$$\text{g-softmax}(f) \triangleq \nabla\Omega^\star(f) = \operatorname*{argmin}_{\alpha \in \mathcal{M}_1^+(\mathcal{Y})} \Phi(\alpha, f)$$

$$\text{where} \quad \Phi(\alpha, f) \triangleq \langle \alpha \otimes \alpha, \exp(-\frac{f \oplus f + C}{2}) \rangle$$

and where g stands for geometric and LSE for log-sum-exp.

The proof can be found in §A.2. $\nabla\Omega^\star(f)$ is the usual Fréchet derivative of $\Omega^\star$, that lies a priori in the topological dual space of $\mathcal{C}(\mathcal{Y})$, i.e. $\mathcal{M}(\mathcal{Y})$. From a Danskin theorem (Bernhard & Rapaport, 1995), it is in fact a probability measure. The probability distribution $\alpha = \nabla\Omega^\star(f)$ is typically *sparse*, as the minimizer of a quadratic on a convex subspace of $\mathcal{M}(\mathcal{Y})$. We call the loss ℓ_Ω generated by the Sinkhorn negentropy **g-logistic** loss.

Special cases. Let $\mathcal{Y} = [d]$ and $C = 1 - I_{d\times d}$ (0-1 cost matrix). From Prop. 1, Ω asymptotically recovers the negative Shannon entropy when $\Omega = \Omega_{\frac{C}{\varepsilon}}$ as $\varepsilon \to 0$ and the negative Gini index when $\Omega = \varepsilon\Omega_{\frac{C}{\varepsilon}}$, as $\varepsilon \to \infty$. $\nabla\Omega^\star$ is then equal to $\text{softmax}(f) \triangleq \frac{\exp f}{\sum_i \exp f_i}$, and to $\text{sparsemax}(f) \triangleq \operatorname{argmin}_{\alpha \in \triangle^d} \|\alpha - f\|^2$ (Martins & Astudillo, 2016), respectively. Likewise, ℓ_Ω recovers the logistic and sparsemax losses. When $\varepsilon \to 0$, because $(\frac{C}{\varepsilon})_{y,y'} = \infty$ if $y \neq y'$ and 0 otherwise, we see that the logistic loss infinitely penalizes inter-class errors. That is, to obtain zero logistic loss, the model must assign probability 1 to the correct class. The limit case $\varepsilon \to 0$ is the only case for which g-softmax always outputs completely dense distributions. In the continuous case, $\varepsilon\Omega^\star_{C/\varepsilon}(f/\varepsilon)$ degenerates into a positive deconvolution objective with simplex constraint:

$$\max_{\alpha \in \mathcal{M}_1^+(\mathcal{Y})} \langle \alpha, f \rangle - \frac{1}{2}\langle \alpha \otimes \alpha, -C \rangle.$$

Fig. 1 shows that $\nabla\Omega^\star$ has indeed a deconvolutional effect.

3.5. Computation

Before studying the g-logistic loss ℓ_Ω and link function $\nabla\Omega^\star(f)$, we now describe practical algorithms for computing $\nabla\Omega^\star(f)$ and $\Omega^\star(f)$ in the discrete and continuous cases. The key element in using the g-LSE as a layer in an arbitrary complex model is to minimize the quadratic function $\Phi_f \triangleq \Phi(\cdot, f)$, on $\mathcal{M}_1^+(\mathcal{Y})$. We can then use the minimum value in the forward pass, and the minimizer in the backward pass, during e.g. SGD training.

Continuous optimisation. In the general case where $\mathcal{Y}$ is compact, we cannot represent $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$ using a finite vector. Yet, we can use a Frank-Wolfe scheme to progressively add spikes, i.e. Diracs to an iterate sequence α_t. For this, we need to compute, at each iteration, the gradient of Φ_f in the sense of measure (5), i.e. the function in $\mathcal{C}(\mathcal{Y})$

$$\nabla\Phi_f(\alpha) = \exp(-\frac{f + T(-f, \alpha)}{2}),$$

that simply requires to compute the C-transform of $-f$ on the measure α, similarly to regularized optimal transport. The simplest Frank-Wolfe scheme then updates

$$y_t \in \operatorname*{argmin}_{\mathcal{Y}} \nabla\Phi_f(\alpha_{t-1}), \alpha_t = \alpha_{t-1} + \frac{2}{t+2}(\delta_{y_t} - \alpha_{t-1}).$$

Indeed, for $h \in \mathcal{C}(\mathcal{Y})$, the minimizer of $\langle h, \cdot \rangle$ on $\mathcal{M}_1^+(\mathcal{Y})$ is the Dirac δ_y where $y \in \mathcal{Y}$ minimizes h. This optimization scheme may be refined to ensure a geometric convergence of $\Phi_f(\alpha_t)$. It can be used to identify Diracs from a continuous distance field f, similar to super-resolution approaches proposed in Bredies & Pikkarainen (2013); Boyd et al. (2017). It requires to work with computer-friendly representation of f, so that we can obtain an approximation of y_t efficiently, using e.g. non-convex optimization. Another approach is to rely on a deep parametrization of a particle swarm, as proposed by Boyd et al. (2018). We leave such an application for future work, and focus on an efficient *discrete* solver for the g-LSE and g-softmax.

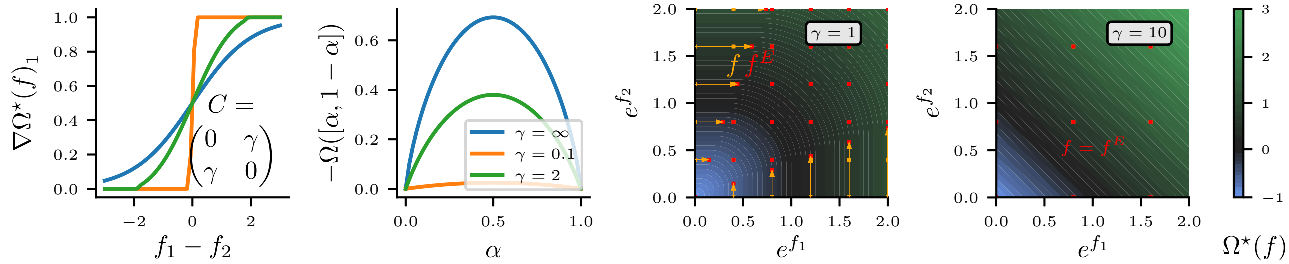

Figure 2. Left: Geometric softmax and Sinkhorn entropy, for symmetric cost matrices, in the binary case. Predictions from the g-softmax are sparse, as the minimizer of a convex quadratic on the simplex. Right: Level sets of the geometric conjugate. Introducing a cost matrix induces a deformation $\triangle^2$, the level-set of the log-sum-exp operator, onto the set of symmetric Sinkhorn potentials $\mathcal{F}$. The geometric conjugate defines an extrapolation operator $f \to f^E$ that replaces the score function onto the cylinder $\mathcal{F} + \mathbb{R}1$.

Discrete optimisation. In the discrete case, we can parametrize $\log \Phi(f, \cdot)$ in logarithmic space, by setting $\alpha = \exp(l) - \text{LSE}(l)$, with $l \in \mathbb{R}^d$. $\Omega^\star(f)$ then reads

$$\max_{l \in \mathbb{R}^d} - \log \sum_{i,j=1}^{d} e^{l_i + l_j - \frac{f_i + f_j + c_{i,j}}{2}} + 2\text{LSE}(l). \qquad (7)$$

This objective is non-convex on $\mathbb{R}^d$ but invariant with translation and convex on $\{h \in \mathbb{R}^d, \text{LSE}(l) = 0\}$. It thus admits a unique solution, that we can find using an unconstrained quasi-Newton solver like L-BFGS (Liu & Nocedal, 1989), that we stop when the iterates are sufficiently stable. For l that maximizes (7), the gradient $\nabla\Omega^\star(f) = \text{softmax}(l)$ is used for backpropagation and at test time. As $\nabla\Omega^\star(f)$ is sparse, we expect some coordinates l_i to go to $-\infty$. In practice, α_i then underflows to 0 after a few iterations.

Two-dimensional convolution. In the discrete case, when dealing with two-dimensional potentials and measures, the objective function (7) can be written with a convolution operator, as $-\log\langle e^{l - \frac{f}{2}}, e^{-\frac{C}{2}} \star e^{l - \frac{f}{2}} \rangle$ where $C \in \mathbb{R}^{(h \times w)^2}$. It is therefore efficiently computable and differentiable on GPUs, especially when the kernel C is separable in height and width, e.g. for the ℓ_2^2 norm, in which case we perform 2 successive one-dimensional convolutions. We use this computational trick in our variational auto-encoder experiments (§5).

4. Geometric and statistical properties

We start by studying the mirror map $\nabla\Omega^\star$, that we expect to invert the mapping $\alpha \to \nabla\Omega(\alpha)$. This study is necessary as we cannot rely on typical conjugate inversion results (e.g., Rockafellar, 1970, Theorem 26.5), that would stipulate that $\nabla\Omega^\star = (\nabla\Omega)^{-1}$ on the domain of $\Omega^\star$. Indeed, this result is stated in finite dimension, and requires that Ω and $\Omega^\star$ be Legendre, i.e. be strictly convex and differentiable on their domain of definition, and have diverging derivative on the boundaries of these domains (see also Wainwright & Jordan, 2008). This is not the case of the Sinkhorn negentropy, which requires novel adjustements. With these at hands, we show that parametric models involving a final g-softmax layer can be trained to minimize a certain well-behaved Bregman divergence on the space of probability measures. Proofs are reported in §A.3 and §A.4

4.1. Geometry of the link function

We have constructed the link function $\nabla\Omega^\star$ in hope that it would allow to go from a symmetric Sinkhorn potential $f = \nabla\Omega(\alpha)$ back to the original measure α. The following lemma states that this is indeed the case, and derives two consequences on the space of symmetric Sinkhorn potentials, defined as $\mathcal{F} \triangleq \{f \in \mathcal{C}(\mathcal{Y}), f = \nabla\Omega(\alpha)\}$.

Lemma 1 (Inversion of the Sinkhorn potentials)**.**

$$\forall\, \alpha \in \mathcal{M}_1^+(\mathcal{Y}),\ \nabla\Omega^\star \circ \nabla\Omega(\alpha) = \alpha.$$
$$\forall\, f \in \mathcal{F}, \quad \nabla\Omega \circ \nabla\Omega^\star(f) = f, \quad \Omega^\star(f) = 0.$$

The computation of the Sinkhorn potential thus inverts the g-LSE operator on the space $\mathcal{F}$, which is included in the 0-level set of $\Omega^\star$. This is similar to the set $\mathcal{F}_{\text{Shannon}} = \{\log\alpha, \alpha \in \triangle^d\}$ being the 0 level set of the log-sum-exp function when using the Shannon negentropy as Ω.

This corollary is not sufficient for our purpose, as we want to characterize the action of $\nabla\Omega^\star$ on *all continuous functions* $f \in \mathcal{C}(\mathcal{Y})$. For this, note that the g-LSE operator $\Omega^\star$ has the same behavior as the log-sum-exp when composed with the addition of a constant $c \in \mathbb{R}$:

$$\Omega^\star(f + c) = \Omega^\star(f) + c, \quad \nabla\Omega^\star(f + c) = \nabla\Omega^\star(f). \quad (8)$$

Therefore, for all $f \in \mathcal{C}(\mathcal{Y})$, $\Omega^\star(f - \Omega^\star(f)) = 0$, which almost makes $f - \Omega^\star(f)$ a part of the space of potentials $\mathcal{F}$. Yet, in contrast with the Shannon entropy case, the inclusion of $(\Omega^\star)^{-1}(0)$ in $\mathcal{F}$ is strict. Indeed, following §3.2 $f \in \mathcal{F}$ implies that there exists $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$ such that $f = -T(-f, \alpha)$ is the image of the C-transform operator. The operator $\nabla\Omega \circ \nabla\Omega^\star$ has therefore an *extrapolation* effect, as it replaces $f - \Omega^\star(f)$ onto the set of Sinhorn potentials. This is made clear by the following proposition.

Proposition 3 (Extrapolation effect of $\nabla\Omega \circ \nabla\Omega^\star$). *For all $f \in \mathcal{C}(\mathcal{Y})$, we define the extrapolation of f to be*

$$f^E \triangleq -T\big(-(f - \Omega^\star(f)), \nabla\Omega^\star(f)\big) + \Omega^\star(f).$$

Then, for all $f \in \mathcal{C}(\mathcal{Y})$, $\nabla\Omega \circ \nabla\Omega^\star(f) = f^E - \Omega^\star(f)$.

The extrapolation operator translates f to $(\Omega^\star)^{-1}(0)$, extrapolates $f - \Omega^\star(f)$ so that it becomes a Sinkhorn potential, then translates back the result so that $\Omega^\star(f^E) = \Omega^\star(f)$. Its effects clearly appears on Figure 2 (right), where we see that f^E is a projection of f onto the cylinder $\mathcal{F} + \mathbb{R}$.

4.2. Relation to Hausdorff divergence

Recall that the Bregman divergence (Bregman, 1967) generated by a strictly convex Ω is defined as

$$D_\Omega(\alpha, \beta) \triangleq \Omega(\alpha) - \Omega(\beta) - \langle \nabla\Omega(f), \alpha - \beta \rangle.$$

When Ω is the classical negative Shannon entropy $\Omega(\alpha) = \langle \alpha, \log \alpha \rangle$, it is well-known that D_Ω equals the Kullback-Leibler divergence and it is easy to check that

$$\ell_\Omega(\alpha, f) = D_\Omega(\alpha, \nabla\Omega^\star(f)) = \mathrm{KL}(\alpha, \mathrm{softmax}(f)).$$

The equivalence between Fenchel-Young loss $\ell_\Omega(\alpha, f)$ and composite Bregman divergence $D_\Omega(\alpha, \nabla\Omega^\star(f))$, however, no longer holds true when Ω is the Sinkhorn negentropy defined in (3). In that case, D_Ω can be interpreted as an *asymmetric* Hausdorff divergence (Aspert et al., 2002; Feydy et al., 2019). It forms a geometric divergence akin to OT distances, and estimates the distance between distribution supports. As we now show, ℓ_Ω provides an upper-bound on the composition of that divergence with $\nabla\Omega^\star$.

Proposition 4 (ℓ_Ω upper-bounds Hausdorff divergence).

$$\begin{aligned} D_\Omega(\alpha, \nabla\Omega^\star(f)) &= \ell_\Omega(\alpha, f^E) \\ &= \ell_\Omega(\alpha, f) - \langle \alpha, f^E - f \rangle \leq \ell_\Omega(\alpha, f) \end{aligned}$$

with equality if $\mathrm{supp}\, \nabla\Omega^\star(f) = \mathrm{supp}\, \alpha$.

In contrast with the KL divergence, the asymmetric Hausdorff divergence is finite even when $\mathrm{supp}\, \alpha \neq \mathrm{supp}\, \beta$, a geometrical property that it shares with optimal transport divergences. We now use Prop. 4 to derive a new consistency result justifying our loss. Let us assume that input features and output distributions follow a distribution $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{M}_1^+(\mathcal{Y}))$. We define the Hausdorff divergence risk and the Fenchel-Young loss risk as

$$\mathcal{E}(\beta) \triangleq \mathbb{E}[D_\Omega(\alpha, \beta(x))] \quad \text{and} \quad \mathcal{R}(g) \triangleq \mathbb{E}[\ell_\Omega(\alpha, g(x))],$$

where the expectation is taken w.r.t. $(x, \alpha) \sim \mathcal{D}$. We define their associated Bayes estimators as

$$\beta^\star \triangleq \operatorname*{argmin}_{\beta :\, \mathcal{X} \to \mathcal{M}_1^+(\mathcal{Y})} \mathcal{E}(\beta) \quad \text{and} \quad g^\star \triangleq \operatorname*{argmin}_{g :\, \mathcal{X} \to \mathcal{C}(\mathcal{Y})} \mathcal{R}(g).$$

The next proposition guarantees calibration of ℓ_Ω with respect to the asymmetric Hausdorff divergence D_Ω.

Proposition 5 (Calibration of the g-logistic loss). *The g-logistic loss ℓ_Ω where Ω is defined in (3) is Fisher consistent with the Hausdorff divergence D_Ω for the same Ω. That is,*

$$\mathcal{E}(\beta^\star) = \mathcal{R}(g^\star), \quad \textit{with} \quad g^\star = \nabla\Omega \circ \beta^\star.$$

The excess of risk in the Hausdorff divergence is controlled by the excess of risk in the g-logistic loss. For all $g : \mathcal{X} \to \mathcal{C}(\mathcal{Y})$, we have

$$\mathcal{E}(\nabla\Omega^\star \circ g) - \mathcal{E}(\beta^\star) \leq \mathcal{R}(g) - \mathcal{R}(g^\star).$$

This result, that follows the terminology of Tewari & Bartlett (2005), shows that ℓ_Ω is suitable for learning predictors that minimize D_Ω.

5. Applications

We present two experiments that demonstrate the validity and usability of the geometric softmax in practical use-cases. We provide a PyTorch package for reusing the discrete geometric softmax layer[1].

5.1. Ordinal regression

We first demonstrate the g-softmax for *ordinal regression*. In this setting, we wish to predict an ordered category among d categories, and we assume that the cost of predicting $\hat{y}$ instead of y is symmetric and depends on the difference between $\hat{y}$ and y. For instance, when predicting ratings, we may have three categories *bad* $\prec$ *average* $\prec$ *good*. This is typically modeled by a cost-function $C(\hat{y}, y) = \phi(|\hat{y} - y|)$, where ϕ is the ℓ_2^2 or ℓ_1 cost. We use the real-world ordinal datasets provided by Gutierrez et al. (2016), using their predefined 30 cross-validation folds.

Experiment and results. We study the performance of the geometric softmax in this discrete setting, where the score function is assumed to be a linear function of the input features $x \in \mathbb{R}^k$, i.e, $g_{W,b}(x) = Wx + b$, with $b \in \mathbb{R}^d$, $x \in \mathbb{R}^k$ and $W \in \mathbb{R}^{d \times k}$. We compare its performance to multinomial regression, and to immediate threshold and all-threshold logistic regression (Rennie & Srebro, 2005), using a reference implementation provided by Pedregosa et al. (2017). We use a cross-validated ℓ_2 penalty term on the linear score model g_θ. To compute the Hausdorff divergence at test time and the geometric loss during training, we set $C(\hat{y}, y) = (\hat{y} - y)^2/2$.

The results, aggregated over datasets and cross-validation folds, are reported in Table 1. We observe that the g-logistic regression performs better than the others for the Hausdorff divergence on average. It performs slightly worse than a simple logistic regression in term of accuracy, but

[1] github.com/arthurmensch/g-softmax

Table 1. Performance of geometric loss as a drop-in replacement in linear models for ordinal regression. Our method performs better w.r.t. its natural metric, the Hausdorff divergence.

	LR	LR(AT)	LR(IT)	**g-logistic**
Haus. div.	.46±.12	.47±.14	.59±.16	**.44±.08**
MAE	.44±.09	**.42±.06**	.44±.08	.45±.09
Acc.	**.66±.07**	.65±.06	.65±.06	.65±.07

slightly better in term of mean absolute error (MAE, the reference metric in ordinal regression). It thus provides a viable alternative to thresholding techniques, that performs worse in accuracy but better in MAE. It has the further advantage of naturally providing a distribution of output given an input x. We simply have, for all $y \in [d]$, $p(Y = y | X = x) = (\text{g-softmax}(g_{W,b}(x)))_y$.

Calibration of the geometric loss. We validate Prop. 5 experimentally on the ordinal regression dataset *car*. During training, we measure the geometric cross-entropy loss and the Hausdorff divergence on the train and validation set. Figure 3 shows that ℓ_Ω is indeed an upper bound of D_Ω, and that the difference between both terms reduces to almost 0 on the train set. Prop. 5 ensures this finding provided that the set of scoring function is large enough, which appears to be approximately the case here.

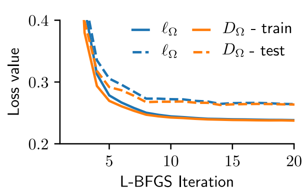

Figure 3. Training curves for ordinal regression on dataset *car*. The difference between the g-logistic loss and the Hausdorff divergence vanishes on the train set.

5.2. Drawing generation with variational auto-encoders

The proposed geometric loss and softmax are suitable to estimate distributions from inputs. As a proof-of-concept experiment, we therefore focus on a setting in which distributional output is natural: generation of hand-drawn doodles and digits, using the Google QuickDraw (Ha & Eck, 2018) and MNIST dataset. We train variational autoencoders on these datasets using, as output layers, (1) the KL divergence with normalized output and (2) our geometric loss with normalized output. These approaches output an image prediction using a softmax/g-softmax over all pixels, which is justified when we seek to output a concentrated distributional output. This is the case for doodles and digits, which can be seen as 1D distributions in a 2D space. It differs from the more common approach that uses a binary cross-entropy loss for every pixel and enables to capture interactions between pixels at the feature extraction level. We use standard KL penalty on the latent space distribution.

Using the g-softmax takes into account a cost between pixels (i, j) and (k, l), that we set to be the Euclidean cost C/σ, where C is the ℓ_2^2 cost and σ is the typical distance of interaction—we choose $\sigma = 2$ in our experiments. We therefore made the hypothesis that it would help in reconstructing the input distributions, forming a non-linear layer that captures interaction between inputs in a non-parametric way.

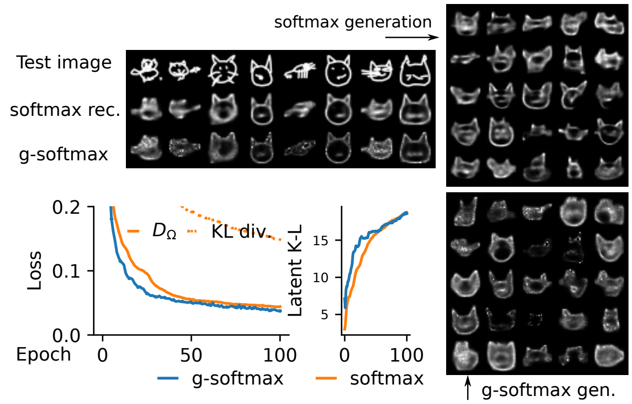

Figure 4. The g-softmax layer permits to generate and reconstruct drawing in a more concentrated manner. For a same level of variational penalty, the g-softmax better and faster minimizes the asymmetric Hausdorff divergence. See also Figure 6.

Results. We fit a simple MLP VAE on 28x28 images from the QuickDraw Cat dataset. Experimental details are reported in Appendix B (see Figure 6). We also present an experiment with 64x64 images and a DCGAN architecture, as well as visualization of a VAE fitted on MNIST. In Figure 4, we compare the reconstruction and the samples after training our model with the g-softmax and simple softmax loss. Using the g-softmax, which has a deconvolutional effect, yields images that are concentrated near the edges we want to reconstruct. We compare the training curves for both the softmax and g-softmax version: using the g-softmax link function and its associated loss better minimizes the asymmetric Hausdorff divergence. The cost of computation is again increased by a factor 10.

6. Conclusion

We introduced a principled way of learning distributional predictors in potentially continuous output spaces, taking into account a cost function in between inputs. We constructed a geometric softmax layer, that we derived from Fenchel conjugation theory in Banach spaces. The key to our construction is an entropy function derived from regularized optimal transport, convex and weak⋆ continuous on probability measures. Beyond the experiments in discrete measure spaces that we presented, our framework opens the doors for new applications that are intrinsically off-the-grid, such as super-resolution.

Acknowledgements

The work of A. Mensch and G. Peyré has been supported by the European Research Council (ERC project NORIA). A. Mensch thanks Jean Feydy and Thibault Séjourné for fruitful discussions.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, pp. 214–223, 2017.

Aspert, N., Santa-Cruz, D., and Ebrahimi, T. Mesh: Measuring errors between surfaces using the hausdorff distance. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 705–708, 2002.

Bernhard, P. and Rapaport, A. On a theorem of Danskin with an application to a theorem of Von Neumann-Sion. *Nonlinear Analysis: Theory, Methods & Applications*, 24(8):1163–1181, 1995.

Blondel, M., Martins, A. F. T., and Niculae, V. Learning classifiers with Fenchel-Young losses: Generalized entropies, margins, and algorithms. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2018.

Blondel, M., Martins, A. F., and Niculae, V. Learning with Fenchel-Young losses. *arXiv:1901.02324*, 2019.

Boyd, N., Schiebinger, G., and Recht, B. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.

Boyd, N., Jonas, E., Babcock, H. P., and Recht, B. DeepLoco: Fast 3D localization microscopy using neural networks. *BioRxiv*, pp. 267096, 2018.

Bredies, K. and Pikkarainen, H. K. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

Bregman, L. M. 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.

Ciliberto, C., Rosasco, L., and Rudi, A. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, pp. 4412–4420, 2016.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 3730–3739, 2017.

Crammer, K. and Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Cuturi, M. and Doucet, A. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pp. 685–693, 2014.

Cuturi, M. and Peyré, G. Semi-dual regularized optimal transport. *SIAM Review*, 60(4):941–965, 2018.

Danskin, J. M. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

Duchi, J. C., Khosravi, K., and Ruan, F. Multiclass classification, information, divergence, and surrogate risk. *Annals of Statistics*, 2018.

Feydy, J. and Trouvé, A. Global divergences between measures: From Hausdorff distance to optimal transport. In *ShapeMI: Shape in Medical Imaging*, 2019.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the International Conference on Artifical Intelligence and Statistics*, 2019.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pp. 2053–2061, 2015.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pp. 3440–3448, 2016.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with Sinkhorn divergences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2018.

Gimpel, K. and Smith, N. A. Softmax-margin CRFs: Training log-linear models with cost functions. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2010.

Gini, C. *Variabilità e Mutuabilità*. C. Cuppini, Bologna, 1912.

Gutierrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., and Hervas-Martinez, C. Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2016.

Ha, D. and Eck, D. A neural representation of sketch drawings. In *International Conference on Learning Representations*, 2018.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. From word embeddings to document distances. In *Proceedings of the International Conference on Machine Learning*, pp. 957–966, 2015.

Liu, D. C. and Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989.

Luenberger, D. G. *Optimization by vector space methods*. John Wiley & Sons, 1997.

Luise, G., Rudi, A., Pontil, M., and Ciliberto, C. Differential properties of Sinkhorn approximation for learning with Wasserstein distance. In *Advances in Neural Information Processing Systems 31*, pp. 5864–5874, 2018.

Martins, A. F. T. and Astudillo, R. F. From Softmax to Sparsemax: A sparse model of attention and multi-label classification. *Proceedings of the International Conference on Machine Learning*, pp. 1614–1623, 2016.

Mensch, A. and Blondel, M. Differentiable dynamic programming for structured prediction and attention. In *Proceedings of the International Conference on Machine Learning*, pp. 3462–3471, 2018.

Mitter, S. K. Convex optimization in infinite dimensional spaces. In *Recent Advances in Learning and Control*, volume 371, pp. 161–179. Springer, 2008.

Montavon, G., Müller, K.-R., and Cuturi, M. Wasserstein training of restricted Boltzmann machines. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3718–3726. Curran Associates, Inc., 2016.

Pedregosa, F., Bach, F., and Gramfort, A. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18(1):1769–1803, 2017.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

Reid, M. D. and Williamson, R. C. Composite binary losses. *Journal of Machine Learning Research*, 11: 2387–2422, 2010.

Rennie, J. D. and Srebro, N. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, pp. 180–186, 2005.

Rockafellar, R. T. Extension of Fenchel duality theorem for convex functions. *Duke Mathematical Journal*, 33 (1):81–89, 1966.

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, 1970.

Rolet, A., Cuturi, M., and Peyré, G. Fast dictionary learning with a smoothed Wasserstein loss. In *Artificial Intelligence and Statistics*, pp. 630–638, 2016.

Santambrogio, F. *Optimal transport for applied mathematicians*. Birkäuser, NY, 2015.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

Székely, G. J. and Rizzo, M. L. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Tewari, A. and Bartlett, P. L. On the consistency of multiclass classification methods. *Learning Theory*, 3559: 143–157, 2005.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

Villani, C. *Optimal transport: Old and new*. Springer, 2008.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Weston, J. and Watkins, C. Support vector machines for multi-class pattern recognition. In *Proceedings of the European Symposium on Artificial Neural Networks*, pp. 219–224, 1999.

Williamson, R. C., Vernet, E., and Reid, M. D. Composite multiclass losses. *Journal of Machine Learning Research*, 17(223):1–52, 2016.