THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'Université Paris Dauphine
et l'École Normale Supérieure

Entropy-Regularized Optimal Transport for Machine Learning

Spécialité SCIENCES

RAPPORTEURS :

Stefanie JEGELKA
Massachusetts Institute of Technology

Lorenzo ROSASCO
Università di Genova

COMPOSITION DU JURY :

Francis BACH
Ecole Normale Supérieure et INRIA Paris

Jean-David BENAMOU
INRIA Paris

Jérémie BIGOT
Université de Bordeaux

Olivier BOUSQUET
Google Brain Zurich

Marco CUTURI
ENSAE et Google Brain Paris

Rémi FLAMARY
Université de Nice Sophia-Antipolis

Jean-Michel LOUBES
Université Toulouse Paul Sabatier

Gabriel PEYRÉ *(Directeur de Thèse)*
Ecole Normale Supérieure

Soutenue par **Aude Genevay**
le 13 mars 2019

Dirigée par **Gabriel Peyré**

DAUPHINE UNIVERSITÉ PARIS | PSL RESEARCH UNIVERSITY PARIS

Abstract

This thesis proposes theoretical and numerical contributions to use Entropy-regularized Optimal Transport (EOT) for machine learning. We introduce Sinkhorn Divergences (SD), a class of discrepancies between probability measures based on EOT which interpolates between two other well-known discrepancies: Optimal Transport (OT) and Maximum Mean Discrepancies (MMD). We develop an efficient numerical method to use SD for density fitting tasks, showing that a suitable choice of regularization can improve performance over existing methods. We derive a sample complexity theorem for SD which proves that choosing a large enough regularization parameter allows to break the curse of dimensionality from OT, and recover asymptotic rates similar to MMD. We propose and analyze stochastic optimization solvers for EOT, which yield online methods that can cope with arbitrary measures and are well suited to large scale problems, contrarily to existing discrete batch solvers.

Résumé

Le Transport Optimal régularisé par l'Entropie (TOE) permet de définir les Divergences de Sinkhorn (DS), une nouvelle classe de distance entre mesures de probabilités basées sur le TOE. Celles-ci permettent d'interpoler entre deux autres distances connues: le Transport Optimal (TO) et l'Ecart Moyen Maximal (EMM). Les DS peuvent être utilisées pour apprendre des modèles probabilistes avec de meilleures performances que les algorithmes existants pour une régularisation adéquate. Ceci est justifié par un théorème sur l'approximation des SD par des échantillons, prouvant qu'une régularisation suffisante permet de se débarrasser de la malédiction de la dimension du TO, et l'on retrouve à l'infini le taux de convergence des EMM. Enfin, nous présentons de nouveaux algorithmes de résolution pour le TOE basés sur l'optimisation stochastique 'en-ligne' qui, contrairement à l'état de l'art, ne se restreignent pas aux mesures discrètes et s'adaptent bien aux problèmes de grande dimension.

Contents

Outline of the Thesis

Comparing probability distributions is a fundamental component of many machine learning problems, both supervised and unsupervised. The main matter of this thesis is to study the behavior of a class of discrepancies between probability distributions, called Sinkhorn Divergences, which are based on entropy-regularized Optimal Transport. We provide both theoretical contributions, regarding their statistical properties, and numerical ones, including solvers to compute Sinkhorn Divergences and use them in machine learning tasks.

Supervised Machine Learning. In *supervised* machine learning, we are provided with a labeled dataset *e.g* $(x_i, y_i)_{i=1\dots n}$ where x_i is the observation in some input space $\mathcal{X}$ (*e.g.* pixel intensities of an image) and y_i is the associated label (*e.g.* the fact that this image represents an apple). A recurrent issue in supervised learning is to learn a classification rule from the data, that takes a new observation x as an input and predicts the associated label y as the output. For instance in nearest-neighbor classification, when provided with a new observation x, one looks for the closest observation x_{i^*} in the dataset and sets $y = y_{i^*}$. This classification rule assumes that if observations are close, they should have the same label. Defining a meaningful notion of distance on the data space $\mathcal{Z}$ is thus crucial. In practice, a lot of data can be represented as histograms on some other space $\mathcal{X}'$: a data point $x \in \mathcal{X}$ is identified to a histogram $\alpha \overset{\text{def.}}{=} \sum_{i=1}^n \boldsymbol{\alpha}_i \delta_{a_i}$, where $(a_i)_i \in \mathcal{X}'$ and $\sum_{i=1}^n \boldsymbol{\alpha}_i = 1$. Since normalized histograms are no more than finite discrete probability distributions, a distance on probability distributions serves as a relevant distance on these data spaces. As a set of representative examples, let us quote: bag-of-visual-words comparison in computer vision (Rubner et al., 2000), color and shape processing in computer graphics (Solomon et al., 2015), bag-of-words for natural language processing (Kusner et al., 2015). Another use of histograms in supervised learning is to associate labels to histograms in multi-label classification (Frogner et al., 2015).

Unsupervised Machine Learning. On the other hand, in *unsupervised* machine-learning, the dataset only consists in observations $(x_i)_{i=1\dots n}$ in the data space $\mathcal{X}$. One way to extract information from the data in an unsupervised setting is to perform

density fitting. The goal is to fit the unknown distribution induced by the dataset with a parametric distribution. This amounts to finding the parameters that minimize some notion of distance between these two distributions (the unknown one from the dataset, and the one from the parametric model). Choosing the right notion of discrepancy between measures here is one of the key issues of the problem. A popular research area in unsupervised learning which emerged in recent years is learning *generative models* (Goodfellow et al., 2014) which can generate new samples resembling the ones in the dataset. The distributions induced by generative models are often assumed to have intrinsic low dimension, and thus do not have a density with respect to a reference measure. The usual Maximum Likelihood Estimation framework can therefore not be used for such models. However, these models are easy to sample from, and thus resorting to a discrepancy on measures which can be robustly computed from samples (from the generative model and the dataset) is essential.

Discrepancies on measures. The most popular frameworks that are used to compare probability distributions are $\varphi-$divergences (Csiszár, 1975), Maximum Mean Discrepancies (MMD) (Gretton et al., 2006) and Optimal Transport (OT) (Kantorovich, 1942). The former are appreciated for their computational simplicity, but they suffer from the major shortcoming of not metrizing weak-convergence. Both MMD and OT have the ability to metrize weak-convergence, but they enjoy different characteristics. MMD can be efficiently estimated from samples of the measures, both statistically since the estimates are robust with a small number or samples (we say it has a good sample complexity) and also numerically, as they are computed in closed form. OT on the other hand, presents none of these advantages, but has the ability to lift a ground metric from the dataspace $\mathcal{X}$ to the set of probability measures on this space and thus take into account the underlying geometry of the data. Its good geometric properties can be strengthened by enforcing structure constraints (Alvarez-Melis et al., 2017) which allows for instance to take into account the class labels in supervised learning. Besides, solving OT also gives a mapping from one measure to the other, which has been successfully used in domain adaptation (Courty et al., 2014). As a unifying alternative to these discrepancies, we introduce Sinkhorn Divergences, based on entropy-regularized Optimal Transport. We prove they interpolate between MMD (with infinitely strong regularization) and OT (with no regularization). In particular, Sinkhorn Divergences preserve the good geometric properties of OT, and also provide a mapping from one measure to the other. However unlike OT – but similarly to MMD – they benefit from good statistical properties and efficient computation.

We now get into more details regarding the technical aspects of our work, formalizing key concepts and outlining the main contributions of this thesis.

Chapter 1: Entropy-regularized Optimal Transport

This introductory chapter is both a review of existing tools commonly used in machine learning to compare probability distributions, and a presentation of key properties of regularized optimal transport (containing both new results and existing ones from the literature), which serves as a basis for the work presented in subsequent chapters.

Previous Works. In machine learning, the first discrepancies that were introduced to compare two probability distributions are **φ-divergences** (Csiszár, 1975), which can be seen as a weighted average (by φ) of the odds-ratio between the two measures. Consider φ a convex, lower semi-continuous function such that $\varphi(1) = 0$, the φ-divergence D_φ between two probability measures α and β is defined by:

$$D_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\Big(\frac{\mathrm{d}\alpha(x)}{\mathrm{d}\beta(x)}\Big)\mathrm{d}\beta(x).$$

up to a corrective (possibly infinite) term if α is not absolutely continuous with respect to β. The computational simplicity of $\varphi-$divergences made them quite popular – the most widely used being the Kullback-Leibler divergence for $\varphi(x) = x\log(x)$. However, they suffer from the major drawback of not metrizing weak-convergence (or convergence in law). A measure α_n weakly converges to α (denoted $\alpha_n \rightharpoonup \alpha$) if and only if $\int f(x)\mathrm{d}\alpha_n(x) \to \int f(x)\mathrm{d}\alpha_n(x)$ for all continuous bounded functions f ; and a loss $\mathcal{L}$ metrizes weak-convergence if and only if $\mathcal{L}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha$. The metrization of weak-convergence is instrumental for discrepancies on measure, as it ensures that the losses remain stable under small perturbations of the support of the measures. As an example, consider the case on $\mathbb{R}$ where $\alpha = \delta_0$ a Dirac mass in 0 and $\alpha_n = \delta_{1/n}$ a Dirac mass in $1/n$. Then $D_\varphi(\alpha_n|\alpha)$ is a constant for all n, although it seems natural to say that when n goes to infinity, α_n gets closer to α. This failure case in $\mathbb{R}$ becomes very problematic in higher dimension, when comparing probability distributions that are supported on low-dimensional manifolds for instance.

The two main classes of discrepancies that satisfy this requirement are **Maximum Mean Discrepancies (MMD)**(Gretton et al., 2006) and **Optimal Transport (OT)** (Santambrogio, 2015) based losses. MMD are a special instance of Integral Probability Metrics (Müller, 1997). Given a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with kernel k; MMD between two probability measures α and β are defined as follows:

$$\begin{aligned} MMD_k^2(\alpha,\beta) &\overset{\text{def.}}{=} \left(\sup_{\{f|\|f\|_{\mathcal{H}}\leqslant 1\}} |\mathbb{E}_\alpha(f(X)) - \mathbb{E}_\beta(f(Y))|\right)^2 \\ &= \mathbb{E}_{\alpha\otimes\alpha}[k(X,X')] + \mathbb{E}_{\beta\otimes\beta}[k(Y,Y')] - 2\mathbb{E}_{\alpha\otimes\beta}[k(X,Y)]. \end{aligned} \tag{0.1}$$

If the kernel k is universal (i.e. its RKHS is dense in the space of continuous functions), they are positive definite, and under some further technical assumptions, they

metrize weak-convergence (Sriperumbudur et al., 2010). This family of losses presents the advantage of being efficiently computed from samples – both in a computational and statistical sense (Gretton et al., 2006). OT-based losses on the other hand behave particularly well in problems that are intrinsically geometric (e.g. shapes or image processing). They rely on the choice of a ground cost c which reflects the geometry of the input space in the following way:

$$W_c(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)\mathrm{d}\pi(x,y), \tag{$\mathcal{P}$}$$

where the feasible set is composed of joint probability distributions with fixed marginals α,β. A typical choice is $c = d^p$, where d is the natural distance on $\mathcal{X}$, for which W_c metrizes weak-convergence when $p > 1$ (Santambrogio, 2015). However, these losses suffer from a computational burden – solving OT requires solving a linear program in the discrete case – and a curse of dimensionality, meaning their approximation from sampled measures degrades quickly in high dimension (Weed and Bach, 2017).

Entropy-regularized OT has recently emerged as a solution to the computational issue of OT (Cuturi, 2013). The regularized problem reads:

$$W_{c,\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon H(\pi|\alpha\otimes\beta), \tag{$\mathcal{P}_\varepsilon$}$$

where

$$H(\pi|\alpha\otimes\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{X}} \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\pi(x,y). \tag{0.2}$$

is the relative entropy of the transport plan π with respect to the product measure $\alpha\otimes\beta$. It has an equivalent dual formulation, which is unconstrained (contrarily to standard OT):

$$\begin{aligned} W_{c,\varepsilon}(\alpha,\beta) = \max_{\substack{u\in\mathcal{C}(\mathcal{X})\\ v\in\mathcal{C}(\mathcal{X})}} &\int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{X}} v(y)\mathrm{d}\beta(y) \\ &- \varepsilon\int_{\mathcal{X}\times\mathcal{X}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon. \end{aligned} \tag{$\mathcal{D}_\varepsilon$}$$

In the case of finite discrete measures, iteratively optimizing over each dual variables yields a fast converging algorithm, called Sinkhorn's algorithm (Sinkhorn, 1967). Besides, the resulting distance happens to perform well in various machine learning tasks as proved in the seminal paper by Cuturi (2013), which opened the way to the use of entropy-regularized OT in the community.

Contributions. The main objective of this thesis is to prove theoretically and numerically that the benefits of entropy-regularized OT extend far beyond this fast algorithm

for finite discrete measures, and in this chapter we review the bases that will be required for our main contributions presented in subsequent chapters. However, this collection of results also includes some original contributions on regularized OT which consist in

(i) **Regularization of OT using relative entropy with respect to the product measure of the marginals**: The seminal paper by Cuturi (2013) deals with the discrete case and uses entropy $H(\pi) \overset{\text{def.}}{=} \sum_{i,j} \log(\pi_{ij})\,\pi_{ij}$ as a regularizer. We suggest instead to use the entropy with respect to the product of marginals defined in equation (0.2), as it allows to formulate the dual problem as the maximization of an expectation :

$$W_{c,\varepsilon}(\alpha,\beta) = \max_{\substack{u\in\mathcal{C}(\mathcal{X})\\ v\in\mathcal{C}(\mathcal{X})}} \mathbb{E}_{\alpha\otimes\beta}\left[f_\varepsilon^{XY}(u,v)\right] + \varepsilon,$$

where $f_\varepsilon^{xy}(u,v) \overset{\text{def.}}{=} u(x)+v(y)-\varepsilon e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}$ and X, Y are distributed according to α and β respectively. This formulation is key to deriving statistical properties of entropy-regularized OT in Chapter 3 and new solvers in Chapter 4.

(ii) **Semi-Dual formulation**: When one of the measures is a weighted sum of n dirac masses, the associated dual variable is a n dimensional vector. Assume (without loss of generality, since the problem is symmetric) that it is the case of the second measure: $\beta \overset{\text{def.}}{=} \sum_{i=1}^n \boldsymbol{\beta}_i \delta_{x_i}$. We exploit the joint convexity of the dual problem, by using the optimality condition over the first dual variable to derive a so-called *Semi-Dual formulation* of entropy-regularized OT:

$$W_{c,\varepsilon}(\alpha,\beta) = \max_{\mathbf{v}\in\mathbb{R}^n} \int_{\mathcal{X}} -\varepsilon \log\left(\sum_{i=1}^n e^{\frac{\mathbf{v}_i - c(x,y_i)}{\varepsilon}} \boldsymbol{\beta}_i\right) \mathrm{d}\alpha(x) + \sum_{i=1}^n \mathbf{v}_i \boldsymbol{\beta}_i. \qquad (\mathcal{S}_\varepsilon)$$

This problem is an optimization problem over $\mathbb{R}^d$, which can also be rewritten as the maximum of an expectation. We make use of this formulation in Chapter 4, resorting to stochastic optimization to solve this problem.

(iii) **Generalization of previous proofs of existence of solutions to the dual problem** ($\mathcal{D}_\varepsilon$): A proof of existence of dual potentials already exist in the discrete case (Franklin and Lorenz, 1989), and for Schrödinger's problem (Chen et al., 2016) which shares strong links with regularized OT. Relying on the same proof technique, i.e. proving that dual potentials are fixed point of contractions for the Hilbert metric, we extend the proof to any arbitrary probability measures, and a bounded cost function.

(iv) **Extension of entropy-regularized OT**: This thorough introduction to entropy-regularized OT is also an opportunity to generalize some of our results to regularizers other than entropy, replacing $H(\pi|\alpha\otimes\beta)$ by $D_\varphi(\pi|\alpha\otimes\beta)$ in ($\mathcal{P}_\varepsilon$), where

D_φ is any $\varphi-$divergence. Besides, we also extend these formulations to unbalanced transport, which extends the notion of OT to positive Radon measures with arbitrary mass (Liero et al., 2018),(Chizat et al., 2018) whenever possible (e.g. regularization, formulation as an expectation).

Chapter 2: Learning with Sinkhorn Divergences

This chapter is based on (Genevay et al., 2018).

Unsupervised machine learning often boils down to fitting a parametric model to a dataset, i.e. estimating the parameters of a chosen model that *fits* observed data in some meaningful way. Formally, given a dataset of samples with unknown distribution β, we want to learn a parametric measure α_{θ^*} such that

$$\theta^* \in \operatorname*{argmin}_\theta \mathcal{L}(\alpha_\theta, \beta)$$

where $\mathcal{L}$ is some loss on measures. Note that β is unknown, and can only be accessed via a finite number of samples $(y_1, \dots, y_N) \in \mathcal{X}^N$ constituting the dataset. The standard approach for models with a density is Maximum Likelihood Estimation (MLE), setting $\mathcal{L}(\alpha_\theta, \beta) = -\sum_j \log \frac{\mathrm{d}\alpha_\theta}{\mathrm{d}x}(y_j)$, where $\frac{\mathrm{d}\alpha}{\mathrm{d}x}$ is the density of α_θ with respect to a fixed reference measure. However this approach does not work for *generative models*, obtained as the mapping of a low dimensional reference measure ζ through a non-linear parametric *pushforward function* g_θ with values in a high dimensional space (e.g. a neural network). These models are easy to sample from: a sample x from α_θ is obtained by drawing a sample z from ζ and taking $x = g_\theta(x)$. However, their density is singular in the sense that it is typically supported on a low-dimensional "manifold" of the data space $\mathcal{X}$, thus making the MLE unusable.

Previous Works. To fit generative models, several likelihood-free alternatives exist. Pioneer approaches include variational autoencoders (VAE) (Kingma and Welling, 2013) and generative adversarial networks (GAN) (Goodfellow et al., 2014) which lead to numerous variations including combinations of both ideas (Larsen et al., 2016). The adversarial GAN approach can be viewed as a two-player game where player one optimizes its parameter θ to fool player two whose goal is to discriminate between samples from the model measure α_θ and samples from the true measure β by optimizing a parametric discriminator D_w. Formally, this is equivalent to minimizing the dual of the Jensen-Shannon divergence (expressed as the maximum over a class of parametric functions D_w) between α_θ and β. This min-max approach can be extended to any given $\varphi-$divergences (Nowozin et al., 2016). Another approach is to minimize the MMD between the distribution of the data and the model. It was shown in relevant work (Li et al., 2015; Dziugaite et al., 2015) that the effectiveness of the MMD in that setting

hinges on the ability to find a relevant kernel function, which is nontrivial. The Wasserstein distance, long known to be a powerful tool to compare probability distributions with non-overlapping supports, a has recently emerged as a serious contender. Although the use of Wasserstein metrics for inference in generative models was considered over ten years ago in (Bassetti et al., 2006), that development remained exclusively theoretical until a recent wave of papers managed to implement that idea more or less faithfully using several workarounds: entropic regularization over a discrete space (Montavon et al., 2016), approximate Bayesian computations (Bernton et al., 2017) and a neural network parameterization of the dual potential arising from the dual OT problem when considering the 1-Wasserstein distance (Arjovsky et al., 2017). As opposed to this dual way to compute gradients of the fitting energy, we advocate for the use of a primal formulation, which is numerically stable, because it does not involve differentiating the (dual) solution of an OT sub-problem, as also pointed out in (Bousquet et al., 2017).

Contributions. The main contributions of this chapter include a theoretical contribution regarding a new OT-based loss for generative models, and a simple numerical scheme to learn under this loss.

(i) **Sinkhorn Divergence**: We introduce the Sinkhorn Divergence, based on regularized optimal transport with an entropy penalty:

$$SD_{c,\varepsilon}(\alpha,\beta) \stackrel{\text{def.}}{=} W_{c,\varepsilon}(\alpha,\beta) - \frac{1}{2}W_{c,\varepsilon}(\alpha,\alpha) - \frac{1}{2}W_{c,\varepsilon}(\beta,\beta), \tag{0.3}$$

where $W_{c,\varepsilon}$ is the loss induced by entropy-regularized OT. This corrects the bias introduced by entropy to ensure that $SD_\varepsilon(\alpha,\alpha) = 0$. We conjectured in the early stages of our work on Sinkhorn Divergence, based on empirical evidence, this normalization of regularized OT enforces positive-definiteness. It was recently proved in subsequent work by Feydy et al. (2019 (to appear) along with the fact that Sinkhorn Divergences metrize the weak-convergence of measures under some assumptions on the cost.

(ii) **Interpolation property**: We prove that when the smoothing parameter $\varepsilon = 0$ we recover pure OT loss whereas letting $\varepsilon = +\infty$ leads to MMD with kernel $-c$ (i.e. minus the ground cost of OT):

Theorem 1. *Consider the Sinkhorn Divergence defined in (0.3), then it has the following asymptotic behavior in ε:*

$$\text{as } \varepsilon \to 0, \quad SD_{c,\varepsilon}(\alpha,\beta) \to W_c(\alpha,\beta), \tag{0.4}$$

$$\text{as } \varepsilon \to +\infty, \quad SD_{c,\varepsilon}(\alpha,\beta) \to \frac{1}{2}MMD^2_{-c}(\alpha,\beta). \tag{0.5}$$

Note that to define a proper MMD, $-c$ needs to induce a positive definite kernel. This is the case when $c = \|\cdot\|_2^p$ for $0 < p < 2$, and the associated MMD yields the Energy Distance (Sejdinovic et al., 2013). This interpolation property is further studied in Chapter 3, where we prove that the sample complexity of Sinkhorn Divergences also interpolates between that of OT and MMD, alleviating the curse of dimensionality brought by OT when ε is sufficiently large. It is also supported by empirical evidence in this chapter.

(iii) **Learning generative models under a Sinkhorn Divergence**: We consider the density fitting problem with a Sinkhorn Divergence as a loss:

$$\theta^* \in \operatorname*{argmin}_{\theta} SD_{c,\varepsilon}(\alpha_\theta, \beta).$$

We solve the inference problem by making two key simplifications: (i) approximate $SD_\varepsilon(\alpha_\theta, \beta)$ by a size-m mini-batch sampling $SD_\varepsilon(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$ to make it amenable to stochastic gradient descent ; (ii) approximate $SD_\varepsilon(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$ by L-steps of the Sinkhorn algorithm (Cuturi, 2013) to obtain an algorithmic loss $SD_\varepsilon^{(L)}(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$ which is amenable to automatic differentiation. Numerical experiments, both on simulated and real data, show that Sinkhorn Divergences are able to capture the geometry of the data in a more powerful way than the Energy Distance, which tends to ignore extreme points.

(iv) **Adversarially learning the cost function**: Similarly to what is done for kernel functions in (Dziugaite et al., 2015), we propose to learn the cost function c adversarially. This is crucial for applications in which there is no natural distance between samples, like in computer vision where there is no universal meaningful metric between images. We parametrize the cost function in the following way:

$$c_\varphi(x, y) \overset{\text{def.}}{=} \|f_\varphi(x) - f_\varphi(y)\|^p \quad \text{where} \quad f_\varphi : \mathcal{X} \to \mathbb{R}^{d'},$$

where f_φ can be seen as a feature extractor that reduces the dimensionality of $\mathcal{X}$ through a mapping onto $\mathbb{R}^{d'}$. This cost function should make the discrepancy large, to be able to discriminate well between the model α_θ and the true distribution β, we then solve the min-max problem:

$$\min_{\theta} \max_{\varphi} SD_{c_\varphi,\varepsilon}(\alpha_\theta, \beta).$$

Shortly after the submission of this work, we came across the recent work by (Salimans et al., 2018) which shares several ideas with our method. One distinction lies in the fact that they do not back-propagate errors across the Sinkhorn iterations, but rather use an estimate of the optimal transport matrix to compute an upper-bound on the Sinkhorn

Divergence, as was done for instance in (Cuturi and Doucet, 2014).

Chapter 3: Sample Complexity of Sinkhorn Divergences

This chapter is based on (Genevay et al., 2019 (to appear).

The numerical experiments in Chapter 2 further support what was first observed in (Cuturi, 2013): entropy-regularized OT breaks the curse-of-dimensionality from which OT suffers when the regularization parameter is large enough. The goal of this chapter is to make this more formal through a sample complexity theorem. We also provide a convergence rate of entropy-regularized transport to standard transport, proving that there is a tradeoff between a faithful estimation of OT and good sample complexity.

Previous Works. The central theoretical contribution of Chapter 2 (see Theorem 1) states that Sinkhorn Divergences, based on regularized OT, interpolate between OT and MMD. These two metrics, which emerged as popular candidates to compare probability measures, differ on a fundamental aspect: their sample complexity. The definition of sample complexity of a loss function that we choose here is the convergence rate of the loss evaluated on empirical measures to the loss evaluated on the "true" measures, as a function of the number of samples. This notion is crucial in machine learning, as bad sample complexity induces overfitting and high gradient variance when using these divergences for parameter estimation. In that context, it is well known that the sample complexity of MMD is independent of the dimension, scaling as $\frac{1}{\sqrt{n}}$ (Gretton et al., 2006) where n is the number of samples. In contrast, it is well known that standard OT suffers from the curse of dimensionality (Dudley, 1969): considering a probability measure $\alpha \in \mathcal{M}(\mathbb{R}^d)$ and its empirical estimation $\hat{\alpha}_n$, we have $\mathbb{E}[W_p(\alpha, \hat{\alpha}_n)] = O(n^{-1/d})$. Its sample complexity is thus exponential in the dimension of the ambient space d. Although it was recently proved that this result can be refined to d being the intrinsic dimension of data (Weed and Bach, 2017), the sample complexity of OT is now the major bottleneck for the use of OT in high-dimensional machine learning problems.

A solution to this shortcoming comes, once again, from entropic-regularization. Sinkhorn Divergences (0.3), have been empirically observed to be less prone to overfitting, as a certain amount of regularization can improve performance in simple learning tasks (Cuturi, 2013). The interpolation property in Theorem 1 also suggests that for large regularizations, Sinkhorn Divergences should behave similarly to MMD. However, aside from a recent central limit theorem in the case of measures supported on finite discrete spaces (Bigot et al., 2017), the convergence of empirical Sinkhorn Divergences, and more generally their sample complexity, remains an open question.

Contributions. This chapter contains the main theoretical contributions of this thesis, in the form of three theorems exhibiting theoretical properties of Sinkhorn Diver-

gences.

(i) **Bound on the speed of convergence of regularized OT to standard OT:** On a bounded domain of $\mathbb{R}^d$ and with a Lipschitz cost-function c, Theorem 2 quantifies the speed of convergence of the value of regularized OT to that of standard OT with respect to the regularization parameter ε

Theorem 2. *Let α and β be probability measures on $\mathcal{X}$ and $\mathcal{Y}$ bounded subsets of $\mathbb{R}^d$ such that $|\mathcal{X}|$ and $|\mathcal{Y}| \leqslant D$ and assume that c is L-Lipschitz w.r.t. x and y. It holds*

$$0 \leqslant W_\varepsilon(\alpha,\beta) - W(\alpha,\beta) \leqslant 2\varepsilon d \log\left(\frac{e^2 \cdot L \cdot D}{\sqrt{d}\cdot\varepsilon}\right) \tag{0.6}$$

$$\underset{\varepsilon\to 0}{\sim} 2\varepsilon d \log(1/\varepsilon). \tag{0.7}$$

(ii) **The dual potentials lie in a Sobolev (RKHS) ball** We then prove that optimizers of the dual regularized optimal transport problem ($\mathcal{D}_\varepsilon$) lie in a Sobolev ball which is independent of the measures:

Theorem 3. *When $\mathcal{X}$ and $\mathcal{Y}$ are two bounded sets of $\mathbb{R}^d$ and the cost c is $\mathcal{C}^\infty$, then optimal Sinkhorn potentials (u,v) (i.e. a pair of maximizers of ($\mathcal{D}_\varepsilon$)) are uniformly bounded in the Sobolev space $\mathbf{H}^s(\mathbb{R}^d)$ and their norms satisfy*

$$\|u\|_{\mathbf{H}^s} = O\left(1 + \frac{1}{\varepsilon^{s-1}}\right) \text{ and } \|v\|_{\mathbf{H}^s} = O\left(1 + \frac{1}{\varepsilon^{s-1}}\right),$$

with constants that only depend on $|\mathcal{X}|$ (or $|\mathcal{Y}|$ for v), d, and $\left\|c^{(k)}\right\|_\infty$ for $k = 0,\dots,s$. In particular, we get the following asymptotic behavior in ε: $\|u\|_{\mathbf{H}^s} = O(1)$ as $\varepsilon \to +\infty$ and $\|u\|_{\mathbf{H}^s} = O(\frac{1}{\varepsilon^{s-1}})$ as $\varepsilon \to 0$.

This allows us to rewrite the Sinkhorn Divergence as an expectation maximization problem *in a RKHS ball* and thus justify the use of kernel-SGD for regularized OT as advocated in Chapter 4 (see contribution (iii)).

(iii) **Sample complexity of Sinkhorn Divergences:** As a consequence of this reformulation (maximization over a RKHS ball), we derive a sample complexity result. We focus on the influence of the sample size and the regularization parameter on the convergence rate of the empirical Sinkhorn Divergence (i.e., computed from samples of two continuous measures) to the continuous Sinkhorn Divergence. We show that the Sinkhorn Divergence benefits from the same sample complexity as MMD, scaling in $\frac{1}{\sqrt{n}}$ but with a constant that depends on the inverse of the regularization parameter:

Theorem 4. *Consider the Sinkhorn Divergence between two measures α and β on*

$\mathcal{X}$ and $\mathcal{Y}$ two bounded subsets of $\mathbb{R}^d$, with a $\mathcal{C}^\infty$, L-Lipschitz cost c. One has

$$\mathbb{E}|W_\varepsilon(\alpha,\beta) - W_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)| = O\left(\frac{e^{\frac{\kappa}{\varepsilon}}}{\sqrt{n}}\left(1+\frac{1}{\varepsilon^{\lfloor d/2\rfloor}}\right)\right),$$

where $\kappa = 2L|\mathcal{X}| + \|c\|_\infty$ and constants only depend on $|\mathcal{X}|$, $|\mathcal{Y}|$, d, and $\left\|c^{(k)}\right\|_\infty$ for $k = 0 \dots \lfloor d/2\rfloor$. In particular, we get the following asymptotic behavior in ε:

$$\mathbb{E}|W_\varepsilon(\alpha,\beta) - W_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)| = O\left(\frac{e^{\frac{\kappa}{\varepsilon}}}{\varepsilon^{\lfloor d/2\rfloor}\sqrt{n}}\right) \qquad \text{as } \varepsilon \to 0$$
$$\mathbb{E}|W_\varepsilon(\alpha,\beta) - W_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)| = O\left(\frac{1}{\sqrt{n}}\right) \qquad \text{as } \varepsilon \to +\infty.$$

Sample complexity worsens when getting closer to standard OT and there is therefore a tradeoff between a good approximation of OT (small regularization parameter) and fast convergence in terms of sample size (larger regularization parameter).

Chapter 4: Stochastic Optimization for Large-Scale Optimal Transport

This chapter is based on(Genevay et al., 2016).

Taking advantage of the formulation of dual regularized OT as the maximization of an expectation presented in Chapter 1 (see (i) in contributions), we propose a class of provably convergent stochastic optimization solvers. Contrarily to existing methods, which only apply to discrete measures, ours can handle both discrete and continuous distributions, with the sole requirement that one can *sample* from them.

Previous Works. The prevalent way to compute OT distances is by solving the so-called Kantorovitch problem (Kantorovich, 1942) (introduced in Chapter 1) which boils down to a large-scale linear program when dealing with discrete distributions (i.e., finite weighted sums of Dirac masses). This linear program can be solved using network flow solvers with $(n^3 \log(n))$ computational complexity (n being the number of points in the measure), which can be further refined to assignment problems when comparing measures of the same size with uniform weights (Burkard et al., 2009). Regularized approaches that solve the OT with an entropic penalization, as introduced in Chapter 1, have been shown to be efficient to approximate OT solutions at a low computational cost by applying Sinkhorn's algorithm (Sinkhorn, 1964). Its main computational advantage over competing solvers is that each iteration boils down to matrix-vector multiplications, which results in a $O(n^2)$ complexity. These operations can be easily parallelized, stream extremely well on GPU, and enjoy linear-time implementation on regular grids or triangulated domains (Solomon et al., 2015). It can also be easily extended to solve

other problems involving optimal-transport, such as the computation of Wasserstein barycenters or multimarginal optimal transport (Benamou et al., 2015).

This method is however purely discrete and cannot cope with continuous densities. The only known class of methods that overcome this limitation are so-called semi-discrete solvers (Aurenhammer et al., 1998), that can be implemented efficiently using computational geometry primitives (Mérigot, 2011). They compute distance between a discrete distribution and a continuous density, but are restricted to the Euclidean squared cost, and can only be implemented in low dimensions. Lastly, let us point out that there is currently no method that can compute OT distances between two continuous densities, which is thus an open problem we tackle in this chapter.

Contributions. This chapter introduces a new class of *online stochastic optimization algorithms* to deal with large-scale (discrete measures with a very large number of points) and/or high dimensional OT problems. They can handle arbitrary distributions (discrete or continuous) as long as one is able to draw samples from them. This alleviates the need to discretize these densities, which introduces an important bias in high dimension, while giving access to provably convergent methods. These algorithms rely on one key idea which is that the dual ($\mathcal{D}_\varepsilon$) and semi-dual ($\mathcal{S}_\varepsilon$) OT problems can be re-cast as the maximization of an expectation. When $\beta = \sum_{j=1}^m \boldsymbol{\beta}_j \delta_{y_j}$ is a discrete measure, the semi-dual problem ($\mathcal{S}_\varepsilon$) in expectation form is

$$W_\varepsilon(\alpha, \beta) = \max_{\mathbf{v} \in \mathbb{R}^m} \mathbb{E}_\alpha \left[g_\varepsilon^X(\mathbf{v}) \right], \tag{0.8}$$

where $X \sim \alpha$ and

$$g_\varepsilon^x(\mathbf{v}) = \sum_{j=1}^m \mathbf{v}_j \boldsymbol{\beta}_j + \begin{cases} -\varepsilon \log(\sum_{j=1}^m \exp(\frac{\mathbf{v}_j - c(x,y_j)}{\varepsilon}) \boldsymbol{\beta}_j) & \text{if } \varepsilon > 0, \\ \min_j \left(c(x, y_j) - \mathbf{v}_j \right) & \text{if } \varepsilon = 0. \end{cases}$$

We exploit this formulation in the discrete-discrete and semi-discrete setups, and rely on the standard dual ($\mathcal{D}_\varepsilon$) when neither measures are discrete:

(i) **Comparing two finite discrete measures**: When $\alpha = \sum_{i=1}^n \boldsymbol{\alpha}_i \delta_{x_j}$ and $\beta = \sum_{j=1}^m \boldsymbol{\beta}_j \delta_{y_j}$, the semi-dual regularized OT problem ($\mathcal{S}_\varepsilon$) becomes the maximization of a sum of n functions. This can be efficiently solved thanks to stochastic gradient methods with variance reduction – we use **Stochastic Averaged Gradients** (SAG) in our experiments. The iterates of SAG can be summarized by the following formula

$$\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + \frac{C}{n} \sum_{i=1}^n z_i^{(k)},$$

where an index $i^{(k)}$ is selected at random in $\{1 \dots n\}$ and

$$z_i^{(k)} = \begin{cases} \nabla g_\varepsilon^{x_i}(\mathbf{v}^{(k)}) & \text{if } i = i^{(k)} \\ z_i^{(k-1)} & \textit{otherwise.} \end{cases}$$

At each iteration an index $i^{(k)}$ is selected at random in $\{1 \dots n\}$ to compute $\nabla g_\varepsilon^{x_{i^{(k)}}}(\mathbf{v}^{(k)})$, the gradient corresponding to the sample $x_{i^{(k)}}$ at the current estimate $\mathbf{v}^{(k)}$. SAG keeps in memory a copy of that gradient and computes an *average* of all gradients stored so far which provides a better proxy of full gradient $\nabla \bar{\mathbb{E}}_\alpha[g_\varepsilon^X]$. Compared to Sinkhorn, which can be viewed as a batch method, SAG is an online algorithm which reduces the complexity of each iteration to $O(m)$, with a $O(1/k)$ convergence rate (since our objective is not strongly convex). There is thus a tradeoff in iteration complexity vs. convergence rate to consider when using Sinkhorn or SAG. The latter is thus more efficient for problems with a very large m – i.e. discrete measures with a very large number of points.

(ii) **Comparing a finite discrete measure to an arbitrary probability measure**: We solve the semi-dual problem ($\mathcal{S}_\varepsilon$) in expectation form defined in (0.8) thanks to the **Stochastic Gradient Descent** (SGD) algorithm. The idea of SGD is fairly intuitive : at each iteration, a sample x_k is drawn from α and its gradient $\nabla g_\varepsilon^{x_k}$ is computed at the current iterate $\mathbf{v}^{(k)}$ to serve as a proxy for the full gradient $\nabla \bar{G}_\varepsilon$. The iterations are given by:

$$\mathbf{v}^{(k+1)} = \mathbf{v}(k) + \frac{C}{\sqrt{k}} \nabla_v g_\varepsilon^{x_k}(\mathbf{v}^{(k)}) \quad \text{where} \quad x_k \sim \alpha.$$

Since samples from α are drawn *online*, i.e. without prior discretization, this method avoids the discretization bias introduced when using a discrete solvers. It has a $O(1/\sqrt{k})$ convergence rate along with a $O(m)$ complexity per iteration. This online semi-discrete algorithm has been successfully applied to texture synthesis in image processing (Galerne et al., 2018), and to the computation of Wasserstein Barycenters (Staib et al., 2017).

(iii) **Comparing two arbitrary probability measures**: When neither measures are finite discrete ones, we resort to the dual formulation

$$W_\varepsilon(\alpha, \beta) = \max_{\substack{u \in \mathcal{C}(\mathcal{X}) \\ v \in \mathcal{C}(\mathcal{Y})}} \mathbb{E}_{\alpha \otimes \beta} \left[f_\varepsilon^{XY}(u, v) \right] + \varepsilon,$$

where $f_\varepsilon^{xy}(u, v) \overset{\text{def.}}{=} u(x) + v(y) - \varepsilon e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}$. We propose a **stochastic gradient descent over a Reproducing Kernel Hilbert Space** (RKHS), by using the fundamental property of a RKHS $\mathcal{H}$ with kernel k: $u \in \mathcal{H} \Leftrightarrow u(x) = \langle u, k(x, \cdot) \rangle$.

Theorem 3 from Chapter 3, stating that the dual potentials are in a RKHS ball, allows to prove the convergence of this method. We also introduce an approximate feature approach (via incomplete Cholesky decomposition (Wu et al., 2006) or Random Fourier features (Rahimi and Recht, 2007)) to significantly reduce computational time, going from quadratic to linear in the number of iterations. This is currently the only known method to solve entropy-regularized OT between arbitrary measures.

Notations

Ambient space. For a metric space $\mathcal{X}$, we denote by :

- $\mathcal{C}(\mathcal{X})$ the space of continuous functions on $\mathcal{X}$,
- $\mathcal{C}_b(\mathcal{X})$ the space of continuous bounded functions on $\mathcal{X}$,
- $\mathcal{C}^\infty(\mathcal{X})$ the space of continuous functions, infinitely differentiable with continuous derivatives on $\mathcal{X}$,
- $\mathcal{M}_+(\mathcal{X})$ the set of positive Radon measures on $\mathcal{X}$,
- $\mathcal{M}_+^1(\mathcal{X})$ the set of positive Radon probability measures (i.e. of mass 1) on $\mathcal{X}$.

When $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$, we denote by $|\mathcal{X}|$ its diameter, defined by $|\mathcal{X}| \stackrel{\text{def.}}{=} \max_{x,x' \in \mathcal{X}} \|x - x'\|$.

Measures. We use upper-cases to denote random variables (e.g. X). We denote by $X \sim \alpha$ the fact that a random variable X follows a distribution $\alpha \in \mathcal{M}_+^1(\mathcal{X})$. We write $\mathbb{E}_\alpha(f(X)) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x)$, the expectation of the random variable $f(X)$, for any measurable function f on $\mathcal{X}$. The Dirac measure at point x is δ_x. We denote by $\hat{\alpha}_n$ the empirical measure obtained from n i.i.d. samples $(x_1, \dots x_n)$ of α, *i.e.* $\hat{\alpha}_n \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$, $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, we define

$$\Pi(\alpha, \beta) \stackrel{\text{def.}}{=} \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \,|\, \forall (A, B) \subset \mathcal{X} \times \mathcal{Y}, \pi(A \times \mathcal{Y}) = \alpha(A), \pi(\mathcal{X} \times B) = \beta(B)\},$$

the set of joint probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals α and β. For some continuous map $g : \mathcal{Z} \to \mathcal{X}$, we denote $g_\sharp : \mathcal{M}_+^1(\mathcal{Z}) \to \mathcal{M}_+^1(\mathcal{X})$ the associated push-forward operator, which is a linear map between distributions. This corresponds to defining, for $\zeta \in \mathcal{M}_+^1(\mathcal{Z})$ and $B \subset \mathcal{X}$, $(g_\sharp \zeta)(B) = \zeta(g^{-1}(B))$; or equivalently, that $\int_{\mathcal{X}} \varphi \mathrm{d}(g_\sharp \zeta) = \int_{\mathcal{Z}} \varphi \circ g \mathrm{d}\zeta$ for continuous functions φ on $\mathcal{X}$. A random sample x from $g_\sharp \zeta$ can be obtained as $x = g(z)$ where z is a random sample from ζ, i.e. $g_\sharp \zeta$ is the law of $g(Z)$, where $Z \sim \zeta$.

Vectors and matrices. We use bold lower-case for vectors (e.g. $\mathbf{a}$) and bold upper-case for matrices (e.g. $\mathbf{A}$). For a matrix $\mathbf{A}$, $\mathbf{A}^\top$ denotes its transpose. Element-wise multiplication of vectors is denoted by $\odot$. For two vectors (or matrices) $\langle \mathbf{u}, \mathbf{v} \rangle \stackrel{\text{def.}}{=} \sum_i \mathbf{u}_i \mathbf{v}_i$ is the canonical inner product (the Frobenius dot-product for matrices). We denote $\mathbb{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^n$ and $\mathbb{0}_n = (0, \dots, 0)^\top \in \mathbb{R}^n$. The probability simplex of n bins is $\Sigma_n = \{\boldsymbol{\alpha} \in \mathbb{R}_+^n \,;\, \sum_i \boldsymbol{\alpha}_i = 1\}$.

Others. We use the notation $\varphi(x) = O(1 + x^k)$ to say that $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is bounded by a polynomial of order k in x with positive coefficients.

Chapter 1

Entropy-regularized Optimal Transport

This chapter is a collection of fundamental results on discrepancies between probability measures, with a focus on entropy-regularized optimal transport. Many problems in machine learning boil down to comparing probability measures, thus the question of the right notion of discrepancy between these measures is itself a crucial matter.

We start by a review of three popular candidates: φ-divergences, Maximum Mean Discrepancies (MMD) and Optimal Transport (OT). While φ-divergences are appreciated for their simplicity, they do not metrize weak convergence. This shortcoming is overcome by MMDs, defined as Integral Probability Metrics on the ball of Reproducing Kernel Hilbert Spaces (RKHS), which can also be efficiently estimated through samples. As for OT, its ability to capture the geometry of the data makes it an interesting candidate but the fact that it suffers from a curse of dimensionality and its computational burden make it impractical.

The recent introduction of Entropy-regularized OT (EOT) has alleviated both shortcomings of OT (statistical and computational). We detail here the three formulations of EOT: primal, dual and semi-dual *along with basic results which are the common base to the remainder of this thesis. Our thorough introduction, both theoretical and algorithmic, includes original contributions:*

(i) the regularization of OT using relative entropy with respect to the product measure of the marginals, which allows to have a dual formulation as an expectation useful to derive statistical properties in Chapter 3 and new solvers in Chapter 4,

(ii) the semi-dual formulation and some key properties, which are exploited in Chapter 4,

(iii) a generalization of the proof of existence of solutions to the dual problem,

(iv) an extension of our results to regularizers other than entropy and unbalanced OT.

1 Introduction

Comparing probability distributions is a fundamental issue arising in many machine learning problems, both supervised and unsupervised. In *unsupervised* machine-learning, one of the most popular research areas which emerged in recent years is learning generative models (Goodfellow et al., 2014). The goal is to fit the distribution of a parametric generative model to the unknown distribution induced by the dataset, to then be able to generate new samples which resemble the ones in the dataset. Choosing the right loss to be minimized between these two distributions is one of the key issues of the problem. On the *supervised* side of things, when one wants to learn a classifier for instance, choosing a meaningful distance on the data space is crucial. Many types of data can be represented as histograms, for instance: bag-of-visual-words comparison in computer vision (Rubner et al., 2000), color and shape processing in computer graphics (Solomon et al., 2015), bag-of-words for natural language processing (Kusner et al., 2015) and multi-label classification (Frogner et al., 2015). Normalized histograms are no more than finite discrete probability distributions, thus a good distance to compare histograms requires a good distance on measures.

Previous Works. In machine learning, the first candidates were φ-divergences, which can be seen as a weighted average (by φ) of the odds-ratio between the two measures (Csiszár, 1975). Their computational simplicity made them very popular, although they suffer from the major drawback of being oblivious to geometry, and they do not metrize weak-convergence. The latter is solved by Integral Probability Metrics (IPMs) (Müller, 1997), of which Maximum Mean Discrepancies (Gretton et al., 2006) are the most popular instance in machine learning applications as they can be computed efficiently in closed form with samples of the two measures. Another class of losses are Optimal Transport (OT) based losses – of which the Wasserstein Distance is a particular case. They behave particularly well in problems that are intrinsically geometric (e.g. shapes or image processing). However, they are expensive to compute and suffer from a curse of dimensionality, meaning their approximation from sampled measures degrades quickly in high dimension. A solution to the computational issue was introduced in Cuturi (2013), thanks to the regularization of the original OT problem with entropy. It allows to derive an efficient solver for finite discrete measures, and the resulting distance happens to perform well in various machine learning tasks as proved in this seminal paper.

Contributions. The object of this thesis is to showcase that the benefits of entropy-regularized OT extend far beyond fast algorithms for finite discrete measures. Before giving both theoretical and empirical evidence that it solves both the computational (Chapters 2 and 4) and statistical (Chapter 3) burdens of OT, we exhibit in this re-

view chapter the basics of regularized OT which are exploited in the remainder of this thesis: primal, dual, semi-dual formulations, existence of solutions, convergence of the regularized problem. We also provide a detailed account on Sinkhorn's algorithm, the state-of-the-art solver for discrete entropy-regularized OT. This chapter is different from the subsequent ones, as it is a collection of existing results, including some original contributions on regularized OT. They consist in *(i)* the regularization of OT using relative entropy with respect to the product measure of the marginal (instead of entropy with respect to the uniform measure in (Cuturi, 2013)), which allows us to formulate the dual problem as the maximization of an expectation – useful to derive statistical properties in Chapter 3 and new solvers in Chapter 4, *(ii)* the semi-dual formulation and its key properties, which are exploited in Chapter 4, *(iii)* and to a lesser extent a proof of existence of solutions to the dual problem for arbitrary measures, building on Franklin and Lorenz (1989), which provides a proof in the discrete setting and Chen et al. (2016) which provides a proof in the continuous case for Schrödinger's problem, which shares strong links with OT. These contributions were originally given in (Genevay et al., 2016) and (Genevay et al., 2019 (to appear) (on which Chapters 4 and 3 are respectively based), but it seems more natural to add them to the collection of the results used in subsequent chapters, to provide a unified and thorough introduction to regularized OT. Eventually, another contribution of this chapter is *(iv)* the extension of our results for regularizers other than entropy, and links with unbalanced transport (Chizat et al., 2018) whenever possible, which was not previously done in published work.

2 Distances Between Probability Measures and Weak-Convergence

This section introduces three types of discrepancies between measures, which are not all distances strictly-speaking, but they all define some sort of closeness between probability measures. We review $\varphi-$divergences, Maximum Mean Discrepancy(MMD) (which comes from the larger class of Integral Probability Metrics) and Optimal Transport (OT) distances of which the Wasserstein Distance is a special instance, as these are all popular losses in machine learning problems.

2.1 φ-divergences

The simplest tool to compare two measures are φ-divergences. Roughly speaking, they compare $\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}(x)$ to 1 trough the following formulation:

Definition 1. ***(φ-divergence)*** *(Csiszár, 1975) Let φ be a convex, lower semi-continuous function such that $\varphi(1) = 0$.*

Table 1.1 – Examples of φ-divergences

Kullback-Leibler	$D_{KL}(\alpha\|\beta) = \int_{\mathcal{X}} \log(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}(x))\mathrm{d}\beta(x)$	$\leftrightarrow$	$\varphi(x) = x\log(x)$
Jensen-Shannon	$D_{JS}(\alpha\|\beta) = D_{KL}(\alpha\|\frac{1}{2}(\alpha+\beta))$ $+D_{KL}(\beta\|\frac{1}{2}(\alpha+\beta)))$	$\leftrightarrow$	$\varphi(x) = x\log(x)$ $-(1+x)\log\frac{1+x}{2}$
Hellinger	$D_{H^2}(\alpha\|\beta) = \int_{\mathcal{X}}(\sqrt{d\alpha} - \sqrt{d\beta})^2$	$\leftrightarrow$	$\varphi(x) = (\sqrt{x}-1)^2$
Total Variation	$D_{TV}(\alpha\|\beta) = \sup_{A\in\mathcal{B}(\mathcal{X})} \lvert\alpha(A) - \beta(A)\rvert$	$\leftrightarrow$	$\varphi(x) = \frac{1}{2}\lvert x-1\rvert$

The φ-divergence D_φ *between two probability measures* α *and* $\beta \in \mathcal{M}_+^1$ *is defined by:*

$$D_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}(x)\right)\mathrm{d}\beta(x) + \varphi_\infty \alpha^\perp(\mathcal{X}),$$

where and $\varphi_\infty \overset{\text{def.}}{=} \lim_{x\to+\infty}\frac{\varphi(x)}{x}$ *and* $\alpha^\perp(\mathcal{X})$ *denotes the mass of the part of* α *that is not absolutely continuous with respect to* β *in the Radon-Nikodym decomposition of* α*, i.e.* $\alpha = \frac{\mathrm{d}\alpha}{\mathrm{d}\beta}(x)\beta + \alpha^\perp$.
Besides, D_φ *is jointly convex in both variables and if* φ *is strictly convex at* 1 *then* D_φ *is non-negative i.e.*

$$D_\varphi(\alpha|\beta) \geqslant 0 \quad \textit{and} \quad D_\varphi(\alpha|\beta) = 0 \Leftrightarrow \alpha = \beta.$$

The best-known φ-divergence is the so-called *Kullback-Leibler divergence* (see Table 1.1 for examples), which is widely used in machine learning problems (see Chapter 2 for an overview of learning with φ-divergences). However, it is equal to $+\infty$ if both measures do not share the same support, which causes discontinuity issues. For instance, consider the case on $\mathbb{R}$ where $\alpha = \delta_0$ a Dirac mass in 0 and $\alpha_n = \delta_{1/n}$ a Dirac mass in $1/n$. Then $D_{KL}(\alpha_n|\alpha) = +\infty$ for all n, although it would seem natural to say that when n goes to infinity, α_n gets closer to α. As for D_{TV} (which is a norm) and D_{H^2} (which is the square of a distance), they are both finite constants for all n, when considering this same case. This issue is simply an illustration of the fact that φ-divergences do not metrize weak-convergence.

Definition 2. ***(Weak-convergence)*** *We say that a sequence of measures* $(\alpha_n)_n$ weakly converges *to* α *(or converges in law) if*

$$\int_{\mathcal{X}} f(x)d\alpha_n(x) \to \int_{\mathcal{X}} f(x)d\alpha(x) \quad \forall f \in \mathcal{C}_b(\mathcal{X}), \tag{2.1}$$

where $\mathcal{C}_b(\mathcal{X})$ *denotes the set of continuous bounded functions on* $\mathcal{X}$.

We say that a discrepancy d metrizes *the weak-convergence of measures if*

$$d(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha,$$

where $\rightharpoonup$ denotes weak-convergence (or convergence in law, for $X_n \rightharpoonup X$ where $X_n \sim \alpha_n$ and $X \sim \alpha$).

Remark 1. As pointed out in Sec. 5.1 of (Ambrosio et al., 2006), it is sufficient to check (2.1) on any subset Ω of bounded continuous functions whose linear envelope $span(\Omega)$ is uniformly dense (i.e. dense in the uniform topology induced by the infinity norm) in $\mathcal{C}_b(\mathcal{X})$.

The fact that $\varphi-$divergences do not metrize weak convergence is a major issue and makes them poor candidates for learning problems, in spite of their appreciated computational simplicity. We discuss this in details in Chapter 2, where focus on finding a good notion of distance between measures to fit a (generative) parametric model to a dataset. For now, let us introduce another class of distances between measures which can metrize weak convergence under some assumptions.

2.2 Integral Probability Metrics and Maximum Mean discrepancy

The notion of Integral Probability Metrics (IPMs) was introduced by (Müller, 1997) as a class of maximization problems on certain sets of functions, regrouping some well known distances:

Definition 3. ***(Integral probability metrics)*** *(Müller, 1997) Consider two probability distributions α and β on a space $\mathcal{X}$. Given a set of measurable functions $\mathcal{F}$, the* integral probability metric *$d_{\mathcal{F}}$ is defined as*

$$d_{\mathcal{F}}(\alpha, \beta) \overset{\text{def.}}{=} \sup_{f \in \mathcal{F}} |\mathbb{E}_\alpha(f(X)) - \mathbb{E}_\beta(f(Y))|. \tag{2.2}$$

Let us now give a sufficient condition on $\mathcal{F}$ so that the associated IPM metrizes weak convergence:

Proposition 1. *If $span(\mathcal{F})$ is uniformly dense in $\mathcal{C}_b(\mathcal{X})$, then $d_{\mathcal{F}}$ metrizes weak convergence.*

Proof. $d_{\mathcal{F}}$ metrize weak convergence if and only if $d_{\mathcal{F}}(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha$. Using the definition of $d_{\mathcal{F}}$ (2.2) and the definition of weak convergence (2.1) we can rewrite this as:

$$\sup_{f \in \mathcal{F}} \Big| \int_{\mathcal{X}} f(x) \mathrm{d}\alpha_n(x) - \int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) \Big| \to 0$$
$$\Leftrightarrow \Big| \int_{\mathcal{X}} f(x) \mathrm{d}\alpha_n(x) - \int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) \Big| \to 0 \quad \forall f \in \mathcal{C}_b(\mathcal{X}).$$

Table 1.2 – Examples of Integral Probability Metrics

Total Variation	$\mathcal{F} = \{f \| \|f\|_\infty \leqslant 1\}$	functions upper-bounded by 1
Maximum Mean discrepancy	$\mathcal{F} = \{f \| \|f\|_{\mathcal{H}} \leqslant 1\}$	unit ball of the RKHS $\mathcal{H}$
Wasserstein-1	$\mathcal{F} = \{f \| \|f\|_{Lip} \leqslant 1\}$	functions with Lipschitz constant smaller than 1

Besides,

$$\sup_{f\in\mathcal{F}} \Big| \int_{\mathcal{X}} f(x)\mathrm{d}\alpha_n(x) - \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x)\Big| \to 0$$
$$\Leftrightarrow \Big| \int_{\mathcal{X}} f(x)\mathrm{d}\alpha_n(x) - \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x)\Big| \to 0 \quad \forall f \in \mathcal{F}.$$

Remark 1 yields the desired conclusion. □

We give some examples of well-known IPMs in Table 1.2. The Wasserstein-1 distance, which is the IPM for the set of 1-Lipschitz functions can be reformulated using Kantorovich-Rubinstein duality as:

$$W_1(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|_2 \mathrm{d}\pi(x,y),$$

where $\Pi(\alpha,\beta)$ is the set of probability distributions over the product set $\mathcal{X}\times\mathcal{X}$ with marginals α and β. This formulation is known as an optimal transport problem between α and β with cost function $c(x,y) = \|x-y\|_2$. It is well known that W_1 metrizes weak convergence of measures. We will get back to a more general definition of Wasserstein distance with other cost functions in the following section, since it extends beyond the frame of IPMs. As for TV, which is both a $\varphi-$divergence and an IPM, it does not metrize weak convergence: convergence in TV implies weak-convergence but not the other way around. We now focus on Maximum Mean Discrepancy for a while. Maximum Mean Discrepancies are IPMs on the unit ball of a *Reproducing Kernel Hilbert Space (RKHS)*, where the norm is the one induced by its *kernel function* k. The fact that MMDs metrize weak convergence requires some conditions on the kernel k. Let us start by introducing these concepts in more detail:

Definition 4. ***(Reproducing Kernel Hilbert Space)*** *Consider a Hilbert space $\mathcal{H}$ of real-valued functions on a space $\mathcal{X}$. Let L_x be the evaluation operator, such that $L_x(f) \overset{\text{def.}}{=} f(x)$. Then $\mathcal{H}$ is a* Reproducing Kernel Hilbert Space *if and only if L_x is continuous.*

From this definition, the role of the *reproducing kernel* in the *reproducing kernel Hilbert space* is not obvious. We first give the definition of a reproducing kernel.

Definition 5. ***(Reproducing Kernel)*** *Consider a Hilbert space $\mathcal{H}$ of real-valued functions on a space $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a* reproducing kernel of $\mathcal{H}$ *if it verifies:*

1. $\forall x \in \mathcal{X}, \quad k(x, \cdot) \in \mathcal{H}$,
2. $\forall f \in \mathcal{H}, \quad f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}}$.

Proposition 2. *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a* reproducing kernel *if and only if it is positive definite,* i.e *for all $(x_1, \ldots, x_n) \in \mathcal{X}^n$, $(a_1, \ldots, a_n) \in \mathbb{R}^n$,*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) \geqslant 0.$$

We can now state a theorem giving an equivalent definition for RKHS.

Theorem 5. *A Hilbert space $\mathcal{H}$ of real-valued functions on a space $\mathcal{X}$ is a Reproducing Kernel Hilbert Space if and only if it has a reproducing kernel. Besides, this reproducing kernel is unique.*

Thanks to this theorem, it is possible to define the RKHS associated to any positive definite kernel k.

Remark 2. The proof of any RKHS having a reproducing kernel is made thanks to Riesz representer theorem. Since the evaluation function is linear and continuous, there exists a function $k_x \in \mathcal{H}$ such that $f(x) = \langle f, k_x\rangle_{\mathcal{H}}$. Defining the bilinear function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by $k(x, y) = k_x(y)$ we clearly have that k is a reproducing kernel of $\mathcal{H}$.

The reproducing property of RKHS allows to derive a much simpler expression for their associated IPM, which becomes a closed form formula.

Proposition 3. ***(Maximum Mean discrepancy)*** *(Gretton et al., 2006) Consider two probability measures α and $\beta \in \mathcal{M}_+^1(\mathcal{X})$. Then, denoting by MMD_k the Maximum Mean Discrepancy on the Reproducing Kernel Hilbert Space $\mathcal{H}$ with kernel k, we have that*

$$\begin{aligned} MMD_k^2(\alpha, \beta) &\overset{\text{def.}}{=} \left(\sup_{\{f \,|\, \|f\|_{\mathcal{H}} \leqslant 1\}} |\mathbb{E}_\alpha(f(X)) - \mathbb{E}_\beta(f(Y))|\right)^2 \\ &= \mathbb{E}_{\alpha\otimes\alpha}[k(X, X')] + \mathbb{E}_{\beta\otimes\beta}[k(Y, Y')] - 2\mathbb{E}_{\alpha\otimes\beta}[k(X, Y)]. \end{aligned} \tag{2.3}$$

Proof. Using the fact that any function f in the RKHS satisfies $f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}}$, we can rewrite MMD as follows:

$$\begin{aligned} \sup_{f \,|\, \|f\|_{\mathcal{H}} \leqslant 1} |\mathbb{E}_\alpha(f(X)) - \mathbb{E}_\beta(f(Y))| &= \sup_{f \,|\, \|f\|_{\mathcal{H}} \leqslant 1} |\mathbb{E}_\alpha(\langle f, k(X, \cdot)\rangle_{\mathcal{H}}) - \mathbb{E}_\beta(\langle f, k(Y, \cdot)\rangle_{\mathcal{H}})| \\ &= \sup_{f \,|\, \|f\|_{\mathcal{H}} \leqslant 1} |\langle f, \mathbb{E}_\alpha k(X, \cdot) - \mathbb{E}_\beta k(Y, \cdot)\rangle_{\mathcal{H}}| \\ &\leqslant \|\mathbb{E}_\alpha k(X, \cdot) - \mathbb{E}_\beta k(Y, \cdot)\|_{\mathcal{H}}, \end{aligned}$$

and this upper bound is reached for $f = \mathbb{E}_\alpha k(X, \cdot) - \mathbb{E}_\beta k(Y, \cdot)$.

□

When α and β are finite discrete measures, i.e. $\alpha \overset{\text{def.}}{=} \sum_{i=1}^n \boldsymbol{\alpha}_i \delta_{x_i}$ and $\beta \overset{\text{def.}}{=} \sum_{i=1}^n \boldsymbol{\beta}_i \delta_{y_i}$, (2.3) becomes

$$\sum_{i,j=1}^n k(x_i, x_j)\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j + \sum_{i,j=1}^n k(y_i, y_j)\boldsymbol{\beta}_i\boldsymbol{\beta}_j - 2\sum_{i,j=1}^n k(x_i, y_j)\boldsymbol{\alpha}_i\boldsymbol{\beta}_j.$$

Thus, MMD can be efficiently estimated with samples from α and β. We discuss this in Chapter 3 when we compare sample complexity for MMD, Wasserstein distance, and entropy-regularized optimal transport.

We now give some conditions on k to ensure that MMD_k metrizes weak convergence

Theorem 6. ***(MMD and weak convergence)*** *(Sriperumbudur et al., 2010) Consider Maximum Mean Discrepancy with kernel k between two measures α and β on some space $\mathcal{X}$, as defined in (2.3).*

(i) Let $\mathcal{X}$ be a compact space. If the kernel k is universal *(i.e. its associated RKHS is dense in the space of continuous functions), then MMD_k metrizes weak convergence on $\mathcal{M}_+^1(\mathcal{X})$.*

(ii) Let $\mathcal{X} = \mathbb{R}^d$ and $k(x, y) = \kappa(x - y)$ where κ is a bounded strictly positive-definite function. If $\exists l \in \mathbb{N}$ such that:

$$\int_{\mathbb{R}^d} \frac{1}{\hat{\kappa}(\omega)(1 + \|\omega\|_2)^l} < \infty,$$

then MMD_k metrizes weak convergence on $\mathcal{M}_+^1(\mathcal{X})$.

The most widely used kernel is the Gaussian kernel $k(x, y) = \exp(\frac{-\|x-y\|^2}{\sigma^2})$, which is a *universal kernel*. According to Theorem 6 it metrizes weak convergence on a compact set, but it does not verify the required hypotheses on $\mathbb{R}^d$. They are however *characteristic*, meaning that $MMD_k(\alpha, \beta) = 0 \leftrightarrow \alpha = \beta$. An example of kernels that verify the hypotheses of Theorem 6 on $\mathbb{R}^d$ are the so-called Matern kernels, whose associated RKHS are Sobolev spaces. We further discuss the use of various kernels for learning problems in Chapter 2 and for function estimation in Chapters 3 and 4.

2.3 Optimal Transport

We consider two probability measures $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and β on $\mathcal{M}_+^1(\mathcal{Y})$. The Kantorovich formulation (Kantorovich, 1942) of *Optimal Transport* (OT) between α and β is defined by:

$$W_c(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\mathrm{d}\pi(x, y), \qquad (\mathcal{P})$$

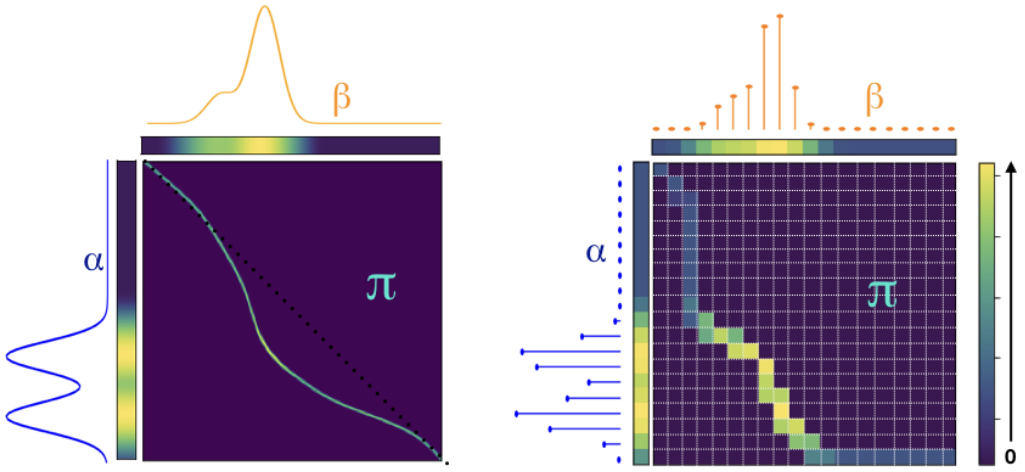


Figure 1.1 – Illustration of optimal transport between two measures α and β in the continuous case (left) and discrete case (right). In the continuous case, the transport plan is a probability distribution on $\mathcal{X} \times \mathcal{Y}$ while in the discrete case it is a matrix. For the latter, each entry π_{ij} corresponds to how much mass is moved from i to j.

where the feasible set is composed of probability distributions over the product space $\mathcal{X} \times \mathcal{Y}$ with fixed marginals α, β:

$$\Pi(\alpha, \beta) \overset{\text{def.}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \; ; \; P_{1\sharp}\pi = \alpha, P_{2\sharp}\pi = \beta \right\},$$

where $P_{1\sharp}\pi$ (resp. $P_{2\sharp}\pi$) is the marginal distribution of π for the first (resp. second) variable, using the projection maps $P_1(x, y) = x; P_2(x, y) = y$ along with the push-forward operator $\sharp$.

An optimizer π is called the *transport plan* between α and β, and quantifies how mass is optimally moved from α to β, see Figure 1.1. The cost function c represents the cost to move a unit of mass from x to y, and $W_c(\alpha, \beta)$ represents the total cost of moving all mass from α to β.

Remark 3 (p-Wasserstein distance). When $\mathcal{X} = \mathcal{Y}$ is endowed with a distance $d_{\mathcal{X}}$, choosing $c(x, y) = d_{\mathcal{X}}(x, y)^p$ where $p \geqslant 1$ yields the p-th power of the *p-Wasserstein distance*. It defines an actual distance between probability measures, which metrizes the weak-convergence.

For other cost functions c, $W_c(\alpha, \beta)$ is not necessarily a distance, since it does not always satisfy the triangle inequality but it still symmetric and positive under natural assumptions on the cost function (e.g. $c(x, y) = 0 \Leftrightarrow x = y, c(x, y) \geqslant 0$).

Optimal transport is a powerful tool to capture the underlying geometry of the measures, by relying on the cost function c which encodes the geometry of the space $\mathcal{X}$, and they have the ability to make meaningful comparisons even when the supports of the measures do not overlap (which is not the case for Kullback-Leibler divergence for

instance). Besides, the transport plan π gives a mapping between measures which can be used for instance in domain adaptation (Courty et al., 2014). More structure can be enforced with extensions of OT (Alvarez-Melis et al., 2017), which can for instance take into account labels of the data in supervised learning. However, OT suffers from a computational and statistical burden:

- **Computing OT is costly**: Solving OT when dealing with discrete distributions (i.e., finite weighted sums of Dirac masses) amounts to solving a large-scale linear program. This can be done using network flow solvers, which can be further refined to assignment problems when comparing measures of the same size with uniform weights (Burkard et al., 2009). The computational complexity is $O(n^3 log(n))$ where n is the number of points in the discrete measure (see also the monograph on Computational OT by Peyré et al. (2017) for a detailed review of OT solvers).

- **OT suffers from a curse of dimensionality**: considering a probability measure $\alpha \in \mathcal{M}_+^1(\mathbb{R}^d)$ and its empirical estimation $\hat{\alpha}_n$, we have $\mathbb{E}[W_p(\alpha, \hat{\alpha}_n)] = O(n^{-1/d})$ (see (Weed and Bach, 2017) for refined convergence rates depending on the support of α). Thus the error made when approximating the Wasserstein distance from samples grows exponentially fast with the dimension of the ambient space.

These two issues have caused OT to be neglected in machine learning applications for a long time in favor of simpler $\varphi-$divergences or MMD.

Let us conclude this section on OT with a recent extension introduced in (Chizat et al., 2018), (Liero et al., 2018). While OT is restricted to positive measures of mass 1, *Unbalanced Optimal Transport* can compare any two arbitrary positive measures. The marginal constraints are relaxed, as they are replaced with φ-divergences:

Definition 6. ***(Unbalanced Optimal Transport)*** *Consider two positive measures $\alpha \in \mathcal{M}_+(\mathcal{X})$ and $\beta \in \mathcal{M}_+(\mathcal{Y})$. Unbalanced Optimal Transport is defined as the following minimization problem*

$$\min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + D_{\psi_1}(P_{1\sharp}\pi|\alpha) + D_{\psi_2}(P_{2\sharp}\pi|\beta), \tag{2.4}$$

where ψ_1 and ψ_2 are positive, lower-semi-continuous functions such that $\psi_1(1) = 0$ and $\psi_2(1) = 0$.

Note that there is no constraint on the transport plan besides positivity: it is not required to have marginals equal to α and β nor to have mass 1. For specific choices of c, ψ_1, ψ_2, unbalanced OT defines a distance on $\mathcal{M}_+(\mathcal{X})$. This extension is popular in several applications due to the fact that it can compare any arbitrary positive measures. Whenever possible, we extend our results on regularized OT to the unbalanced case.

3 Regularized Optimal Transport

We introduce regularized optimal transport, which consists in regularizing the original problem by penalizing it with the φ-divergence of the transport plan with respect to the product measure:

$$W_{c,\varepsilon}^{\varphi}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y), \tag{$\mathcal{P}_{\varepsilon,\varphi}$}$$

where φ is a convex function with domain $\mathbb{R}^+$.

Entropic regularization, which is the main focus of this thesis, corresponds to the case $\varphi(w) = w\log(w) - w + 1$ (or alternatively $\varphi(w) = w\log(w)$) but one may choose the squared penalty $\varphi(w) = \frac{w^2}{2} + \iota_{\mathbb{R}^+}(w)$, where ι denotes the convex indicator function. However, most of the properties we derive for regularized optimal transport – in particular fast numerical solvers and improved sample complexity – are specific to the entropic regularization.

3.1 Dual Formulation

An advantage to consider regularized OT is to get an unconstrained dual problem. The dual of standard OT reads:

$$W_c(\alpha,\beta) = \sup_{(u,v)\in\mathcal{U}(c)} \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(x)\mathrm{d}\beta(y), \tag{$\mathcal{D}$}$$

where the constraint set $\mathcal{U}(c)$ is defined by

$$\mathcal{U}(c) \overset{\text{def.}}{=} \{(u,v) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) | u(x) + v(y) \leqslant c(x,y),\ \forall (x,y) \in \mathcal{X}\times\mathcal{Y}\}.$$

while the dual of regularized OT is given by an unconstrained maximization problem:

Proposition 4. *Consider OT between two probability measures α and β with a convex regularizer φ with domain $\mathbb{R}^+$. Then strong duality holds and ($\mathcal{P}_{\varepsilon,\varphi}$) is equivalent to the following dual formulation:*

$$\begin{aligned} W_{c,\varepsilon}^{\varphi}(\alpha,\beta) = \sup_{u,v\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} & \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(x)\mathrm{d}\beta(y) \\ & -\varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi^*\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y), \end{aligned} \tag{$\mathcal{D}_{\varepsilon,\varphi}$}$$

where φ^* is the Legendre transform of φ defined by $\varphi^*(p) \overset{\text{def.}}{=} \sup_w wp - \varphi(w)$.

Remark 4. (Strong Duality) Before getting into details on the derivation of the dual problem, note that strong duality holds, thanks to the application of Fenchel-Rockafellar theorem to the dual problem, which also guarantees existence of a primal solution to ($\mathcal{P}_{\varepsilon,\varphi}$)

(see (Chizat, 2017), Prop. 3.5.6 for technical details). The existence of maximizers for the dual problem ($\mathcal{D}_{\varepsilon,\varphi}$) is not guaranteed in general, and we give a proof of existence in the case of entropic regularization in Sec. 4.1 of this Chapter.

Proof. The primal problem reads

$$\min_{\pi} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right) \mathrm{d}\alpha(x)\mathrm{d}\beta(y))$$

under the constraint that $P_{1\sharp}\pi = \alpha, P_{2\sharp}\pi = \beta$ Introducing the Lagrange multipliers u and v associated to these constraints, the Lagrangian reads

$$\begin{aligned}\mathcal{L}(\pi,u,v) = &\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right) \mathrm{d}\alpha(x)\mathrm{d}\beta(y) \\ &+ \int_{\mathcal{X}} u(x)\left(\mathrm{d}\alpha(x) - \int_{\mathcal{Y}} \mathrm{d}\pi(x,y)\right) + \int_{\mathcal{Y}} v(y)\left(\mathrm{d}\beta(y) - \int_{\mathcal{X}} \mathrm{d}\pi(x,y)\right)\end{aligned}$$

The dual Lagrange function is given by $g(u,v) = \min_\pi \mathcal{L}(\pi,u,v)$ and thus rearranging terms we get

$$\begin{aligned} g(u,v) = &\int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(x)\mathrm{d}\beta(y) \\ &+\varepsilon \min_{\pi}\left(\int_{\mathcal{X}\times\mathcal{Y}}\left(\varphi\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right) - \frac{u(x)+v(y)-c(x,y)}{\varepsilon}\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y)\right) \\ &= \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(x)\mathrm{d}\beta(y) - \varepsilon\int_{\mathcal{X}\times\mathcal{Y}} \varphi^*\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y), \end{aligned}$$

where φ^* the Legendre transform of φ is given by

$$\varphi^*(p) = \sup_w wp - \varphi(w) = -\inf_w \varphi(w) - wp.$$

□

Remark 5. (Primal-Dual Relationship) The primal-dual relationship is given by

$$\begin{aligned} &\pi = \mathrm{argmin}_\pi\, \varphi(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}) - \frac{u(x)+v(y)-c(x,y)}{\varepsilon}\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)} \\ \Leftrightarrow\quad &\mathrm{d}\pi(x,y) = (\varphi')^{-1}(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})\mathrm{d}\alpha(x)\mathrm{d}\beta(y), \end{aligned}$$

when (φ') is invertible.

The smoothing effect of regularization is clear when looking at the dual of standard OT , since the constraint on the dual problem is replaced by a smooth penalization. The term $\int_{\mathcal{X}\times\mathcal{Y}} \varphi^*(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$ penalizes large positive values of $u(x)+v(y)-c(x,y)$. Ideally, to get a regularized problem that stays true to standard OT, we want φ^* to go quickly to large positive values when $u(x)+v(y)-c(x,y)$ grows. A good choice for

such a function is $\varphi^*(w) = e^w$, which actually corresponds to the entropic regularization (see section 4 for more details). A weaker penalization can also be considered using $\varphi^*(w) = \max(w, 0)^2/2$, which corresponds to the quadratic regularization.

3.2 The Case of Unbalanced OT

Unbalanced OT can also be regularized with a $\varphi-$divergence, which gives the following problem:

$$\min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + D_{\psi_1}(P_{1\sharp}\pi|\alpha) + D_{\psi_2}(P_{2\sharp}\pi|\beta) + \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right) \mathrm{d}\alpha(x)\mathrm{d}\beta(y).$$

As previously done with balanced OT, we can compute the dual of this problem:

Proposition 5. *The dual of regularized unbalanced OT with a convex regularizer φ with domain $\mathbb{R}^+$ is given by*

$$\sup_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} -\int_{\mathcal{X}} \psi_1^*(-u(x))\mathrm{d}\alpha(x) - \int_{\mathcal{Y}} \psi_2^*(-v(y))\mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi^*\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right) \mathrm{d}\alpha(x)\mathrm{d}\beta(y).$$

Proof. The proof is essentially similar to the derivation of the regularized dual in Proposition 4 except the problem is unconstrained. The primal problem reads

$$\min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \int_{\mathcal{X}} \psi_1\left(\frac{P_{1\sharp}\pi(x)}{\mathrm{d}\alpha(x)}\right) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} \psi_2\left(\frac{P_{2\sharp}\pi(y)}{\mathrm{d}\beta(y)}\right) \mathrm{d}\beta(y) + \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right) \mathrm{d}\alpha(x)\mathrm{d}\beta(y)).$$

We introduce slack variables a and b such that $a = P_{1\sharp}\pi$ and $b = P_{2\sharp}\pi$. This gives the following constrained problem:

$$\min_{\substack{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y}) \\ (a,b)\in\mathcal{M}_+(\mathcal{X})\times\mathcal{M}_+(\mathcal{Y})}} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \int_{\mathcal{X}} \psi_1\left(\frac{\mathrm{d}a(x)}{\mathrm{d}\alpha(x)}\right) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} \psi_2\left(\frac{\mathrm{d}b(y)}{\mathrm{d}\beta(y)}\right) \mathrm{d}\beta(y) + \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right) \mathrm{d}\alpha(x)\mathrm{d}\beta(y)),$$

subject to $a = P_{1\sharp}\pi$ and $b = P_{2\sharp}\pi$. Introducing the Lagrange multipliers u and v

associated to the constraints, the Lagrangian reads

$$\begin{aligned}\mathcal{L}(\pi, a, b, u, v) = &\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \int_{\mathcal{X}} \psi_1\left(\frac{\mathrm{d}a(x)}{\mathrm{d}\alpha(x)}\right)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} \psi_2\left(\frac{\mathrm{d}b(y)}{\mathrm{d}\beta(y)}\right)\mathrm{d}\beta(y)\\ &+ \int_{\mathcal{X}} u(x)\left(\mathrm{d}a(x) - \int_{\mathcal{Y}} \mathrm{d}\pi(x,y)\right) + \int_{\mathcal{Y}} v(y)\left(\mathrm{d}b(y) - \int_{\mathcal{X}} \mathrm{d}\pi(x,y)\right)\\ &+ \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y)).\end{aligned}$$

The dual Lagrange function is given by $g(u,v) = \min_{\pi,a,b} \mathcal{L}(\pi, a, b, u, v)$, and since the problem is separable we get three distinct minimization problems for each variable:

$$\begin{aligned}g(u,v) = &\min_a \Big[\int_{\mathcal{X}} u(x)\mathrm{d}a(x) + \psi_1\left(\frac{\mathrm{d}a(x)}{\mathrm{d}\alpha(x)}\right)\mathrm{d}\alpha(x)\Big]\\ &+ \min_b \Big[\int_{\mathcal{Y}} v(y)\mathrm{d}b(y) + \psi_2\left(\frac{\mathrm{d}b(x)}{\mathrm{d}\beta(y)}\right)\mathrm{d}\beta(x)\Big]\\ &+ \min_\pi \Big[\int_{\mathcal{X}\times\mathcal{Y}} (c(x,y) - u(x) - v(y))\mathrm{d}\pi(x,y)\\ &\qquad + \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y))\Big].\end{aligned}$$

The three minimization problems are actually the expression of the Legendre transform for ψ_1,ψ_2, and φ and so the dual function can be rewritten as:

$$\begin{aligned}g(u,v) = &-\int_{\mathcal{X}} \psi_1^*(-u(x))\mathrm{d}\alpha(x) - \int_{\mathcal{Y}} \psi_2^*(-v(y))\mathrm{d}\beta(y)\\ &- \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi^*\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y),\end{aligned}$$

where f^* is the Legendre transform of f defined by $f^*(p) = \sup_w wp - f(w)$. □

3.3 Dual Expectation Formulation

Another benefit of the regularization introduced above is the fact that is can be rewritten as the maximization of an expectation with respect to the product measure $\alpha \otimes \beta$

Proposition 6. *The dual of regularized OT ($\mathcal{D}_{\varepsilon,\varphi}$) has the following equivalent formulation:*

$$W^{\varphi}_{c,\varepsilon}(\alpha,\beta) = \sup_{u,v\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \mathbb{E}_{\alpha\otimes\beta}[f^{XY}_{\varepsilon}(u,v)],$$

where $f^{xy}_{\varepsilon} \overset{\text{def.}}{=} u(x) + v(y) - \varphi^*(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})$.

Since many machine learning problems (e.g. risk minimization) are formulated as the maximization of an expectation, this formulation of the dual of regularized OT allows us to apply well-known techniques from machine learning to study statistical properties

of regularized OT in Chapter 3 and use stochastic optimization to solve it in Chapter 4.

Note that the formulation as an expectation of the dual problem is only available for $\varepsilon > 0$. Indeed, the dual of standard OT has a constraint $(u + v - c \leqslant 0)$ whose indicator function cannot be put inside the expectation.

Remark 6. (Generalization to Unbalanced OT) Recall the dual of regularized unbalanced OT with regularizer φ:

$$\sup_{u,v} -\int_{\mathcal{X}} \psi_1^*(-u(x))\mathrm{d}\alpha(x) - \int_{\mathcal{Y}} \psi_2^*(-v(y))\mathrm{d}\beta(y) \\ -\varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi^*(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})\mathrm{d}\alpha(x)\mathrm{d}\beta(y).$$

Thus, it can also be cast as the maximization of an expectation with respect to the product measure $\alpha \otimes \beta$

$$\sup_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} \mathbb{E}_{\alpha\otimes\beta}\left[\psi_1^*(-u(X)) - \psi_2^*(-v(Y)) + \varphi^*(\frac{u(X)+v(Y)-c(X,Y)}{\varepsilon})\right].$$

4 Entropy-Regularized Optimal Transport

Entropic regularization is the main focus of this thesis, as it presents several specific properties:

- closed-form primal-dual relationship, allowing to recover the transport plan π after solving the simpler (unconstrained) dual problem,
- a fast numerical solver for finite discrete measures, Sinkhorn's algorithm (see Sec. 4.2),
- a discrepancy between measures interpolating between standard OT and MMD (see Chapter 2),
- an improved sample complexity compared to OT, breaking the curse of dimensionality for a regularization parameter large enough (see Chapter 3),
- reformulation of the dual as the maximization of an expectation in a Reproducing Kernel Hilbert Space (RKHS) ball of finite radius, allowing to solve the dual problem with a kernel version of stochastic gradient descent (see Chapter 3 and Chapter 4),
- semi-dual formulation ($\mathcal{S}_\varepsilon$), allowing to solve semi-discrete OT with online descent algorithms (see Chapter 4).

Let us rewrite the primal and dual problems ($\mathcal{P}_{\varepsilon,\varphi}$) and ($\mathcal{D}_{\varepsilon,\varphi}$) derived in Proposition 4 with the entropic regularization:

Proposition 7. *Consider OT between two probability measures α and β with entropic regularization:*

$$W_{c,\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon H(\pi|\alpha\otimes\beta), \tag{$\mathcal{P}_\varepsilon$}$$

where

$$H(\pi|\alpha\otimes\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \left(\log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right) - 1\right)\mathrm{d}\pi(x,y) + 1 \tag{4.1}$$

is the relative entropy of the transport plan π with respect to the product measure $\alpha\otimes\beta$. It is equivalent to this dual formulation:

$$\begin{aligned} W_{c,\varepsilon}(\alpha,\beta) = &\max_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\beta(y) \\ &- \varepsilon\int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon \qquad (\mathcal{D}_\varepsilon) \\ = &\max_{u\in\mathcal{C}(X),v\in\mathcal{C}(Y)} \mathbb{E}_{\alpha\otimes\beta}\left[f_\varepsilon^{XY}(u,v)\right] + \varepsilon \qquad (4.2) \end{aligned}$$

where $f_\varepsilon^{xy}(u,v) = u(x) + v(y) - \varepsilon e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}$.
Besides, the primal-dual relationship is given by

$$\mathrm{d}\pi(x,y) = \exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y).$$

Proof. This is a direct application of Proposition 4, using $\varphi(w) = w\log w - w + 1$, in which case $\varphi^*(p) \overset{\text{def.}}{=} \sup_w wp - \varphi(w) = e^p + 1$. Note that the sup is a max is this case, and we prove the existence of optimizers in Sec. 4.1 below. □

Remark 7. (Equivalent formulation of ($\mathcal{P}_\varepsilon$)) The transport plan π is constrained to be a probability measure which imposes $\int_{\mathcal{X}\times\mathcal{Y}}\mathrm{d}\pi(x,y) = 1$, so the primal problem ($\mathcal{P}_\varepsilon$) can be simplified to:

$$\min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right). \tag{$\mathcal{P}_\varepsilon$}$$

However when computing the dual directly from this formulation, we get a primal-dual relationship that is less elegant:

$$\mathrm{d}\pi(x,y) = e^{\frac{u(x)+v(y)-c(x,y)-1}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y),$$

which is why we prefer to formally state Proposition 7 with H defined in (4.1).

Remark 8. (Dual Potentials and Exponential Scalings) As commonly done in the literature on OT, we refer to the variables of the dual problem (u,v) as the *dual (Kantorovitch) potentials*. We will also often use the so-called *exponential scalings* of the dual variables (a,b) defined by $a \overset{\text{def.}}{=} e^{\frac{u}{\varepsilon}}$ and $b \overset{\text{def.}}{=} e^{\frac{v}{\varepsilon}}$.

Entropic regularization of optimal transport was first introduced with the following formulation of entropy: $H(\pi) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}x\mathrm{d}y}\right)\mathrm{d}\pi(x,y)$ (Cuturi, 2013). The resulting dual problem is slightly different:

$$\max_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\beta(y) - \varepsilon\int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}\mathrm{d}x\mathrm{d}y + \varepsilon.$$

Here, the third term in the dual is an integral with respect to the Lebesgue measure, while with relative entropy, the integral is taken with respect to the product measure $\alpha\otimes\beta$. The formulation with simple entropy yields an unconstrained dual problem which can be solved efficiently (see Proposition. 10 for details) but it can not be formulated as the maximization of an expectation which, as already mentioned, is crucial for the results presented in Chapters 3 and 4. Thus we only use relative-entropy as a regularizer, as it keeps all the benefits brought by simple entropy with the added benefit of the expectation formulation.

4.1 Solving the Regularized Dual Problem

As most of the methods we develop in this thesis rely on the dual formulation of entropy-regularized OT, we give a proof of existence of a solution to this problem, for a general setting. We discuss further the regularity of the dual potentials in Chapter 3. This section is dedicated to proving the following existence theorem:

Theorem 7. ***(Existence of a dual solution)*** *Consider the dual of entropy-regularized OT, with marginals $\alpha,\beta \in \mathcal{M}_+^1(\mathcal{X})\times\mathcal{M}_+^1(\mathcal{Y})$ supported on two subsets of $\mathbb{R}^d$, and with a cost function c bounded on $\mathcal{X}\times\mathcal{Y}$. Let $L^\infty(\alpha) \overset{\text{def.}}{=} \{f:\mathcal{X}\to\mathbb{R} | \exists C>0$ such that $f(x)\leqslant C$ α-a.e.$\}$. Then the dual problem has solutions $(u^*,v^*)\in L^\infty(\alpha)\times L^\infty(\beta)$ which are unique $\alpha-$ and $\beta-$a.e. up to an additive constant.*

It is straightforward to see that for any solution (u^*,v^*) to the dual problem, the pair (u^*+k,v^*-k) for $k\in\mathbb{R}$ is also a solution to the dual problem. Besides, modifying the values of u^* and v^* outside of the support of the measures does not have any effect on the value of the problem.

The proof of existence of a solution to the dual problem essentially amounts to rewriting the optimality condition as a fixed point equation, and proving that a fixed point exists. To do so, we show that the operator in the fixed point equation is a contraction for a certain metric, called the Hilbert metric. This proof is based on the same idea from that of the existence of a solution to Schrodinger's system (which shares strong links with regularized OT) in (Chen et al., 2016), inspired from the original proof of (Franklin and Lorenz, 1989) which deals with discrete regularized OT. We prove the existence of potentials in a general framework, as we consider arbitrary measures α and β and any bounded regular cost function c.

The dual problem is unconstrained, and it is jointly concave in both variables. Thus, we can fix one and optimize over the other, and the first order condition for u gives:

$$u(x) = -\varepsilon \log \left(\int_{\mathcal{Y}} e^{\frac{v(y)-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y) \right) \quad \text{for a.e. } x \in \mathcal{X}, \tag{4.3}$$

and similarly for v:

$$v(y) = -\varepsilon \log \left(\int_{\mathcal{X}} e^{\frac{u(x)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(y) \right) \quad \text{for a.e. } y \in \mathcal{Y}. \tag{4.4}$$

Remark 9. Although the optimality conditions (4.3) and (4.4) only fix the value of the optimal potentials (u^*, v^*) on the supports of α and β respectively, they allow to extrapolate the values of the potentials outside of this support.

4.1.1 Hilbert Metric

We start with a few definitions and properties of the Hilbert metric, which will be useful later on. Proof of these results can be found in (Bushell, 1973).

Definition 7. ***(Hilbert metric)*** *Consider $\mathcal{K}$ a closed solid cone on a real Banach space $\mathcal{B}$ i.e. $\mathcal{K}$ satisfies the 4 following properties:*

1. *the interior of $\mathcal{K}$ is not empty,*
2. $\mathcal{K} + \mathcal{K} \subseteq \mathcal{K}$,
3. $\alpha\mathcal{K} \subseteq \mathcal{K} \forall \alpha \geqslant 0$,
4. $\mathcal{K} \cap -\mathcal{K} = \{0\}$.

We use the partial order induced by the cone, meaning $x \leqslant y \Leftrightarrow y - x \in \mathcal{K}$, and define the following quantities

$$M(a,b) \overset{\text{def.}}{=} \inf\{\lambda | a \leqslant \lambda b\} \quad \textit{and} \quad m(a,b) \overset{\text{def.}}{=} \sup\{\lambda | a \leqslant \lambda b\} \quad \textit{for } a, b \in \mathcal{K}^+ \overset{\text{def.}}{=} \mathcal{K} \setminus \{0\}.$$

Then the Hilbert metric d_H on $\mathcal{K}$ is given by

$$d_H(a,b) \overset{\text{def.}}{=} \log \frac{M(a,b)}{m(a,b)}. \tag{4.5}$$

Note that the Hilbert metric is projective, meaning that it is invariant by multiplication by a positive factor: $d_H(a,b) = d_H(\alpha a, b) = d_H(a, \alpha b)$, $\forall\, \alpha > 0$.

The Hilbert metric is a pseudo-metric on the interior of the cone $\mathring{\mathcal{K}}$, and a metric on the restriction of $\mathring{\mathcal{K}}$ to the unit sphere:

Theorem 8. *$(\mathring{\mathcal{K}}, d_H)$ is a pseudo-metric space and $(\mathring{\mathcal{K}} \cap S(0,1), d_H)$ is a metric space, where $S(0,1)$ is the unit sphere in $\mathcal{B}$*

To use Banach's fixed point theorem on $(\mathring{\mathcal{K}} \cap B(0,1), d_H)$, we need to introduce the notion of contraction ratio:

Definition 8. *We say that an operator $\mathcal{E}$ is a* positive map *in the cone if $\mathcal{E}(\mathcal{K}^+) \subset \mathcal{K}^+$. For a positive map $\mathcal{E}$, we denote its* projective diameter *by*

$$\Delta(\mathcal{E}) \stackrel{\text{def.}}{=} \sup\{d_H(\mathcal{E}(a), \mathcal{E}(b)) \mid a, b \in \mathcal{K}^+\},$$

and its contraction ratio

$$\kappa(\mathcal{E}) \stackrel{\text{def.}}{=} \inf\{\lambda \mid d_H(\mathcal{E}(a), \mathcal{E}(b)) \leqslant \lambda d_H(x, y) \forall x, y \in \mathcal{K}^+\}.$$

In the case where the mapping is linear, we have a relation between the contraction ratio and the projective diameter.

Proposition 8. *Consider a* linear *positive map $\mathcal{E}$ on $\mathcal{K}$, then*

$$\kappa(\mathcal{E}) \leqslant \tanh\left(\frac{1}{4}\Delta(\mathcal{E})\right),$$

and $\Delta(\mathcal{E}) \leqslant 2\sup_a\{d_H(\mathcal{E}(a), 1)) \mid a \in \mathcal{K}^+\}$.

Since $|\tanh(x)| < 1$ for $|x| < +\infty$, this means that if the projective diameter of a positive mapping is finite, then it is a contraction. The proof of the first inequality is given in (Bushell, 1973) while the second is a direct application of the triangle inequality.

4.1.2 Fixed Point Theorem

Now let us rewrite the optimality condition as a fixed point equation. We consider the *exponential scalings* (a, b) of the dual variables (u, v). At optimality we have that

$$a(x) = \left(\int_{\mathcal{Y}} b(y) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y)\right)^{-1} \quad \text{and} \quad b(y) = \left(\int_{\mathcal{X}} a(x) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\right)^{-1}. \tag{4.6}$$

We define the operators $\varphi^{\varepsilon,\alpha}$ and $\varphi^{\varepsilon,\beta}$ such that

$$\varphi^{\varepsilon,\alpha}(f) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} f(x) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x) \quad \text{and} \quad \varphi^{\varepsilon,\beta}(f) \stackrel{\text{def.}}{=} \int_{\mathcal{Y}} f(y) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y), \tag{4.7}$$

and we denote by $\mathcal{E}$ the operator such that $\mathcal{E}(a) \stackrel{\text{def.}}{=} 1/a$.

Proposition 9. *The optimal exponential scalings (a^*, v^*) satisfy the following fixed-point equations:*

$$a^* = \Phi(a^*) \quad \textit{where} \quad \Phi \stackrel{\text{def.}}{=} \mathcal{E} \circ \varphi^{\varepsilon,\beta} \circ \mathcal{E} \circ \varphi^{\varepsilon,\alpha}, \tag{4.8}$$

and

$$b^* = \tilde{\Phi}(b^*) \quad \textit{where} \quad \tilde{\Phi} \stackrel{\text{def.}}{=} \mathcal{E} \circ \varphi^{\varepsilon,\alpha} \circ \mathcal{E} \circ \varphi^{\varepsilon,\beta}. \tag{4.9}$$

To prove the existence of solutions to (4.6) we first need to prove the following lemma

Lemma 1. *Consider the operators Φ defined in (4.8) and $\tilde{\Phi}$ defined in (4.9), and let $L_+^\infty(\mathcal{X}) \overset{\text{def.}}{=} \{a \in L^\infty(\mathcal{X}) \mid a(x) > 0, \forall x \in \mathcal{X}\}$. Then Φ and $\tilde{\Phi}$ are contractions on $L_+^\infty(\mathcal{X})$ with contraction ratio $\Delta(\Phi) \leqslant \tanh\left(\frac{1}{4}\log(e^{\frac{2\|c\|_\infty}{\varepsilon}})\right) < 1$.*

Proof. (Lemma 1) We consider the space of positive bounded functions $L_+^\infty(\alpha) \overset{\text{def.}}{=} \{f \in L^\infty(\alpha) | f(x) > 0 \forall x \in \mathcal{X}\}$ and $L_+^\infty(\beta)$. It is easy to check that it is a cone with non-empty interior and we can thus endow $L_+^\infty(\alpha)$ and $L^\infty(\beta)$ with Hilbert's metric. We also have that $\mathcal{E}$, $\varphi^{\varepsilon,\alpha}$ and $\varphi^{\varepsilon,\beta}$ are positive maps mapping L_+^∞ to itself, $L_+^\infty(\alpha)$ to $L_+^\infty(\beta)$ and $L_+^\infty(\beta)$ to $L_+^\infty(\alpha)$ respectively. To be able to use Banach's fixed point theorem, we restrict $L_+^\infty(\alpha)$ and $L_+^\infty(\beta)$ to the unit sphere, which is not a restriction per se as for any function a that verifies the fixed point equation, $a/\|a\|_\infty$ also verifies it.

To compute the contraction ratio of the composition Φ, we can simply compute the contraction ratio of each of the composing functions and multiply them to get the whole contraction ratio.

The inversion operator $\mathcal{E}$ is an isometry for Hilbert's metric:

$$d_H(\mathcal{E}(a), \mathcal{E}(b)) = \frac{\inf\{\lambda | 1/a \leqslant \lambda 1/b\}}{\sup\{\lambda | 1/a \leqslant \lambda 1/b\}} = \frac{\inf\{\lambda | a \leqslant \lambda b\}}{\sup\{\lambda | b \leqslant \lambda a\}} = d_H(b, a) = d_H(a, b).$$

We are left with computing the contraction ratio of $\varphi^{\varepsilon,\alpha}$ and $\varphi^{\varepsilon,\beta}$. Since they are both linear maps, we can instead consider the quantity $\sup_a\{d_H(\varphi^\varepsilon(a), 1) \mid a \in \mathcal{K}^+\}$ thanks to proposition 8. We focus on $\varphi^{\varepsilon,\alpha}$ as $\varphi^{\varepsilon,\beta}$ behaves the same way. We have that $\forall a \in L^\infty(\alpha)$

$$e^{\frac{-\|c\|_\infty}{\varepsilon}} \int_X a(x) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x) \leqslant \int_X a(x) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x) \leqslant e^{\frac{\|c\|_\infty}{\varepsilon}} \int_X f(x) \mathrm{d}\alpha(x),$$

and thus

$$\Delta(\varphi^{\varepsilon,\alpha}) \leqslant 2 \sup_a \left(\log \frac{\sup \varphi^{\varepsilon,\alpha}(a)}{\inf \varphi^{\varepsilon,\alpha}(a)}\right) \leqslant 2 \log\left(e^{\frac{2\|c\|_\infty}{\varepsilon}}\right) < \infty.$$

Combining all contraction ratios, we get $\Delta(\Phi) \leqslant \tanh\left(\frac{1}{4}\log\left(e^{\frac{2\|c\|_\infty}{\varepsilon}}\right)\right) < 1$ and thus Φ is a contraction for the Hilbert metric. □

Proof. (Theorem 7) Thanks to Lemma 1 and Proposition 9, we can conclude with Banach's fixed point theorem that Φ and $\tilde{\Phi}$ admit a unique fixed point in $(L_+^\infty(\mathcal{X}) \cap S(0,1))$ and complete the proof of Theorem 7. This implies the existence of unique exponential scalings (a^*, b^*) on the unit sphere, but any other pair $(ka^*, b^*/k)$ for $k \in \mathbb{R}^*$ satisfies the optimality conditions. Since dual potentials are essentially the log of these exponential scalings, we therefore have unicity of the potential scalings, up to an additive constant, instead of a multiplicative constant for the exponential scalings. □

The optimal dual potentials can be constructed as fixed points of a contractive map,

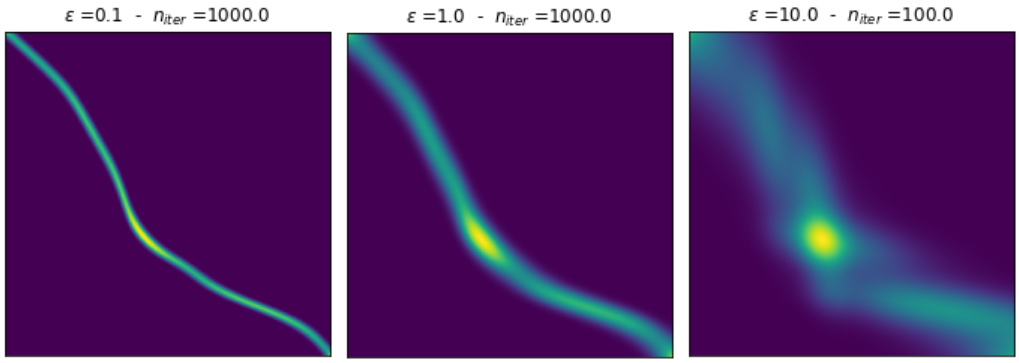


Figure 1.2 – Influence of the regularization parameter ε on the transport plan π computed with Sinkhorn's algorithm. Regularization tends to spread the transport plan, leading to a smoother solution.

which yields an algorithm to compute the potentials along with a speed of convergence for the iterates.

Corollary 1. *Let* $(a^{(\ell)}, b^{(\ell)}) = (\Phi^{(\ell)}(1), \tilde{\Phi}^{(\ell)}(1))$ *where* $\Phi^{(\ell)}$ *is the* ℓ*-fold composition of* Φ *defined in* (4.8)*. Then*

$$d_H(a^{(\ell)}, a^*) = O\left(\tanh\left(\frac{1}{4}\log(e^{\frac{2\|c\|_\infty}{\varepsilon}}\right)^{2\ell}\right).$$

Proof. This is a direct corollary of Banach's fixed point theorem, with the contracting ratio of the operator Φ being $\tanh\left(\frac{1}{4}\log(e^{\frac{2\|c\|_\infty}{\varepsilon}}\right)^2$. □

4.2 Sinkhorn's Algorithm

Since the dual problem is concave in each variable, a natural way to solve it is to iteratively optimize over each variable. In the discrete case, the first order conditions for each of the variables read:

$$\mathbf{u}_i = -\varepsilon \log\left(\sum_{j=1}^{m} e^{\frac{\mathbf{v}_j - c(x_i,y_j)}{\varepsilon}} \boldsymbol{\beta}_j\right) \quad \text{and} \quad \mathbf{v}_j = -\varepsilon \log\left(\sum_{i=1}^{n} e^{\frac{\mathbf{u}_i - c(x_i,y_j)}{\varepsilon}} \boldsymbol{\alpha}_i\right), \tag{4.10}$$

or, using the exponential scalings of the dual variables $\mathbf{a} \overset{\text{def.}}{=} e^{\frac{\mathbf{u}}{\varepsilon}}$ and $\mathbf{b} \overset{\text{def.}}{=} e^{\frac{\mathbf{v}}{\varepsilon}}$:

$$\mathbf{a}_i = \frac{1}{\sum_{j=1}^{m} \mathbf{b}_j e^{\frac{-c(x_i,y_j)}{\varepsilon}} \boldsymbol{\beta}_j} \quad \text{and} \quad \mathbf{b}_j = \frac{1}{\sum_{i=1}^{n} \mathbf{a}_i e^{\frac{-c(x_i,y_j)}{\varepsilon}} \boldsymbol{\alpha}_i.} \tag{4.11}$$

The algorithm corresponding to these alternating maximizations is usually called Sinkhorn's algorithm in the literature, although the denomination IPFP (Iterative Projection Fitting Procedure) can also be found. The latter can be understood as a primal resolution of the problem, consisting in iteratively projecting over each marginal con-

straint for the Kullback-Leibler divergence, but both approaches correspond to the same algorithm. Introducing the exponential scaling of the dual variables $\mathbf{a} \overset{\text{def.}}{=} e^{\frac{\mathbf{u}}{\varepsilon}}$ and $\mathbf{b} \overset{\text{def.}}{=} e^{\frac{\mathbf{v}}{\varepsilon}}$ as well as the exponential scaling of the cost matrix $\mathbf{K}$ such that $\mathbf{K}_{ij} = e^{-\frac{c(x_i,y_j)}{\varepsilon}}$ we can write the iterations with matrix-vector multiplications:

Proposition 10. ***(Sinkhorn Iterations)*** *Consider optimal transport between two finite discrete measures $\alpha \overset{\text{def.}}{=} \sum_{i=1}^n \boldsymbol{\alpha}_i \delta_{x_i}$ and $\beta \overset{\text{def.}}{=} \sum_{j=1}^m \boldsymbol{\beta}_j \delta_{y_j}$ with cost function c. Let $\mathbf{K}_{ij} = e^{-\frac{c(x_i,y_j)}{\varepsilon}}$. Then iterations given by*

$$\mathbf{a}^{(\ell+1)} = \frac{1}{\mathbf{K}(\mathbf{b}^{(\ell)} \odot \boldsymbol{\beta})} \quad \text{and} \quad \mathbf{b}^{(\ell+1)} = \frac{1}{\mathbf{K}^T(\mathbf{a}^{(\ell+1)} \odot \boldsymbol{\alpha})} \tag{4.12}$$

converge to $(\mathbf{a}^*, \mathbf{b}^*)$ the exponential scaling of a solution of the dual problem $\mathcal{D}_\varepsilon$. The optimal transport plan is recovered via the following formula

$$\boldsymbol{\pi}^* = \operatorname{diag}(\mathbf{a}^* \odot \boldsymbol{\alpha}) \mathbf{K} \operatorname{diag}(\mathbf{b}^* \odot \boldsymbol{\beta}),$$

where $\operatorname{diag}(a)$ is the diagonal matrix with vector a on the diagonal and 0 elsewhere.

The complexity of each iteration is $O(n^2)$ if both marginals have the same number of points n. This is a major improvement compared to $O(n^3 log(n))$ needed to solve the linear program induced by standard discrete OT.

Proposition 11. ***(Convergence rate of Sinkhorn Iterations)*** *(Franklin and Lorenz, 1989) Let $\mathbf{a}^{(\ell)}$ the ℓ-th iterate of Sinkhorn's algorithm, $\mathbf{a}^*$ the optimal exponential scaling, and $\boldsymbol{\pi}^{(\ell)} \overset{\text{def.}}{=} \operatorname{diag}(\mathbf{a}^{(\ell)}) K \operatorname{diag}(\mathbf{b}^{(\ell)})$. Then*

$$d_H(\mathbf{a}^{(\ell)}, \mathbf{a}^*) = O(\lambda(K)^{2\ell}) \quad \text{and} \quad d_H(\mathbf{a}^{(\ell)}, \mathbf{a}^*) \leqslant \frac{d_H(\boldsymbol{\pi}^{(\ell)} \mathbb{1}, \boldsymbol{\alpha})}{1 - \lambda(K)}, \tag{4.13}$$

where

$$\lambda(K) = \frac{\sqrt{\eta(K)} - 1}{\sqrt{\eta(K)} - 1} < 1 \qquad ; \qquad \eta(K) = \max_{i,j,k,l} \frac{K_{ik} K_{jl}}{K_{jk} K_{il}},$$

and the same rates hold for the other iterate $\mathbf{b}^{(\ell)}$.

Proof of convergence of the algorithm is a special case of the proof of existence of the dual potentials, using the Hilbert metric d_H on $\mathbb{R}^n$. Inequality (4.13) gives a useful insight on how to monitor convergence of the algorithm in practice, as the marginal constraint violation is an upper bound on the convergence of the exponential scalings. The convergence rate of this algorithm is illustrated in this manner in Figure 1.4. The negative influence of ε on the convergence rate is quite clear in the figure, although the asymptotic rate given in (4.13) is not often sharp in practice.

Remark 10. (Stabilizing Sinkhorn) The algorithm suffers from numerical instability when ε gets too small as some coefficients of the matrix K explode. This issue can

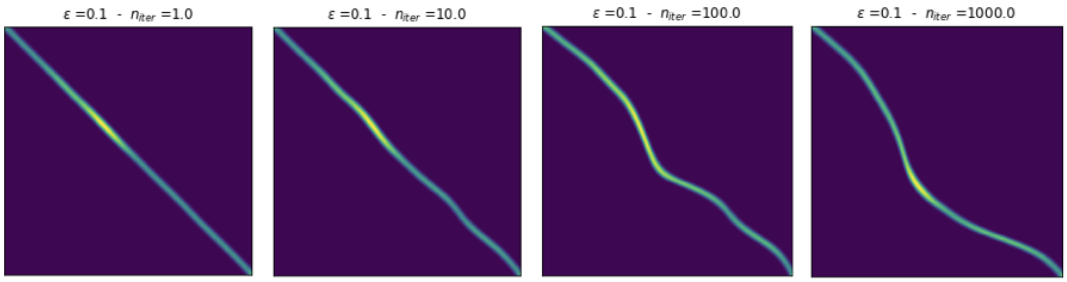


Figure 1.3 – Evolution of the transport plan with the number of iterations for Sinkhorn's algorithm. The algorithm is initialized with $\mathbf{b} = 1_n$ which corresponds to the initial transport plan $\boldsymbol{\pi}^{(0)}$ being the product of marginals.

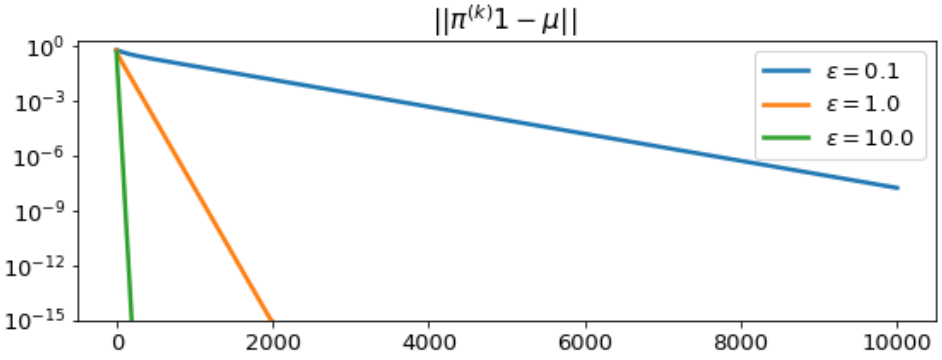


Figure 1.4 – Influence of the regularization parameter ε on the speed of convergence of Sinkhorn's algorithm. Convergence of the algorithm is monitored by looking at the constraint violation on the first marginal $\|\pi^{(\ell)}1 - \boldsymbol{\alpha}\|_1$. The convergence rate worsens dramatically when descreasing ε, and there is thus a tradeoff between getting a fast approximation of OT, or an accurate one.

be solved by writing iterations in the log domain, i.e. on the dual variables (u, v) instead of the exponential scalings (a, b), and by replacing the matrix K at each iteration using $\tilde{K}_{ij}^{(\ell)} \overset{\text{def.}}{=} \exp(\mathbf{u}_i^{(\ell)} + \mathbf{v}_j^{(\ell)} - C_{ij})$ which is numerically more stable. Besides, we can seen in figure 1.4 that convergence becomes slower for small ε. For some applications, getting close to standard optimal transport is important, and thus one might resort to *ε-scaling*. This heuristic consists in starting with a large regularization parameter ε and then rerunning the algorithm with slowly decreasing the value of the parameter with a warm start, reusing the values of a and b obtained with a larger regularization. More details can be found in (Schmitzer, 2016) (sec.3).

4.3 Semi-Dual Formulation

The equations (4.3) and (4.4) giving u a a function of v and conversely can be seen as smoothed versions of the c-transform which links the dual potentials in standard OT. The c-transform of a function v is given by

$$v^c(x) \overset{\text{def.}}{=} \min_{y \in \mathcal{Y}} c(x, y) - v(y). \tag{4.14}$$

In the regularized case, we introduce the c, ε-transform:

$$v^{c,\varepsilon}(x) \overset{\text{def.}}{=} -\varepsilon \log \left(\int_{\mathcal{Y}} e^{\frac{v(y) - c(x,y)}{\varepsilon}} \mathrm{d}\beta(y) \right). \tag{4.15}$$

Note that these smoothed $c-$transforms actually depend on β but we omit it in the notation. We can now derive a *semi-dual* formulation, which is a maximization problem over v only:

Proposition 12. *Consider OT between two probability measures α and β with entropic regularization. Then ($\mathcal{P}_\varepsilon$) is equivalent to the following semi-dual formulation:*

$$W_{c,\varepsilon}(\alpha, \beta) = \max_{v \in \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} v^{c,\varepsilon}(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y) \mathrm{d}\beta(y), \tag{$\mathcal{S}_\varepsilon$}$$

where $v^{c,\varepsilon}$ is the c, ε-transform of v defined in (4.15).

Proof. Replacing u by $v^{c,\varepsilon}$ in the dual, we get

$$W_{c,\varepsilon}(\alpha, \beta) = \int_{\mathcal{X}} v^{c,\varepsilon}(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y) \mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X} \times \mathcal{Y}} e^{\frac{v^{c,\varepsilon}(x) + v(y) - c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x) \mathrm{d}\beta(y) + \varepsilon.$$

Let us focus on the third term, which corresponds to the smooth constraint. We have that $e^{\frac{v^{c,\varepsilon}}{\varepsilon}} = \left(\int_{\mathcal{Y}} e^{\frac{v(y) - c(x,y)}{\varepsilon}} \mathrm{d}\beta(y) \right)^{-1}$ by definition of the c, ε-transform. And thus the terms inside the integral cancel out, leaving just $\int_{\mathcal{X}} \mathrm{d}\alpha(x)$ which il equal to 1 since α is a measure of mass 1. □

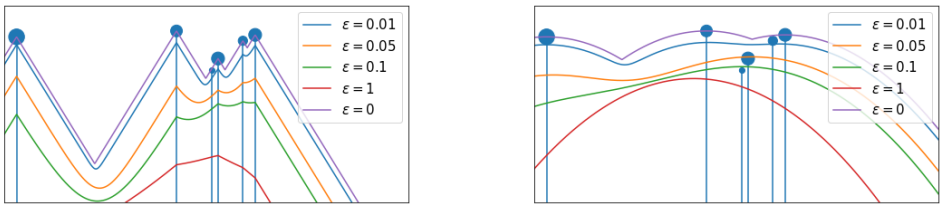


Figure 1.5 – Plot of $v^{c,\varepsilon}$, the c,ε-transform of a discrete vector $\mathbf{v}$ for various values of ε for Euclidean (left) and squared Euclidean (right) cost function c. The blue markers are plotted at $(y_i, \mathbf{v}_i)$ and their diameter is proportional to $\boldsymbol{\beta}_i$.

4.3.1 Case of a Discrete Measure

The semi-dual formulation is mostly interesting in the case where one of the measures is discrete. For instance, if β is a discrete measure $\beta \overset{\text{def.}}{=} \sum_{i=1}^n \boldsymbol{\beta}_i \delta_{y_i}$, the associated dual potential $\mathbf{v}$ is a vector in $\mathbb{R}^n$ and its c,ε-transform is given by $v^{c,\varepsilon}(x) = -\varepsilon \log\left(\sum_{i=1}^n e^{\frac{\mathbf{v}_i - c(x,y_i)}{\varepsilon}} \boldsymbol{\beta}_i\right)$. Thus the semi-dual problem becomes

$$W_{c,\varepsilon}(\alpha,\beta) = \max_{\mathbf{v}\in\mathbb{R}^n} \int_{\mathcal{X}} -\varepsilon \log\left(\sum_{i=1}^n e^{\frac{\mathbf{v}_i - c(x,y_i)}{\varepsilon}} \boldsymbol{\beta}_i\right) \mathrm{d}\alpha(x) + \sum_{i=1}^n \mathbf{v}_i \boldsymbol{\beta}_i. \qquad (\mathcal{S}_\varepsilon)$$

A modified version of the well-known *log-sum-exp* appears in the smooth c,ε-transform in lieu of the max in the c-transform. Here we have a dependence on $\boldsymbol{\beta}$, while the *log-sum-exp* is usually defined by $LSE(w_1,\dots,w_n) \overset{\text{def.}}{=} \log\left(\sum_{i=1}^n e^{w_i}\right)$ while the LSE appears for instance in logistic-regression and is known to be a smooth, convex approximation of the max function. The approximation gets better as the deviations in the w_i get larger. Thus when ε gets small, the values of $\frac{\mathbf{v}_i - c(x,y_i)}{\varepsilon}$ get larger and their deviations increase as well, making the c,ε-transform a sharper approximation of the c-transform. This is illustrated by figure 1.5 which displays the c,ε-transform of a discrete vector $\mathbf{v}$ for various values of ε for a cost function c that is the Euclidean or squared Euclidean norm .

4.3.2 Semi-Dual Expectation Formulation

The semi-dual problem can also be formulated as the maximization of an expectation, with respect to one of the marginals:

$$W_{c,\varepsilon}(\alpha,\beta) = \max_{\mathbf{v}\in\mathbb{R}^n} \mathbb{E}_\alpha[g_\varepsilon^X(\mathbf{v})], \qquad (\mathcal{S}_\varepsilon)$$

where

$$g_\varepsilon^x(\mathbf{v}) \overset{\text{def.}}{=} -\varepsilon \log\left(\sum_{i=1}^n e^{\frac{\mathbf{v}_i - c(x,y_i)}{\varepsilon}} \beta_i\right) + \sum_{i=1}^n \mathbf{v}_i \boldsymbol{\beta}_i. \qquad (4.16)$$

Note that the semi-dual expectation formulation is still valid at the limit when $\varepsilon = 0$, contrarily to the dual expectation formulation (4.2). Indeed, using the c-transform from standard OT, we have that

$$W_c(\alpha, \beta) = \max_{\mathbf{v} \in \mathbb{R}^n} \mathbb{E}_\alpha[g_0^X(\mathbf{v})],$$

where $g_0^x(\mathbf{v}) \stackrel{\text{def.}}{=} \max_{j \in 1 \dots n} \mathbf{v}_j - c(x, y_j) + \sum_{i=1}^n \mathbf{v}_i \boldsymbol{\beta}_i$.

4.3.3 Some Analytic Properties of the Semi-Dual Functional

Since the potential $\mathbf{v}$ is a n-dimensional vector when β is a discrete measure with n diracs, and we can compute the gradient and Hessian of g_ε, deriving some useful properties of the semi-dual function.

Proposition 13. *Consider the semi-dual functional g_ε defined in (4.16). When $\varepsilon > 0$ its gradient is defined by*

$$\nabla_{\mathbf{v}} g_\varepsilon^x(\mathbf{v}) = \boldsymbol{\beta} - \chi_\varepsilon(x),$$

and the hessian is given by

$$\partial^2_{\mathbf{v}} g_\varepsilon^x(v) = \frac{1}{\varepsilon} \left(\chi_\varepsilon(x) \chi_\varepsilon(x)^T - \operatorname{diag}(\chi_\varepsilon(x)) \right),$$

where

$$\chi_\varepsilon(x)_i = \frac{\exp(\frac{\mathbf{v}_i - c(x, y_i)}{\varepsilon})}{\sum_{j=1}^J \exp(\frac{\mathbf{v}_j - c(x, y_j)}{\varepsilon})}.$$

Besides,

$$0 \preceq \partial^2_{\mathbf{v}} g_\varepsilon^x(\mathbf{v}) \preceq \frac{1}{\varepsilon},$$

and thus g_ε^x is a convex function with a Lipschitz gradient.

When $\varepsilon = 0$ (standard OT) g_0 is not smooth and a subgradient is given by

$$\nabla_{\mathbf{v}} g_0(\mathbf{v}, x) = \boldsymbol{\beta} - \chi(x),$$

where

$$\chi(x)_i = \mathbb{1}_{i = j^\star(x)} \quad \text{with } j^\star(x) \in \operatorname{argmin}_{i \in \{1 \dots n\}} c(x, y_i) - \mathbf{v}_i.$$

Note that since the lower bound on the eigenvalues of the Hessian is 0 the semi-dual functional is convex but not strongly convex as strong convexity requires a strictly positive lower-bound on eigenvalues of the Hessian.

Remark 11. (Laguerre Diagrams) Laguerre diagrams are extensions of Voronoi diagrams where each cell j with center y_j has a specific weight $\mathbf{v}_j$. They partition the space $\mathcal{X}$ with n cells $(L_j(\mathbf{v}))_{j=1,\dots,n}$ in the following way:

$$L_j(\mathbf{v}) \stackrel{\text{def.}}{=} \{x \in \mathcal{X} \mid \forall i' \neq i, c(x, y_i) - \mathbf{v}_i \leqslant c(x, y_i') - \mathbf{v}_i'\}.$$

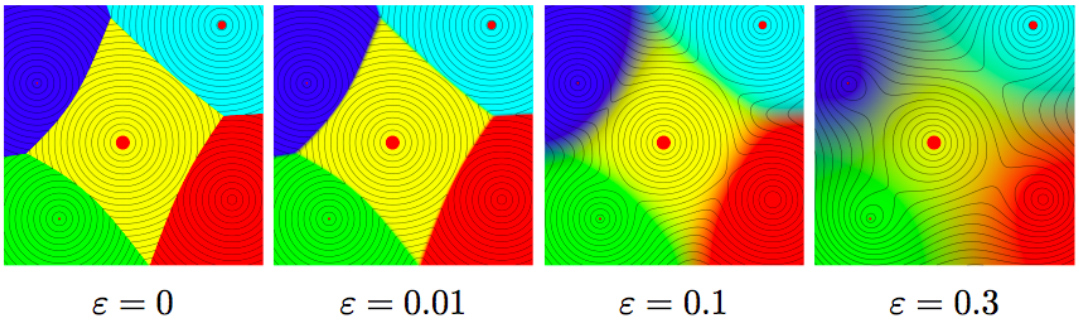


Figure 1.6 – Illustration taken from (Peyré et al., 2017) of the Laguerre cells and their smooth counterpart in 2D, with quadratic cost. The colors indicate the (smoothed) indicator function of the Laguerre cells χ_ε. The red marks are at locations y_i and their size is proportional to v_j. The black lines are the level sets of the $c, \varepsilon-$transform of v.

The function χ appearing in the gradient of the semi-dual is an indicator function corresponding to the Laguerre diagram of the space. More specifically, $\chi(x)_i = 1$ if and only if x belongs to cell i of the Laguerre diagram with distance c and weights v. Its regularized counterpart χ_ε is a smoothed version of the indicator function. Both are represented in Figure 1.6 taken from (Peyré et al., 2017). . The connection between Laguerre cells and semi-discrete OT is presented in (Mérigot, 2011), and we refer to Chapter 4, Sec. 4 for more details.

4.4 Convergence of Entropy-Regularized OT to Standard OT

When using regularized OT as a proxy for OT in various applications, the question of convergence when $\varepsilon \to 0$ naturally arises. For some applications, we are interested in the value of OT, and we thus want to understand how $W_{c,\varepsilon}(\alpha, \beta)$ approximates $W_c(\alpha, \beta)$. For others however, we are interested in the transport plan, and we are thus interested in the convergence of the optimizer π_ε (or the optimizers $(u_\varepsilon, v_\varepsilon)$ for the dual problem, to recover π_ε with the primal-dual relationship.

The convergence of $W_{c,\varepsilon}(\alpha, \beta)$ to $W_c(\alpha, \beta)$ is well known, for instance see Chapter 3, Sec. 3 where we derive convergence rates. However convergence of the optimizers is a more delicate issue. For the primal problem ($\mathcal{P}_\varepsilon$), we have the following theorem from (Carlier et al., 2017) which proves $\Gamma-$convergence of the regularized problem to the unregularized one in the case of a euclidean cost in $\mathbb{R}^d$. $\Gamma-$convergence is a powerful property implying both convergence of the value of the problem and convergence of the minimizers.

Theorem 9. ***(Convergence of Entropy-Regularized OT)*** *(Carlier et al., 2017) Consider the primal problem of entropy-regularized optimal transport ($\mathcal{P}_\varepsilon$) on $\mathbb{R}^d$ with*

cost function $c(x, y) = \|x - y\|^p$*, and denote by* π_ε *its unique minimizer. We have*

$$\lim_{\varepsilon \to 0} W_{c,\varepsilon}(\alpha, \beta) = W_c(\alpha, \beta) \quad \textit{and} \quad \pi_\varepsilon \rightharpoonup \pi,$$

where π *is the minimizer of the unregularized primal* ($\mathcal{P}$) *and* $\rightharpoonup$ *denotes convergence with respect to the weak topology.*

For the dual problem ($\mathcal{D}_\varepsilon$), proof of convergence of the regularized minimizers is given in (Cominetti and Martin, 1994) in the case of discrete measures.

For the semi-dual problem ($\mathcal{S}_\varepsilon$), we proposed in (Genevay et al., 2016) a proof of convergence of the minimizer v_ε in the case where one measure is discrete which is precisely the case where the semi-dual formulation presents an interest.

Proposition 14. ***(Convergence of the semi-dual regularized problem)*** *We assume that* $\forall y \in \mathcal{Y}$, $c(\cdot, y) \in L^1(\alpha)$*, that* $\beta = \sum_{j=1}^m \boldsymbol{\beta}_j \delta_{y_j}$*, and we fix* $x_0 \in \mathcal{X}$*. For all* $\varepsilon > 0$*, let* $\mathbf{v}^\varepsilon$ *be the unique solution of* ($\mathcal{S}_\varepsilon$) *such that* $\mathbf{v}^\varepsilon(x_0) = 0$*. Then* $(\mathbf{v}^\varepsilon)_\varepsilon$ *is bounded and all its converging sub-sequences for* $\varepsilon \to 0$ *are solutions of* ($\mathcal{S}_0$)*.*

We first prove a useful lemma.

Lemma 2. *If* $\forall y$, $x \mapsto c(x, y) \in L^1(\alpha)$ *then* g_ε *converges pointwise to* g_0*.*

Proof. Let $w_j(x) \overset{\text{def.}}{=} \mathbf{v}_j - c(x, y_j)$ and $j^* \overset{\text{def.}}{=} \operatorname{argmax}_j w_j(x)$.
On the one hand, since $\forall j$, $w_j(x) \leqslant w_{j^*}(x)$ we get

$$\varepsilon \log\left(\sum_{j=1}^m e^{\frac{w_j(x)}{\varepsilon}} \beta_j\right) = \varepsilon \log\left(e^{\frac{w_{j^*}(x)}{\varepsilon}} \sum_{j=1}^m e^{\frac{w_j(x) - w_{j^*}(x)}{\varepsilon}} \beta_j\right) \leqslant w_{j^*}(x) + \varepsilon \log\left(\sum_{j=1}^m \beta_j\right) = w_{j^*}(x).$$

On the other hand, since log is increasing and all terms in the sum are non negative we have

$$\varepsilon \log\left(\sum_{j=1}^m e^{\frac{w_j(x)}{\varepsilon}} \beta_j\right) \geqslant \varepsilon \log\left(e^{\frac{w_{j^*}(x)}{\varepsilon}} \beta_{j^*}\right) = w_{j^*}(x) + \varepsilon \log(\beta_{j^*}) \overset{\varepsilon \to 0}{\longrightarrow} w_{j^*}(x).$$

Hence $\varepsilon \log\left(\sum_{j=1}^m e^{\frac{w_j(x)}{\varepsilon}} \beta_j\right) \overset{\varepsilon \to 0}{\longrightarrow} w_{j^*}(x)$ and $\varepsilon \log\left(\sum_{j=1}^m e^{\frac{w_j(x)}{\varepsilon}} \beta_j\right) \leqslant w_{j^*}(x)$.
Since we assumed $x \mapsto c(x, y_j) \in L^1(\alpha)$, then $w_{j^*} \in L^1(\alpha)$ and by dominated convergence we get that $g_\varepsilon(v) \overset{\varepsilon \to 0}{\longrightarrow} g_0(v)$. □

Proof. (Proof of Proposition 14) First, let us prove that $(\mathbf{v}_\varepsilon)_\varepsilon$ has a converging subsequence, where $\mathbf{v}_\varepsilon \overset{\text{def.}}{=} (v_\varepsilon(y_1), \dots, v_\varepsilon(y_n))$. The dual optimal condition gives that $v_\varepsilon(y_i) = -\varepsilon \log\left(\int_{\mathcal{X}} e^{\frac{u_\varepsilon(x) - c(x, y_i)}{\varepsilon}} d\alpha(x)\right)$. We denote by $\tilde{v}_\varepsilon$ the c-transform of u_ε such that $\tilde{v}_\varepsilon(y_i) = \min_{x \in \mathcal{X}} c(x, y_i) - u_\varepsilon(x)$. From standard results on optimal transport (see (Santambrogio, 2015), p.11) we know that $|\tilde{v}_\varepsilon(y_i) - \tilde{v}_\varepsilon(y_j)| \leqslant \omega(\|y_i - y_j\|)$, where

ω is the modulus of continuity of the cost c. Besides, using once again the soft-max argument we can bound $|v_\varepsilon(y) - \tilde{v}_\varepsilon(y)|$ by some constant C. Thus we get that:

$$\begin{aligned}|v_\varepsilon(y_i) - v_\varepsilon(y_j)| &\leqslant |v_\varepsilon(y_i) - \tilde{v}_\varepsilon(y_i)| + |\tilde{v}_\varepsilon(y_i) - \tilde{v}_\varepsilon(y_j)| + |\tilde{v}_\varepsilon(y_j) - v_\varepsilon(y_j)| \\ &\leqslant C + \omega(\|y_i - y_j\|) + C.\end{aligned}$$

Besides, the regularized potentials are unique up to an additive constant. Hence we can set without loss of generality $v_\varepsilon(y_0) = 0$. So from the previous inequality yields:

$$v_\varepsilon(y_i) \leqslant 2C + \omega(\|y_i - y_0\|).$$

So $\mathbf{v}_\varepsilon$ is bounded on $\mathbb{R}^m$ and thus we can extract a subsequence which converges to a certain limit that we denote by $\bar{\mathbf{v}}$.

Let $\mathbf{v}^* \in \operatorname{argmax}_v g_0$. To prove that $\bar{\mathbf{v}}$ is optimal, it suffices to prove that $g_0(\mathbf{v}^*) \leqslant g_0(\bar{\mathbf{v}})$.
By optimality of $\mathbf{v}_\varepsilon$, we have

$$g_\varepsilon(\mathbf{v}^*) \leqslant g_\varepsilon(\mathbf{v}_\varepsilon).$$

The term on the left-hand side of the inequality converges to $g_0(\mathbf{v}^*)$ since g_ε converges pointwise to g_0. We still need to prove that the right-hand term converges to $g_0(\bar{\mathbf{v}})$.

By the Mean Value Theorem, there exists $\tilde{\mathbf{v}}_\varepsilon \overset{\text{def.}}{=} (1 - t^\varepsilon)\mathbf{v}_\varepsilon + t^\varepsilon \bar{\mathbf{v}}$ for some $t^\varepsilon \in [0, 1]$ such that

$$|g_\varepsilon(\mathbf{v}_\varepsilon) - g_\varepsilon(\bar{\mathbf{v}})| \leqslant \|\nabla g_\varepsilon(\tilde{\mathbf{v}_\varepsilon})\| \|\mathbf{v}_\varepsilon - \bar{\mathbf{v}}\|$$

The gradient of g_ε reads

$$\nabla_v g_\varepsilon(\mathbf{v}) = \beta - \pi(\mathbf{v}),$$

where $\pi_i(\mathbf{v}) = \dfrac{\int_{\mathcal{X}} e^{\frac{\mathbf{v}_i - c(x, y_i)}{\varepsilon}} \beta_i d\alpha(x)}{\int_{\mathcal{X}} \sum_{j=1}^m e^{\frac{\mathbf{v}_j - c(x, y_j)}{\varepsilon}} \beta_j d\alpha(x)}$.

It is the difference of two elements in the simplex thus it is bounded by a constant C independently of ε.

Using this bound in (4.4) yields

$$g_\varepsilon(\bar{\mathbf{v}}) - C\|\mathbf{v}_\varepsilon - \bar{\mathbf{v}}\| \leqslant g_\varepsilon(\mathbf{v}_\varepsilon) \leqslant g_\varepsilon(\bar{\mathbf{v}}) + C\|\mathbf{v}_\varepsilon - \bar{\mathbf{v}}\|.$$

By pointwise convergence of g_ε we know that $g_\varepsilon(\bar{\mathbf{v}}) \to g_0(\bar{\mathbf{v}})$, and since $\bar{\mathbf{v}}$ is a limit point of $\mathbf{v}_\varepsilon$ we can conclude that the left and right-hand terms of the inequality converge to $g_0(\bar{\mathbf{v}})$. Thus we get $g_\varepsilon(\mathbf{v}_\varepsilon) \to g_0(\bar{\mathbf{v}})$. □

Chapter 2

Learning with Sinkhorn Divergences

Optimal Transport (OT) metrics and their ability to handle measures with non-overlapping supports have emerged as a promising tool to learn a parametric distribution – for instance a generative model – from a dataset. Yet, training generative models using OT raises formidable computational and statistical challenges, because of (i) *the computational burden of evaluating OT losses,* (ii) *their instability and lack of smoothness,* (iii) *the difficulty to estimate them, as well as their gradients, in high dimension because of the curse of dimensionality from which they suffer.*

In this chapter we introduce Sinkhorn Divergences, based on entropy-regularized OT, which generates a family of losses interpolating between Wasserstein (OT) (when the regularization parameter $\varepsilon = 0$) and Maximum Mean Discrepancy (MMD) losses (when $\varepsilon = \infty$). Aside from the interpolation in terms of cost, which we demonstrate, we also observe empirically that they allow to find a sweet spot leveraging the geometry of OT on the one hand, and the favorable high-dimensional sample complexity of MMD on the other hand (this is formally proved in Chapter 3, Theorem 14).

We use this new discrepancy between measures to train large scale generative models, with an OT-based loss which does not suffer from its usual computational and statistical shortcomings. This is achieved thanks to: (a) *entropic-regularization, which turns the original OT loss into the differentiable and more robust Sinkhorn Divergence, that can be computed efficiently using Sinkhorn fixed point iterations;* (b) *algorithmic (automatic) differentiation, allowing to get stable gradients of these iterations with seamless GPU execution. We further propose an algorithm to learn a cost function on the data space in an adversarial way, similar to what has been done for kernels with MMD.*

This chapter is based on (Genevay et al., 2018), with the addition of some background on the training of generative models, more details on the adversarial learning of the cost functions, and an extensive comparison of various losses on simple models.

1 Introduction

Several important statistical problems boil down to fitting a parametric model to a dataset, i.e. estimating the parameters of a chosen model that fits observed data in some meaningful way. While the standard approach for models with a density is Maximum Likelihood Estimation (MLE), this approach is often flawed in machine learning tasks where the sought after distribution is obtained in a generative fashion. These *generative models* are obtained as the mapping of a low dimensional reference measure through a non-linear function with values in a high dimensional space (e.g. a neural network). These models are easy to sample from, but their density is singular in the sense that it only has positive probability on a low-dimensional "manifold" of the observation space and is zero elsewhere, thus making the usual MLE unusable.

Previous works. For purely generative models, several likelihood-free workarounds exist. Major approaches include variational autoencoders (VAE) (Kingma and Welling, 2013), generative adversarial networks (GAN) (Goodfellow et al., 2014) and numerous variations around these two ideas (Larsen et al., 2016). The adversarial GAN approach computes the best achievable classification accuracy (in which the training and generated datapoints have opposite labels) for a given class of classifiers as a proxy for the distance between two distributions: If accuracy is high distributions are well separated, if accuracy is low they are difficult to tell apart and lie thus at a very close distance. Another approach consists in minimizing a metric between distributions: the maximal mean discrepancy (Gretton et al., 2006), parametrized by a positive-definite kernel function. It was shown in ensuing works that the effectiveness of the MMD to learn generative models (Li et al., 2015; Dziugaite et al., 2015) hinges on the ability to find a relevant kernel, which is a highly nontrivial choice. The Wasserstein or earth mover's distance, which also allows to compare distributions with non-overlapping supports, has recently emerged as a serious contender to train generative models. While it was long disregarded because of its computational burden—in its original form solving OT amounts to solving an expensive network flow problem when comparing discrete measures in metric spaces—recent works have shown that this cost can be largely mitigated by settling for cheaper approximations obtained through strongly convex regularizers, in particular entropy, as detailed in Chapter 1 of this thesis. The benefits of this regularization has opened the path to many applications of the Wasserstein distance in supervised learning problems (Courty et al., 2014; Frogner et al., 2015; Huang et al., 2016; Rolet et al., 2016). Although the use of Wasserstein metrics for inference in generative models was considered over ten years ago in (Bassetti et al., 2006), that development remained exclusively theoretical until a recent wave of papers managed to implement that idea more or less faithfully: using entropic regularization over a discrete space (Montavon et al., 2016), with approximate Bayesian computations (Bernton et al., 2017), and considering

a neural network parameterization of the dual potential in the dual OT problem defining 1-Wasserstein distance (Arjovsky et al., 2017). As opposed to this dual way to compute gradients of the fitting energy, we advocate for the use of a primal formulation, which is numerically stable, because it does not involve differentiating the (dual) solution of an OT sub-problem, as also pointed out in (Bousquet et al., 2017). Additionally, introducing entropic regularization in the formulation of optimal transport allows to interpolate between a pure OT loss and a Maximum Mean Discrepancy loss, thus bridging the gap between these two approaches often presented as opposed points of view. Shortly after the submission of this work, we came across the recent work by (Salimans et al., 2018) which shares several ideas with our method. One distinction lies in the fact that they do not back-propagate errors across the Sinkhorn iterations, but rather use an estimate of the optimal transport matrix to compute an upper-bound on the Sinkhorn divergence, as was done for instance in (Cuturi and Doucet, 2014).

Contributions. The main contributions of this chapter are twofold : (i) a theoretical contribution regarding a new OT-based loss on measures, (ii) a simple numerical scheme to learn generative models under this loss. (i) We introduce the Sinkhorn Divergence, based on regularized optimal transport with an entropy penalty, and we prove that when the smoothing parameter $\varepsilon = 0$ we recover pure OT loss whereas letting $\varepsilon = +\infty$ leads to MMD. The addition of entropy is important to reduce sample complexity and gradient bias, and thus allows us to take advantage of the good geometrical properties of OT without its drawbacks in high-dimensions. (ii) We propose a computationally tractable and stable approach to learn with that Sinkhorn Divergence, which enables inference for any differentiable generative model. It operates by approximating Sinkhorn Divergences with minibatches and L iterations of Sinkhorn's algorithm. As routinely done in standard deep-learning architecture frameworks, the training is then achieved using stochastic gradient descent and automatic differentiation. This provides accurate and stable approximation of the loss and its gradient, at a reasonable extra computational cost, and streams nicely on GPU hardware. When dealing with complex data, we propose to learn the cost function for OT in an adversarial way, similarly to what is done for kernels with MMD in (Li et al., 2017).

Subsequent to this work, Sinkhorn Divergences have successfully been used in a deterministic setting for shape registration (Feydy and Trouvé, 2018), which consists in finding a diffeomorphism matching a deformed image to a target. Sinkhorn Divergences perform better than MMD for this task, as they take the global geometry of the problem into account where MMD is more local and has trouble dealing with parts of the shape that are further away.

The GAN rush. This work was carried out in the early stages of what can be called *the GAN rush*. The interest on GANs has kept growing since the seminal work by

(Goodfellow et al., 2014), and it has clearly exploded in the past couple of years, with a large part of the machine learning community now studying generative models. There are various motivations behind the interest in generative models, but the one that is the most popular is realistic image generation. Thus, although it is a crucial issue, evaluation and comparison of generative models has long been neglected by the community, where most papers simply relied on the quality of generated images to assess their performance with little to no insight on how well they fit the underlying distribution of the data. This resulted in a multitude of papers proposing network architectures, regularization techniques, new losses and other heuristics whose superiority over existing methods is difficult to evaluate. These methods all generate nice images, so their efficiency in terms of computer graphics is clear, but from a statistical point of view, quantifying how well a model fits the unknown distribution of the data remains an open question.The choice of an evaluation metric for GANs is indeed a complex matter, which we quickly discuss in Sec. 4, and a recent survey on GANs (Lucic et al., 2018) suggests a few metrics which should be used for this purpose, with a focus on image-generation tasks. Conducting a large-scale empirical study to compare several state-of-the art GAN models (not including regularized OT- and MMD-based GANs) with their metrics, the authors found out that "most models can reach similar scores with enough hyperparameter optimization and random restarts". Thus in practice, a good network architecture was enough for their assessed models to perform well in terms of image generation, and the loss function itself did not have much effect on the performance. In the GAN rush, we adopted a different position on the subject, distancing ourselves from image generation and rather trying to answer the following question : *what is a robust loss to learn a (possibly high-dimensional) singular distribution from samples?* It is this question that should be kept in mind through this chapter and the following – which gives statistical properties of Sinkhorn Divergences, reinforcing the interpolation theorem presented here. Thus, we consider images generation merely as an application of our density fitting scheme, not a the main goal. For a better overview of the performance of Sinkhorn Divergences in image generation, the reader should refer to (Salimans et al., 2018).

2 Density Fitting

We consider a data set of n (usually very large) observations $(y_1, \dots, y_n) \in \mathcal{X}^n$ generated from an unknown distribution β and we want to learn a generative model that produces samples that are similar to that dataset. Samples $x = g_\theta(z)$ from the generative model are defined by taking as input a sample $z \in \mathcal{Z}$ from some reference measure ζ (typically a uniform or a Gaussian measure in a low-dimensional space $\mathcal{Z}$) and mapping it through a differentiable function $g_\theta : \mathcal{Z} \to \mathcal{X}$. Formally, this corresponds to defining the generative model measure α_θ from which x is drawn as $\alpha_\theta = g_{\theta\#}\zeta$. The goal is to find θ which minimizes a certain loss $\mathcal{L}$ between the model measure α_θ and

the unknown measure of the data β:

$$\theta \in \operatorname*{argmin}_{\theta} \mathcal{L}(\alpha_\theta, \beta). \tag{2.1}$$

Maximum likelihood estimation (MLE) is obtained with $\mathcal{L}(\alpha_\theta, \beta) = -\sum_j \log \frac{\mathrm{d}\alpha_\theta}{\mathrm{d}x}(y_j)$, where $\frac{\mathrm{d}\alpha}{\mathrm{d}x}$ is the density of α_θ with respect to a fixed reference measure (a typical choice is $\mathrm{d}x$ being the Lebesgue measure in $\mathcal{X} = \mathbb{R}^d$). This MLE loss can be seen as a discretized version of the relative entropy (a.k.a. the Kullback-Leibler divergence) as it converges to $D_{KL}(\alpha|\beta)$ when $N \to \infty$. A major issue with this approach is that in general generative models defined this way (when $\mathcal{Z}$ has a much smaller dimensionality than $\mathcal{X}$) have singular distributions (*i.e.* supported on a low-dimensional manifold), without density with respect to a fixed measure, and therefore MLE cannot be considered.

2.1 Learning with φ-divergences

The first idea that emerged in the literature of Generative Adversarial Networks (Goodfellow et al., 2014) was to use the Jensen-Shannon divergence, a special instance of the class of $\varphi-$divergences (see Chapter 1, sec. 2.1 for a thorough introduction), to solve the density fitting problem. Subsequent work by (Nowozin et al., 2016) considers a more general framework to learn a generative model with $\varphi-$divergences, which we describe in this section. The density fitting problem they consider is

$$\min_{\theta} D_\varphi(\alpha_\theta \mid \beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha_\theta(x)}{\mathrm{d}\beta(x)}\right) \mathrm{d}\beta(x), \tag{2.2}$$

where α_θ and β are absolutely continuous with respect to a reference measure $\mathrm{d}x$ and $\mathrm{d}\alpha_\theta$, $\mathrm{d}\beta$ are their respective densities (see Chapter 1, sec. 2.1 for a more general definition in the case of measure that are not absolutely continuous); and is φ a convex, lower-semicontinuous function on $\mathbb{R}^+$ satisfying $\varphi(1) = 0$. However, φ-divergences are hard to estimate through samples, in particular because of the fact that they do not metrize weak convergence (again, see Chapter 1, sec. 2.1 for more details).

To alleviate this shortcoming, (Nguyen et al., 2010) suggests using the following lower bound leveraging the dual definition of φ-divergences:

Proposition 15. *(Lower bound on φ-divergences)*

$$D_\varphi(\alpha \mid \beta) = \sup_{T:\mathcal{X}\to\mathbb{R}} \mathbb{E}_\alpha(g(X)) - \mathbb{E}_\beta(\varphi^* g(Y)) \geqslant \sup_{T\in\mathcal{T}} \mathbb{E}_\alpha(T(X)) - \mathbb{E}_\beta(\varphi^* T(Y)),$$

where $\varphi^*(t) = \sup_u tu - \varphi(u)$ is the Legendre transform of φ, the set $\{T : \mathcal{X} \to \mathbb{R}\}$ is the set of measurable functions from $\mathcal{X}$ to $\mathbb{R}$ and $\mathcal{T}$ is an arbitrary class of measurable functions.

Parametrizing T by a variable w, i.e. setting $\mathcal{T} \overset{\text{def.}}{=} \{T_w \mid w \in \mathcal{W}\}$ and using the

previous lower bound (2.2) yields the following saddle point problem:

$$\min_{\theta} \max_{w} \mathbb{E}_{\beta}(T_w(X)) - \mathbb{E}_{\alpha_\theta}(\varphi^*(T_w(X)). \tag{2.3}$$

The formulation of the original GANs is derived from the above, using $f(u) = u \log u + (u+1)\log(u+1)$ (a slight modification of the Jensen-Shannon divergence given in Table 1.1, Chapter 1) and $T_w = \log D_w$, where D_w is a neural network called the *discriminator*. Then (2.3) becomes

$$\begin{aligned} &\min_{\theta} \max_{w} \mathbb{E}_{\beta}[\log D_w(X)] - \mathbb{E}_{\alpha_\theta}[\log(1 - D_w(X))] \\ \Leftrightarrow &\min_{\theta} \max_{w} \mathbb{E}_{\beta}[\log D_w(X)] - \mathbb{E}_{\zeta}[\log(1 - D_w(g_\theta(Z)))], \end{aligned}$$

using the definition of α_θ as the pushforward of ζ through g_θ. From a game-theory point of view, the general GAN formulation can be seen as finding the equilibrium in a two-player game where player one optimizes its parameter θ to fool the discriminator D_w, whose parameter w is optimized by the player two whose goal is to distinguish between samples from the model measure α_θ and samples from the true measure β.

2.2 Maximum Mean Discrepancy and Optimal Transport

A more robust way to compare measures with disjoint support, is to consider losses which *metrize the weak convergence of measures* (see Chapter 1, Sec. 2.1 for a precise definition). Intuitively, these losses enables the comparison of singular measures by taking into account spatial displacement of the measures. For instance, they avoid the typical failure case of φ-divergences by satisfying $\mathcal{L}(\delta_x, \delta_{x'}) \to 0$ as $x \to x'$ where φ-divergences would be equal to a constant. A classical framework for such a loss function $\mathcal{L}$ are *Integral Probability Metrics* (IPMs) which are thoroughly described in Chapter 1, Sec. 2.2. Given a set of measurable functions $\mathcal{F}$, the IPM $d_{\mathcal{F}}$ is defined as

$$d_{\mathcal{F}}(\alpha, \beta) \overset{\text{def.}}{=} \sup_{f \in \mathcal{F}} |\mathbb{E}_{\alpha}(f(X)) - \mathbb{E}_{\beta}(f(Y))|.$$

Popular IPMs include the 1-Wasserstein distance (with $\mathcal{F} = \{f \,;\, \|\nabla f\|_\infty \leqslant 1\}$ the set of 1-Lipschitz functions) and Maximum Mean Discrepancies (with $\mathcal{F} = \{f \,;\, \|f\|_{\mathcal{H}} \leqslant 1\}$ where $\mathcal{H}$ is a Reproducing Kernel Hilbert Space). Recall from Chapter 1, Proposition 3 that on a RKHS with kernel k, MMD can be rewritten as follows (Gretton et al., 2006):

$$MMD_k^2(\alpha, \beta) = \mathbb{E}_{\alpha\otimes\alpha}[k(X, X')] + \mathbb{E}_{\beta\otimes\beta}[k(Y, Y')] - 2\mathbb{E}_{\alpha\otimes\beta}[k(X, Y)]. \tag{2.4}$$

A different approach, for which we advocate, is to consider Optimal Transport (OT) metrics. The OT metric between two probability distributions $(\alpha, \beta) \in \mathcal{M}_+^1(\mathcal{X}) \times$

$\mathcal{M}_+^1(\mathcal{X})$ is defined as the solution of the (possibly infinite dimensional) linear program:

$$W_c(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)\mathrm{d}\pi(x,y), \tag{2.5}$$

where the set of admissible couplings $\Pi(\alpha,\beta)$ is composed of joint probability distributions over the product space $\mathcal{X}\times\mathcal{X}$ with imposed marginals (α,β). Formula (2.5) corresponds to the celebrated Kantorovitch formulation (Kantorovich, 1942) of OT (see Chapter 1, sec.2.3 for more details). Here $c(x,y)$ is the "ground cost" to move a unit of mass from x to y, and we shall make no assumptions (except for regularity) on its form. When $\mathcal{X}$ is equipped with a distance $d_{\mathcal{X}}$, a typical choice is to set $c(x,y) = d_{\mathcal{X}}(x,y)^p$ where $p > 0$ is some exponent, in which case for $p \geqslant 1$ $W_c^{1/p}$ is the so-called p-Wasserstein distance between probability measures. Two majors obstacles to the use of the Wasserstein-distance for inference, as for many machine learning applications, are its high computational complexity and the curse of dimensionality from which it suffers.

2.3 Regularized OT and Variants of the Regularized OT Loss

As detailed in Chapter 1, Sec. 3 one can resort to regularized optimal transport to alleviate the computational burden of OT. Its primal formulation is given by:

$$\min_{\pi\in\Pi(\alpha,\beta)} \int c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \int \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\pi(x,y). \tag{$\mathcal{P}_\varepsilon$}$$

The primal problem ($\mathcal{P}_\varepsilon$) has a equivalent dual formulation (see Chapter 1, Proposition 7) which consists in solving

$$\max_{(u,v)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon. \tag{$\mathcal{D}_\varepsilon$}$$

The optimizer π_ε of the primal formulation can be recovered from optimizers of the dual problem $(u_\varepsilon, v_\varepsilon)$ via the following formula: $\mathrm{d}\pi_\varepsilon(x,y) = e^{\frac{u_\varepsilon(x)+v_\varepsilon(y)-c(x,y)}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$. In practice, state-of-the-art algorithms (including Sinkhorn, the one we use here, already detailed in Chapter 1, Sec. 4.2) solve the dual problem and use the primal-dual relationship to recover the solution of the primal.

Note that introducing this regularization also breaks the curse of dimensionality for ε large enough, making the estimation of regularized OT more robust to sampling noise, and this is the main matter of Chapter 3.

These problems yield four different costs, which all converge to the value of unregularized OT when $\varepsilon \to 0$:

- *Primal cost with entropy*:

$$\mathcal{L}(\alpha,\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi_\varepsilon(x,y) + \varepsilon \int \log\left(\frac{\mathrm{d}\pi_\varepsilon(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\pi_\varepsilon(x,y).$$

This loss is the value of the primal of the regularized problem. Theoretical results are known for this cost, which is extensively studied in Chapter 3. In particular, we have a convergence rate for its approximation from samples, based on the regularization parameter ε.

- *Dual cost with entropy*:

$$\mathcal{L}(\alpha,\beta) = \int_{\mathcal{X}} u_\varepsilon(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v_\varepsilon(y)\mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u_\varepsilon(x)+v_\varepsilon(y)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon.$$

Since strong duality holds for regularized OT, this loss is equal to the primal with entropy. This is actually this formulation that is used to prove the sample-complexity results from Chapter 3. However in practice, when one uses an algorithm to approximate $(u_\varepsilon, v_\varepsilon)$, the primal and dual problems with entropy yield different values.

- *Primal cost without entropy*:

$$\mathcal{L}(\alpha,\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi_\varepsilon(x,y).$$

This loss is arguably the most widely used in practice. It is this version that is used in (Cuturi, 2013), which first showed the benefits of regularized OT for machine learning. It has recently been studied in (Luise et al., 2018) under the name *Sharp Sinkhorn* – to avoid confusion with the primal loss with entropy. This paper gives an algorithm to compute the gradient of this loss, which can further be used in supervised learning problems.

- *Dual cost without entropy*:

$$\mathcal{L}(\alpha,\beta) = \int_{\mathcal{X}} u_\varepsilon(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v_\varepsilon(y)\mathrm{d}\beta(y).$$

The primal-dual relationship is $\mathrm{d}\pi_\varepsilon = e^{\frac{u_\varepsilon(x)+v_\varepsilon(y)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y)$. The probability constraint on π_ε thus implies $\int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u_\varepsilon(x)+v_\varepsilon(y)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y) = 1$. So the dual cost without entropy is equal to the dual cost with entropy, which once again might not be the case when considering approximations of $(u_\varepsilon, v_\varepsilon)$.

2.4 Sinkhorn Divergences : an Interpolation Between OT and MMD

We denote by $W_{c,\varepsilon}$ the *primal without entropy* variant of the regularized OT loss:

$$W_{c,\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} \int c(x,y)\mathrm{d}\pi_\varepsilon(x,y), \tag{2.6}$$

where π_ε is the optimal coupling for the regularized OT problem ($\mathcal{P}_\varepsilon$), and by H_ε the additional term that comes from the entropic regularization:

$$H_\varepsilon(\alpha,\beta) \overset{\text{def.}}{=} \varepsilon \int \log(\frac{\mathrm{d}\pi_\varepsilon(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)})\mathrm{d}\pi_\varepsilon(x,y), \tag{2.7}$$

so that the *primal loss with entropy* is

$$W^H_{c,\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} W_{c,\varepsilon}(\alpha,\beta) + H_\varepsilon(\alpha,\beta). \tag{2.8}$$

To correct for the fact that $W_{c,\varepsilon}(\alpha,\alpha) \neq 0$ and $W^H_{c,\varepsilon}(\alpha,\alpha) \neq 0$, we introduce the following normalization, which we call Sinkhorn Divergence:

Definition 9. ***(Sinkhorn Divergence)*** *The Sinkhorn Divergence between two probability measures α, β is defined as:*

$$SD_{c,\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} W_{c,\varepsilon}(\alpha,\beta) - \frac{1}{2}W_{c,\varepsilon}(\alpha,\alpha) - \frac{1}{2}W_{c,\varepsilon}(\beta,\beta), \tag{2.9}$$

where $W_{c,\varepsilon}$ is the primal cost without entropy defined in (2.6). Alternatively, we define:

$$SD^H_{c,\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} W^H_{c,\varepsilon}(\alpha,\beta) - \frac{1}{2}W^H_{c,\varepsilon}(\alpha,\alpha) - \frac{1}{2}W^H_{c,\varepsilon}(\beta,\beta), \tag{2.10}$$

where $W^H_{c,\varepsilon}$ is the primal cost with *entropy defined in (2.8)*

Far from simply correcting the bias of $W_{c,\varepsilon}(\alpha,\alpha)$, the Sinkhorn Divergence also appears as an interpolating discrepancy between OT and MMD.

Theorem 10. ***(Asymptotics of Sinkhorn Divergence with Respect to ε)*** *The Sinkhorn Divergence has the following asymptotic behavior in ε:*

(i) as $\varepsilon \to 0$, $\quad SD_{c,\varepsilon}(\alpha,\beta) \to W_c(\alpha,\beta)$,

(ii) as $\varepsilon \to +\infty$, $\quad SD_{c,\varepsilon}(\alpha,\beta) \to \frac{1}{2}MMD^2_{-c}(\alpha,\beta)$.

When $-c$ is a positive definite kernel, MMD_{-c} is the MMD with the kernel that is minus the cost used in the optimal transport problem.

Besides, if $c \in \mathcal{C}^1(\mathcal{X}\times\mathcal{Y})$ and $\mathcal{X}$ and $\mathcal{Y}$ are bounded domains of $\mathbb{R}^d$, the asymptotics also hold for $SD^H_{c,\varepsilon}$.

Remark 12. This theorem is a generalization of (Ramdas et al., 2017, §3.3) for continuous measures, and to the cost with entropy.

Proof. Let us start by proving the property for $SD_{c,\varepsilon}$.

(i) The first part of the assumption comes from the fact that $\pi_\varepsilon \rightharpoonup \pi$ (see Chapter 1, Sec. 4.4 or (Carlier et al., 2017)).

(ii) Letting ε go to infinity in the regularized OT problem amounts to finding the coupling with minimum entropy in the constraint set. The problem becomes

$$\min_{\pi \in \Pi(\alpha,\beta)} \int \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right) \mathrm{d}\pi(x,y),$$

where $\Pi(\alpha,\beta)$ is the set of couplings with marginals α and β. Introducing Lagrange multipliers u and v for these constraints, the dual problem becomes $\max_{u,v} \int u(x)\mathrm{d}\alpha(x) + \int v(y)\mathrm{d}\beta(y) - \int \exp(u(x)+v(y))\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$ and the primal-dual relation is given by $\mathrm{d}\pi(x,y) = \exp(u(x)+v(y))\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$. Solving the dual gives $u = v = 0$ and thus the optimal coupling is simply the product of the marginals i.e. $\pi = \alpha \otimes \beta$. This gives

$$W_{c,+\infty}(\alpha,\beta) = \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y).$$

The proof of this assumption for $SD^H_{c,\varepsilon}$ requires some more assumptions to control the entropy, and is based on results from Chapter 3.

(i) The asympotics for $\varepsilon \to 0$ are a direct consequence of Theorem 12 which proves that

$$W^H_{c,\varepsilon}(\alpha,\beta) - W_c(\alpha,\beta) \underset{\varepsilon\to 0}{\sim} 2\varepsilon d \log(1/\varepsilon).$$

(ii) We want to prove that $W^H_{c,\varepsilon}(\alpha,\beta) \underset{\varepsilon\to+\infty}{\to} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$. Since strong duality holds for regularized OT, we have

$$W^H_{c,\varepsilon}(\alpha,\beta) = \int_{\mathcal{X}} u_\varepsilon(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v_\varepsilon(y)\mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u_\varepsilon(x)+v_\varepsilon(y)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon,$$

where u_ε and v_ε are the dual potentials solving ($\mathcal{D}_\varepsilon$). We know from Chapter 3, Proposition 17 that if the cost function c is $\mathcal{C}^1$ and $\mathcal{X}$ and $\mathcal{Y}$ are bounded, then u_ε and v_ε are Lipschitz with the same constant as c, and thus they are bounded in L^∞ norm on $\mathcal{X}$ and $\mathcal{Y}$ independently of ε. This implies that $u_\varepsilon(x) + v_\varepsilon(y) - c(x,y) = O(1)$ when $\varepsilon \to +\infty$. Using the Taylor expansion of the exponential when $\varepsilon \to +\infty$ we get:

$$\begin{aligned} W^H_{c,\varepsilon}(\alpha,\beta) = &\int_{\mathcal{X}} u_\varepsilon(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v_\varepsilon(y)\mathrm{d}\beta(y) \\ &- \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \left(1 + \frac{u_\varepsilon(x)+v_\varepsilon(y)-c(x,y)}{\varepsilon} + O\left(\frac{1}{\varepsilon^2}\right)\right) \mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon, \end{aligned}$$

which simplifies to

$$W^H_{c,\varepsilon}(\alpha,\beta) = \int_{\mathcal{X}\times\mathcal{Y}} \left(c(x,y) + O\left(\frac{1}{\varepsilon}\right)\right) \mathrm{d}\alpha(x)\mathrm{d}\beta(y).$$

Since we have $\int_{\mathcal{X}\times\mathcal{Y}} O\left(\frac{1}{\varepsilon}\right) \mathrm{d}\alpha(x)\mathrm{d}\beta(y) \to 0$ when $\varepsilon \to +\infty$, we get the desired

conclusion.

□

We proved that $SD_{c,\varepsilon} \to MMD_{-c}$ when $\varepsilon \to +\infty$. However, by definition of MMD, $-c$ has to be a positive definite kernel for MMD_{-c} to be well defined. The following proposition proves that some powers of the Euclidean distance yield a valid cost, and define as special instance of MMD also known as *Energy Distance*.

Proposition 16. ***(Energy Distance)*** *(Sejdinovic et al., 2013) Consider the Euclidean distance $\|.\|_2$ on $\mathbb{R}^d$. Then, for $x_0 \in \mathbb{R}^d$,*

$$k_p(x,y) \overset{\text{def.}}{=} \|x\|_2^p + \|y\|_2^p - \|x-y\|_2^p$$

is a positive definite kernel for $0 < p < 1$. And it induces the following MMD, called Energy Distance :

$$ED_p(\alpha,\beta) \overset{\text{def.}}{=} MMD_{k_p}^2(\alpha,\beta) = 2\mathbb{E}_{\alpha\otimes\beta}[\|X-Y\|_2^p] - \mathbb{E}_{\alpha\otimes\alpha}[\|X-X'\|_2^p] - \mathbb{E}_{\beta\otimes\beta}[\|Y-Y'\|_2^p].$$

Recent work by (Feydy et al., 2019 (to appear) also proves that this normalization of regularized OT enforces positive-definiteness for $SD_{c,\varepsilon}^H$, which we conjectured in the early stages of our work on Sinkhorn Divergence, based on empirical evidence, and that Sinkhorn divergences metrize the weak-convergence of measures.

Theorem 11. ***(Positivity of Sinkhorn Divergence)*** *(Feydy et al., 2019 (to appear) Let $\mathcal{X}$ be a compact metric space with a Lipschitz cost function c, that induces, for $\varepsilon > 0$, a positive universal kernel $k_\varepsilon(x,y) \overset{\text{def.}}{=} \exp(-c(x,y)/\varepsilon)$. Then, $SD_{c,\varepsilon}^H$ defines a symmetric positive definite, smooth loss function that is convex in each of its input variables. It also metrizes the convergence in law (or weak-convergence of measures): for all probability Radon measures α and $\beta \in \mathcal{M}_1^+(\mathcal{X})$,*

$$0 = SD_{c,\varepsilon}^H(\beta,\beta) \leqslant SD_{c,\varepsilon}(\alpha,\beta),$$

$$\alpha = \beta \Leftrightarrow SD_{c,\varepsilon}^H(\alpha,\beta) = 0,$$

$$\alpha_n \rightharpoonup \alpha \Leftrightarrow SD_{c,\varepsilon}^H(\alpha_n,\alpha) \to 0.$$

Remark 13. In particular, these results hold for measures with bounded support on a Euclidean space $\mathcal{X} = \mathbb{R}^d$ endowed with ground cost functions $c(x,y) = \|x-y\|_2$ or $c(x,y) = \|x-y\|_2^2$ which induce Laplacian and Gaussian kernels respectively. Note that their proof only holds for Sinkhorn Divergence based on the primal cost *with* entropy SD_ε^H, positive-definiteness of SD_ε remains an open problem.

Discussion on OT *vs.* MMD. As proved in Theorem 1, the Sinkhorn Divergence interpolates between a pure OT loss for $\varepsilon = 0$ and MMD losses for $\varepsilon = +\infty$. As such, when $\varepsilon \to +\infty$, our loss takes advantage of the good properties of MMD losses, and in particular a favorable sample complexity of $O(1/\sqrt{n})$ (decay rate of the approximation of the true loss with a mini-batch of size n) which is the object of Chapter 3 of this thesis. In contrast, the unregularized OT loss suffers from a sample complexity of $O(1/n^{1/d})$, see (Weed and Bach, 2017) for a recent account on this point. Using MMD to train generative models has been shown to be successful in (Dziugaite et al., 2015; Li et al., 2015). The improved Wasserstein GAN approach (Gulrajani et al., 2017) (which penalizes the squared norm of the gradient of the dual potential) is similar to an MMD in the sense that both are IPMs. By tuning the ε parameter, our method is able to take the best of both worlds, to blend the non-flat geometry of OT with the high-dimensional rigidity of MMD losses. Additionally, the Sinkhorn Divergence, as is the case for the original OT problem, can be defined with any cost c, whereas MMD losses are only meaningful when used with positive definite kernels k. We discuss the geometric properties of these losses further in Section 4, where we compare them on various fitting tasks.

3 Sinkhorn AutoDiff Algorithm

We now consider density fitting with Sinkhorn Divergence as a loss:

$$\min_{\theta} E_\varepsilon(\theta) \quad \text{where} \quad E_\varepsilon(\theta) \stackrel{\text{def.}}{=} SD_{c,\varepsilon}(\alpha_\theta, \beta).$$

Computing an approximation of $\nabla_\theta SD_{c,\varepsilon}(\alpha_\theta, \beta)$ is itself a difficult problem. When $\varepsilon = 0$, and when $c = \|x - y\|$ (the case of the 1-Wasserstein distance) a workaround is to use, instead of differentiating the "primal" formula (2.5), the optimum of the "dual" formula, resulting in $\nabla SD_0(\alpha_\theta, \beta) = \int_{\mathcal{Z}} \nabla[h \circ g_\theta](z) \mathrm{d}\zeta(z)$, where h is an optimal dual continuous potential for $\alpha = \alpha_\theta$. This is the problem tackled in (Arjovsky et al., 2017) which uses a deep-network expansion to approximate the continuous dual potential h. While the dual formalism is appealing (in particular because it involves only integration over $\mathcal{Z}$ and not the product space $\mathcal{Z} \times \mathcal{X}$), the resulting gradient formula requires differentiating the dual potential, which tends to be difficult to compute and unstable. A very similar conclusion is reached by (Bousquet et al., 2017) (see in particular their Proposition 3).

We propose a different route, by making two key simplifications: (i) approximate $SD_{c,\varepsilon}(\alpha_\theta, \beta)$ by a size-m mini-batch sampling $SD_{c,\varepsilon}(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$ to make it amenable to stochastic gradient descent ; (ii) approximate $SD_{c,\varepsilon}(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$ by L-steps of the Sinkhorn algorithm (Cuturi, 2013) to obtain an algorithmic loss $SD^{(L)}_{c,\varepsilon}(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$ which is amenable to automatic differentiation.

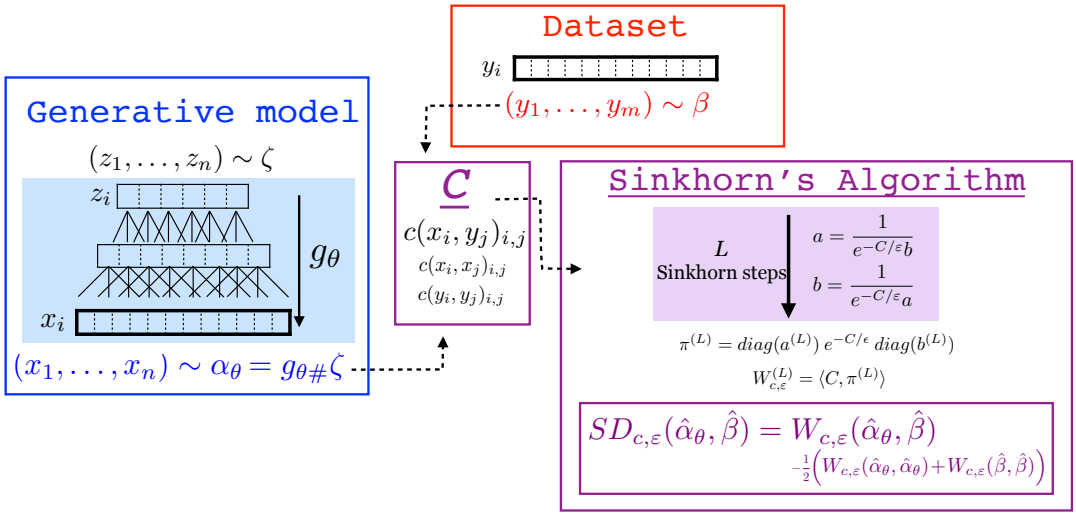


Figure 2.1 – Flow diagram for the computation of the proxy of the Sinkhorn Divergence estimated from samples $SD_{c,\varepsilon}^{(L)}(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$. Samples from the generative model α_θ are obtained by applying the push-forward function g_θ to samples of the initial low-dimensional measure ζ (blue block). These samples are combined with real data (red block) to compute a pairwise distance matrix C, which is in turn used in the Sinkhorn iterations. The resulting loss is the one on which automatic differentiation is applied to perform parameter learning. The display shows a simple 2-layer neural network $g_\theta : z \mapsto x$, but this applies to any generative model.

3.1 Mini-batch Sampling Loss

We approximate $SD_{c,\varepsilon}(\alpha_\theta, \beta)$ by an estimation with empirical measures $SD_{c,\varepsilon}(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$ which leads to consider:

$$\min_\theta SD_{c,\varepsilon}(\hat{\alpha}_{\theta m}, \hat{\beta}_m) \tag{3.1}$$

$$\text{and} \quad \begin{cases} \hat{\alpha}_{\theta m} \overset{\text{def.}}{=} \frac{1}{m}\sum_{i=1}^m \delta_{x_i}, \\ \hat{\beta}_m \overset{\text{def.}}{=} \frac{1}{m}\sum_{i=1}^m \delta_{y_j}, \end{cases} \quad \begin{cases} (z_i)_{i=1}^m \overset{\text{i.i.d}}{\sim} \zeta, \\ \forall\, i,\ x_i \overset{\text{def.}}{=} g_\theta(z_i). \end{cases}$$

As m increases, $\mathbb{E}(SD_{c,\varepsilon}(\hat{\alpha}_{\theta m}, \hat{\beta}_m))$ approaches $SD_{c,\varepsilon}(\alpha_\theta, \beta)$, and convergence of minimizers is studied in (Bernton et al., 2017).

At a given iterate of this stochastic gradient descent scheme (see pseudo-code 2), one draws a mini-batch $(z_i)_{i=1}^m \overset{\text{i.i.d}}{\sim} \zeta$ and a subset of m observations from the dataset, and aims at computing the gradient of $SD_{c,\varepsilon}(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$. In the case where both input measures are discrete (sums of Dirac masses), couplings π can be treated as matrices $\boldsymbol{\pi} \in \mathbb{R}^{m \times n}$, namely $\pi = \sum_{i,j} \boldsymbol{\pi}_{i,j} \delta_{(x_i, y_j)} \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{X})$.

3.2 Sinkhorn Iterates

One major advantage of regularizing the optimal transport problem is that it becomes solvable efficiently using Sinkhorn's algorithm (Sinkhorn, 1964) (when dealing with discrete measures) and leads to a differentiable loss function (as first noticed in (Cuturi, 2013; Cuturi and Doucet, 2014)). Sinkhorn's algorithm is presented in details in Chapter 1, Sec. 4.2 but we give a quick reminder here in the specific case of empirical measures. Recall that the entropic regularization is equivalent to restricting the search space in ($\mathcal{P}_\varepsilon$) to couplings having the so-called scaling form

$$\boldsymbol{\pi}_{i,j} = \mathbf{a}_i \mathbf{K}_{i,j} \mathbf{b}_j \quad \text{where} \quad \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\mathbf{C}_{i,j}/\varepsilon} \quad \text{where} \quad \mathbf{C} \stackrel{\text{def.}}{=} \Big(c(g_\theta(z_i), y_j) \Big)_{ij}.$$

Note that $\mathbf{K}$ depends implicitly on θ (because matrix $\mathbf{C}$ does), and contains therefore all of the geometric information related to the ability of θ to sample points near the dataset. The main computational burden of the procedure, detailed in Algorithm 1 are the matrix-vector multiplication, which stream extremely well on GPU architectures, and therefore nicely add to a typical deep network architecture with L additional layer of linear operations (K can be interpreted as a localized linear filtering) and entry-wise non-linear operations (here divisions).

For a given budget L of iterations, the primal cost is then obtained by using $\boldsymbol{\pi}^{(L)} \stackrel{\text{def.}}{=} \text{diag}(\mathbf{a}^{(L)}) \mathbf{K} \, \text{diag}(\mathbf{b}^{(L)})$ as a proxy for the optimal transport coupling, and thus

$$W_{c,\varepsilon}^{(L)}(\hat{\alpha}_{\theta m}, \hat{\beta}_m) \stackrel{\text{def.}}{=} \langle \mathbf{C}, \boldsymbol{\pi}^{(L)} \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{C}_{i,j} \mathbf{a}_i^{(L)} \mathbf{b}_j^{(L)} \mathbf{K}_{i,j} \tag{3.2}$$

where it is once again important to remind that $\mathbf{K}, \mathbf{C}, \mathbf{a}^{(L)}, \mathbf{b}^{(L)}$ depend on θ. As $L \to +\infty$, one can show that the $\boldsymbol{\pi}^{(L)}$ computed by Sinkhorn's iterates approaches a solution to ($\mathcal{P}_\varepsilon$), with linear convergence rate (deteriorating as $\varepsilon \to 0$), so that $W_{c,\varepsilon}^{(L)}(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$ is a smooth proxy for $W_{c,\varepsilon}(\alpha_\theta, \beta)$ which can be differentiated in a fast and stable way, while converging to $W_{c,\varepsilon}(\alpha_\theta, \beta)$ when $(m, L) \to +\infty$. It is important to realize that for large scale and high dimensional learning applications, empirical considerations (Cuturi, 2013; Kusner et al., 2015; Frogner et al., 2015) suggest that, unlike relevant applications of the same scheme in graphics (Solomon et al., 2015), a relatively strong regularization – a large ε – leads to faster convergence, but also better generalization so that the value for L can be set quite low. This is further backed-up by recent theoretical results – detailed in Chapter 3 – showing that the curse of dimensionality of OT is broken by using regularized OT with a large enough ε.

Algorithm 1 Regularized Primal Loss without Entropy $W^{(L)}_{c_\varphi,\varepsilon}(x_1^m, y_1^m)$

Input: $\varphi, (x_i)_{i=1}^m, (y_j)_{j=1}^m, \varepsilon$

Output: W

$\mathbf{C}_{i,j} \overset{\text{def.}}{=} \|f_\varphi(x_i) - f_\varphi(y_j)\|^p \; \forall\,(i,j)$ (compute the cost matrix $\mathbf{C}$)

$\mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}}$

$\mathbf{b} \leftarrow \mathbb{1}_m,$

for $\ell = 1, 2, \ldots, L$ **do** (L steps of Sinkhorn's algorithm)

$\mathbf{a} \leftarrow \frac{\mathbb{1}_m}{\mathbf{K}\mathbf{b}}$; $\mathbf{b} \leftarrow \frac{\mathbb{1}_m}{\mathbf{K}^\top \mathbf{a}}$

end for

$\boldsymbol{\pi} \leftarrow \mathrm{diag}(\mathbf{a})\mathbf{K}\,\mathrm{diag}(\mathbf{b})$

return $W_{c_\varphi,\varepsilon} = \langle \boldsymbol{\pi}, \mathbf{C} \rangle$ (see (3.2))

3.3 Learning the Cost Function Adversarially

Aside from the regularization parameter, a key element of the Sinkhorn Divergence is the choice of the ground cost c on the data space. In some cases, using a simple metric such as the ℓ^2 norm is sufficient to compare two data points, but when dealing with high-dimensional objects, choosing c is more critical. In such cases, we propose to learn the cost c with the following parametrization

$$c_\varphi(x, y) \overset{\text{def.}}{=} \|f_\varphi(x) - f_\varphi(y)\|^p \quad \text{where} \quad f_\varphi : \mathcal{X} \to \mathbb{R}^{d'},$$

where f_φ can for instance be modeled by a neural network – see numerical experiments below, and can be seen as a feature extractor that reduces the dimensionality of $\mathcal{X}$ through a mapping onto $\mathbb{R}^{d'}$.

The procedure to learn the cost function here is the same as learning a parametric kernel in an MMD model, as done in (Li et al., 2017). The idea, as suggested in (Fukumizu et al., 2009) for MMD, is to learn a cost function (or kernel in their case) that will allow the Sinkhorn Divergence (or MMD in their case) to discriminate well between samples generated by the model distribution α_θ and samples from the data set. In their setting, which is two-sample test, they want to set a threshold τ such that if the MMD evaluated on samples verifies $MMD_k(\hat{\alpha}_\theta, \hat{\beta}) < \tau$ they accept the hypothesis that $\alpha_\theta = \beta$. Thus their kernel function should maximize the value of the discrepancy, so that the equality hypothesis is not wrongfully accepted. Similarly in our case, we want the parameter of the cost function φ to maximize the Sinkhorn Divergence in order to get a strong signal when $\alpha_\theta \neq \beta$. The optimization problem becomes a min-max problem over (θ, φ) instead of a simple minimization problem over θ

$$\min_\theta \max_\varphi SD_{c_\varphi,\varepsilon}(\alpha_\theta, \beta),$$

where in practice $SD_{c_\varphi,\varepsilon}$ is approximated by minibatches and Sinkhorn, as mentioned above. We will give details on the optimization algorithm used for this min-max problem below.

3.4 The Optimization Procedure in Practice

Let us first describe the optimization procedure when the cost function c is fixed. The original problem we want to solve is

$$\min_\theta SD_{c,\varepsilon}(\alpha_\theta, \beta).$$

Since α_θ and β are only available through finite samples, the idea is to use Stochastic Gradient Descent (SGD). At each step we draw a minibatch $(x_i, y_i)_{i=1\dots m} \sim \alpha_\theta \otimes \beta$ and approximate $\nabla_\theta SD_{c,\varepsilon}(\alpha_\theta, \beta)$ by $\nabla_\theta SD_{c,\varepsilon}(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$. In practice, $\nabla_\theta SD_{c,\varepsilon}(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$ is further approximated by $\nabla_\theta SD^{(L)}_\varepsilon(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$ where the latter is computed by backpropagation through the computational graph of $SD^{(L)}_\varepsilon(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$.

Our approximation scheme is summarized in Figure 2.1. Samples from the generative model α_θ are obtained by applying the push-forward function g_θ to samples of the initial low-dimensional measure ζ (blue block). These samples are combined with real data (red block) to compute a pairwise distance matrix C. This matrix, as in MMD-GAN's approach (Li et al., 2015) is all we need to compute the loss. In the purple block of the figure a finite number of Sinkhorn steps (consisting of matrix-vector multiplications) are used to approximate the Sinkhorn Divergence. These Sinkhorn steps are used to evaluate (forward pass) and compute the gradient (backward pass) of our proxy $SD^{(L)}_\varepsilon(\hat{\alpha}_{\theta m}, \hat{\beta}_m)$.

Note that the procedure $\texttt{AutoDiff}_\theta$ corresponds to classical reverse mode automatic differentiation of L steps of the Sinkhorn iteration, and has therefore naturally the same complexity as Sinkhorn, *i.e.* $O(Lm^2)$ operations, with an extra storage cost required to run the backward iteration with no additional computational overhead.

When combining this with the adversarial learning of the cost function, the min-max optimization procedure is the same as (Arjovsky et al., 2017),(Li et al., 2017) and consists in alternating n_c optimization steps to train the cost function f_φ (or the dual network in (Arjovsky et al., 2017)) and an optimization step to train the generator g_θ. Following advice from these papers, we clip the weights φ to ensure a bounded gradient in the maximization and use RMSProp as an optimizer.

A discussion on biased gradients. Convergence of SGD relies on unbiased estimates of the gradient : when optimizing a function F, SGD approximates $\nabla F(\theta^{(\ell)})$ with a proxy $\nabla f_\ell(\theta^{(\ell)})$ at iteration (ℓ), where $\mathbb{E}(\nabla f_\ell(\theta)) = \nabla F(\theta)$. In the case where the gradient and the expectation can be inverted, a differentiable unbiased estimator of F yields an unbiased gradient estimate. The question of biased gradient estimates for MMD- and Wasserstein-GAN was first raised in (Bellemare et al., 2017). Subsequently,

(Bińkowski et al., 2018) demonstrated that IPM gradients always have a downward bias. Consider an IPM $d_{\mathcal{F}}(\alpha,\beta) \overset{\text{def.}}{=} \sup_{f\in\mathcal{F}} |\mathbb{E}_\alpha(f(X)) - \mathbb{E}_\beta(f(Y))|$. For a fixed function f, $\frac{1}{n}\sum_{i=1}^n f(x_i) - \frac{1}{n}\sum_{j=1}^n f(y_j)$ is an unbiased estimator of $\mathbb{E}_\alpha(f(X)) - \mathbb{E}_\beta(f(Y))$ (where $(x_i)_i$ and $(y_j)_j$ are sampled according to α and β respectively. But when the optimal dual function f is unknown, and thus approximated by another function $\hat{f}$ as in MMD-of W-GAN, then $\frac{1}{n}\sum_{i=1}^n \hat{f}(x_i) - \frac{1}{n}\sum_{j=1}^n \hat{f}(y_j)$ is a biased estimator of $d_{\mathcal{F}}(\alpha,\beta)$. Thus the gradient estimates are also biased. Although Sinkhorn Divergence does not fall in the framework of IPMs, it also suffers from biased gradients. The empirical Sinkhorn Divergence, which consists in computing the Sinkhorn Divergence between the empirical measures $\hat{\alpha}_m$, $\hat{\beta}_m$, is a biased estimator of the Sinkhorn Divergence. Thus, our gradient estimates are biased but the algorithm still does well in practice (see Section 4).

Algorithm 2 SGD with Auto-diff

Input: θ_0, φ_0, $(y_j)_{j=1}^n$ (the real data), m (batch size), L (fixed number of Sinkhorn iterations), ε (regularization parameter), τ (learning rate)

Output: θ (parameters of the generative model), φ (parameters of the cost function)

$\theta \leftarrow \theta_0$, $\varphi \leftarrow \varphi_0$,

for $k = 1, 2, \dots$ **do**

  for $t = 1, 2, \dots, n_c$ **do** (inner loop to update cost function)

    Sample $(y_j)_{j=1}^m$ from the dataset

    Sample $(z_i)_{i=1}^m \overset{\text{i.i.d}}{\sim} \zeta$, $(x_i)_{i=1}^m \overset{\text{def.}}{=} g_\theta(z_1^m)$

    $SD_{\varphi,\varepsilon}^{(L)}(x_1^m, y_1^m) \overset{\text{def.}}{=} \left(2W_{\varphi,\varepsilon}^{(L)}(x_1^m, y_1^m) - W_{\varphi,\varepsilon}^{(L)}(x_1^m, x_1^m) - W_{\varphi,\varepsilon}^{(L)}(y_1^m, y_1^m)\right)$

    (compute Sinkhorn Divergence with Algo. 1)

    $\texttt{grad}_\varphi \leftarrow \texttt{AutoDiff}_\varphi\left(SD_{\varphi,\varepsilon}^{(L)}(x_1^m, y_1^m)\right)$ (gradient evaluation with autodiff)

    $\varphi \leftarrow \varphi + \tau\texttt{RMSProp}(\texttt{grad}_\varphi)$. (gradient step with RMSprop)

    $\varphi \leftarrow \texttt{clip}(\varphi, -c, c)$

  end for

  Sample $(y_j)_{j=1}^m$ from the dataset

  Sample $(z_i)_{i=1}^m \overset{\text{i.i.d}}{\sim} \zeta$, $(x_i)_{i=1}^m \overset{\text{def.}}{=} g_\theta(z_1^m)$

  $SD_{\varphi,\varepsilon}^{(L)}(x_1^m, y_1^m) \overset{\text{def.}}{=} \left(2W_{\varphi,\varepsilon}^{(L)}(x_1^m, y_1^m) - W_{\varphi,\varepsilon}^{(L)}(x_1^m, x_1^m) - W_{\varphi,\varepsilon}^{(L)}(y_1^m, y_1^m)\right)$

  (compute Sinkhorn Divergence with Algo. 1)

  $\texttt{grad}_\theta \leftarrow \texttt{AutoDiff}_\theta\left(SD_{\varphi,\varepsilon}^{(L)}(x_1^m, y_1^m)\right)$ (gradient evaluation with autodiff)

  $\theta \leftarrow \theta - \tau\texttt{RMSProp}(\texttt{grad}_\theta)$. (gradient step with RMSprop)

  $\theta \leftarrow \texttt{clip}(\theta, -c, c)$

end for

4 Applications

We start by comparing ED_p (MMD induced by Euclidean cost, see (16)), W_ε and $SD_{c,\varepsilon}$ on a simple fitting task on synthetic data in 2D and 3D. We then consider two

popular problems in machine learning to illustrate the versatility of our method. The first one relies on fitting labeled data with uniform distributions supported on ellipses (note that this could be any parametric shape but ellipses here were a good fit). The second problem consists in tuning a neural network to generate images, first with a fixed cost (on MNIST dataset) and then with a parametric cost (on CIFAR10 dataset). In both cases, we used simple initializations (see details below) and the algorithm yielded similar results when rerun, meaning that the results displayed are representative of the performance of the algorithm and that the procedure is quite stable.

4.1 Benchmark on Synthetic Problems

Since evaluating the performance of generative models is a complicated issue on real data (see discussion in Sec. 4.3), we construct a synthetic framework to get a reliable comparison of different losses. Although our method is meant to be used in high dimensions, we apply it here to two simple problems to be able to visualize the results:

- Deterministic setting : fitting a point cloud in 2D,
- Probabilistic setting: fitting an ellipse in 3D.

In the deterministic setting, we fit a finite discrete measure, to give us a better idea of the geometry of the costs without the sampling noise. We then see how our observations carry out in a probabilistic setting, with sampling and stochastic gradient descent as described in Algorithm 2.

Both problems require the use of losses that are smooth for the weak convergence: for the first one, the model measure is singular, as it is supported on points, while for the second, the model measure measure has a bounded support which changes during the optimization procedure.

Deterministic setting: Fitting a point cloud in 2D. Given a dataset $(y_1, \dots, y_n)$ of points in 2D, we want to fit the empirical measure $\beta \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. The parametric measure that we consider is thus $\alpha_\theta \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$, where the parameters θ_i are the positions of the Dirac masses. To get rid of the sampling noise and better observe the geometry of the losses, we use a full gradient descent to estimate θ (or, in other words, we use minibatches of size $m = n$, to stick to the scheme of Algorithm 2). The results are given in Figure 2.2. We use a cost $c = \|\cdot\|_2^{1.5}$ to ensure fair comparison with ED_p which is well defined for $p < 2$ only. We use $L = 5$ iterations to to Sinkhorn's algorithm. We can see that $W_{c,\varepsilon}$ successfully captures the extreme points, but yields parameters that collapse to a mean values in the dense area. The Energy Distance, on the other hand, fails to capture the extreme points but the points in the dense region are well distributed. Sinkhorn Divergences get the best of both worlds by successfully capturing both the extreme points and the dense area for $\varepsilon = 1$, but when ε gets larger, we recover the behavior of Energy Distance and the extreme points are not recovered anymore.

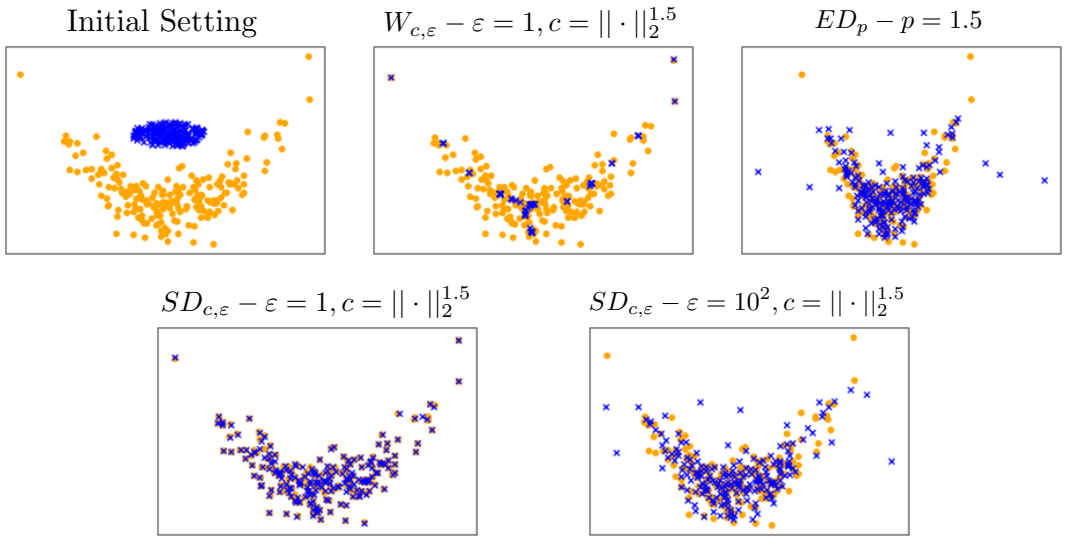


Figure 2.2 – Comparison of $SD_{c,\varepsilon}$ to $W_{c,\varepsilon}$ and ED_p on a deterministic task: fitting a point cloud in 2D. The orange circles represent the target distribution β, and the blue crosses the fitted distribution α_{θ^*}. The top left image represents the initial distribution α_{θ_0}.

Probabilistic setting: fitting an ellipse in 3D. We consider a parametric measure α_θ that generates points uniformly inside an ellipse. The ellipse parametrized by a 3×3 matrix A (the square root of its covariance matrix) and a center $\omega \in \mathbb{R}^3$, so that $\theta = (A, \omega)$. The reference measure ζ is a uniform on the unit ball of dimension 3, and a point is sampled from α_θ thanks to $g_\theta(z) = Az + \omega$. We generate $n = 200$ datapoints from α_{θ_0}, where $\theta_0 = (A_0, \omega_0)$, and thus the distribution of the data, denoted by β in the previous sections, is known and equal to α_{θ_0}. This allows us to evaluate the performance of the loss by looking at the inferred parameters θ^* for each loss and comparing them to θ_0.

To illustrate the effect of the normalization introduced by Sinkhorn Divergences, and the interpolation property, we consider the following losses:

- Sinkhorn Divergences: $SD_{c,\varepsilon}$ for $c = \|\cdot\|_2^p$ and $p \in \{1.5, 2\}$
- Entropy-Regularized OT: $W_{c,\varepsilon}$ (primal cost without entropy) for $c = \|\cdot\|_2^p$ and $p \in \{1.5, 2\}$
- Energy Distance: ED_p for $p \in \{1.5, 2\}$

The choice $c = \|\cdot\|_2^p$ with $p = 2$ is called the *quadratic cost* and is used "by default" for OT, as the associated distance (the 2-Wasserstein distance) is well studied in literature and known to have good properties (see (Santambrogio, 2015) for details). However, by Proposition 16 it doesn't induce a positive definite kernel, thus the associated Energy

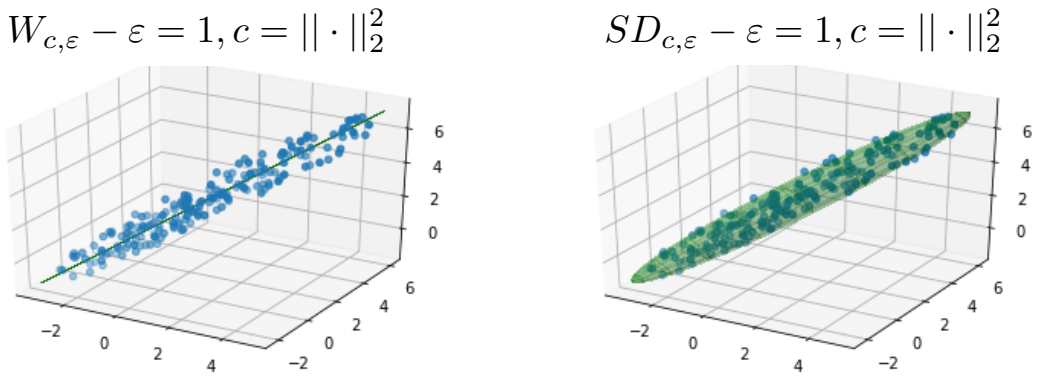


Figure 2.3 – Comparison of $SD_{c,\varepsilon}$ to $W_{c,\varepsilon}$ on a synthetic benchmark : fitting data generated uniformly inside an ellipse. The blue dots represent the points generated from the reference ellipse with covariance A_0 and center ω_0, while the green ellipse is drawn with the inferred parameters A^*, ω^* for each loss (see Table 2.1).

Distance is not positive definite. We thus resort to another cost function $c = \|\cdot\|_2^p$ with $p = 1.5$.

The results of the fitting procedure are given in Figures 2.3 and 2.4. When comparing $SD_{c,\varepsilon}$ to $W_{c,\varepsilon}$ the benefits of the normalization are clear (Figure 2.3). While $W_{c,\varepsilon}$ yields good performance for small values of ε, when for larger values the fitted ellipse collapses to the mean of the values along one axis, before collapsing to a centroid for even larger values. On the other hand, the results obtained with $SD_{c,\varepsilon}$ are robust to the value of ε. Since Sinkhorn's algorithm converges much faster for larger values of ε (the convergence rates can be found in Chapter 1, Sec. 4.2), using $SD_{c,\varepsilon}$ with larger ε allows a consequent computational speedup in the inference procedure, which more than compensates for the added time to compute the normalizing factors.

For $p = 2$, we observe that the Energy Distance yields a degenerate ellipse (Figure 2.4, top-left). As this is the limit case of $SD_{c,\varepsilon}$ for $\varepsilon \to +\infty$, we also observe this behavior for the Sinkhorn Divergence with large values of ε (Figure 2.4, bottom-left). However for smaller values, the fitting is correct (Figure 2.3, right). Now for $p = 1.5$, we can fairly compare Sinkhorn Divergences (Figure 2.4, bottom-right) and the Energy Distance (Figure 2.4, top-right), since the latter is positive definite. From a visual point of view, both losses yield satisfactory results, as ellipses don't collapse to a single point when ε grows. To get a better insight, we consider the values of the inferred parameters. We observe that the best results are for $\varepsilon = 1$. To ensure the robustness of this observation, we run the inference procedure multiple times for each loss, and use the same dataset $\mathcal{D}_i$ for all losses in each trial run i.

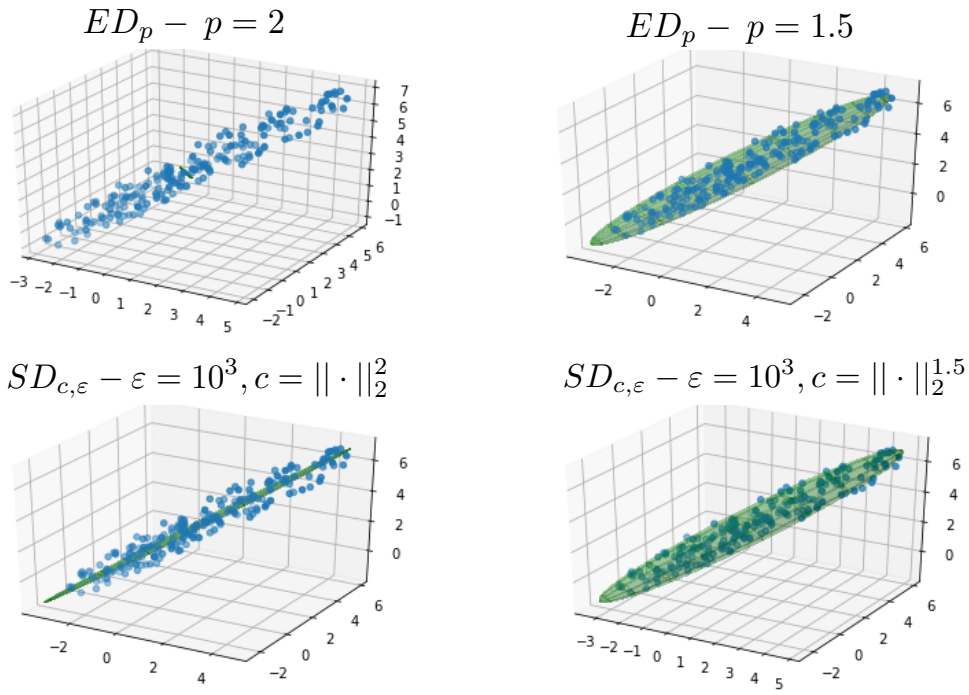


Figure 2.4 – Comparison of $SD_{c,\varepsilon}$ and ED_p on a synthetic benchmark (same setting as Figure 2.3. In the case where the ellipse is not visible, it has collapses to its centroid ω^*, with A^* very close to 0 (see Table 2.1 for numerical values).

4.2 Data Clustering with Ellipses

As already mentioned a strength of the Wasserstein distance is its ability to fit a singular probability distribution to an empirical measure (data). That singular probability may be supported on a subset of the space on a lower dimensional manifold, or simply have a degenerate density that becomes null for some subsets of the original space. To illustrate this principle, we consider in what follows a simple 3D example that can easily be visualized.

We use the Iris dataset (3 classes, 50 observations each in 4 dimensions) projected in 3D using PCA. This defines the dataset $(y_1, \ldots, y_n)$ in $\mathbb{R}^3$, with $n = 150$. If we were to find a probability distribution α_θ bound to be itself an empirical measure of K atoms (in that case parameter θ would contain exactly the locations of those K points in addition to their weight), then minimizing the 2-Wasserstein distance of α_θ to β would be *strictly equivalent* to the K-means problem (Canas and Rosasco, 2012). In that sense, quantization can be regarded as the most elementary example of Wasserstein loss minimization of families of singular probability distributions.

The model we consider is instead composed of $K = 3$ ellipses with uniform density. As in the previous benchmark section, each ellipse is parametrized by a 3×3 matrix A_k (the square root of its covariance matrix) and a center $\omega_k \in \mathbb{R}^3$, so that $\theta = (A_k, \omega_k)_k$.

loss	ED_p	ED_p	$SD_{c,\varepsilon}$
p, ε	2,-	1.5,-	$2,10^3$
A^*	$-0.09 \quad -0.04 \quad 0.05$ $0.06 \quad 0.03 \quad 0.05$ $-0.09 \quad -0.17 \quad -0.11$	$3.12 \quad 1.74 \quad 2.08$ $2.25 \quad 2.83 \quad 2.09$ $2.30 \quad 1.74 \quad 3.07$	$1.56 \quad 2.23 \quad 2.69$ $1.44 \quad 2.31 \quad 2.72$ $1.40 \quad 2.22 \quad 2.86$
ω^*	(0.68, 1.78 , 2.72)	(0.63 , 1.75 , 2.75)	(0.74 , 1.81 , 2.76)

loss	$SD_{c,\varepsilon}$	$SD_{c,\varepsilon}$	**ground truth**
p, ε	$1.5,10^3$	2, 1	
A^*	$2.95 \quad 2.08 \quad 2.05$ $2.05 \quad 3.17 \quad 1.95$ $2.12 \quad 2.15 \quad 3.00$	$2.90 \quad 1.96 \quad 2.13$ $2.02 \quad 3.03 \quad 2.10$ $2.06 \quad 1.95 \quad 3.03$	$3 \quad 2 \quad 2$ $2 \quad 3 \quad 2$ $2 \quad 2 \quad 3$
ω^*	(0.73 ,1.83, 2.76)	(0.94 , 1.96 , 2.90)	**(1,2,3)**

Table 2.1 – Comparison of the inferred parameters A^*, ω^* for the losses displayes in Figures 2.3 and 2.4, the gound truth A_0, ω_0 (parameters used to generate the dataset) is in bold, on the right.

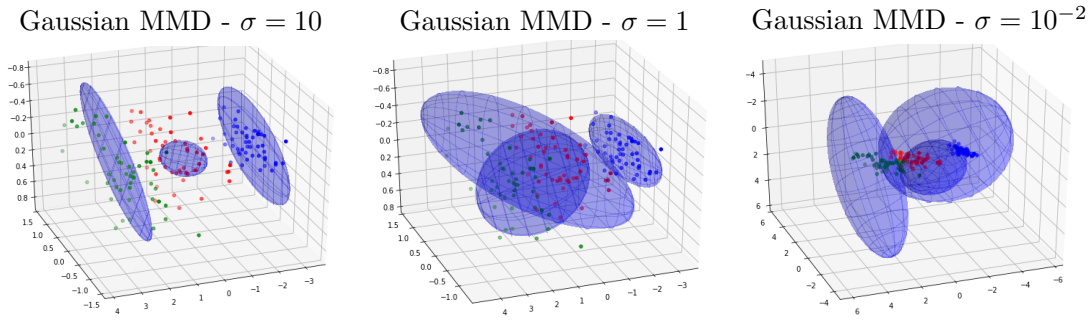


Figure 2.5 – Ellipses after convergence of the stochastic gradient descent, with Gaussian MMD.

Therefore, our results can not be directly compared to that of clustering algorithms, in the sense that we do automatically recover, within such ellipses, entire areas of interest (and not voronoi cells). We assume in this illustration that each ellipse has equal mass $1/K$. To recover these ellipses through a push forward, we use a uniform ground density ζ over 3 centered unit balls, translated and dilated for each ellipse using the push-forward defined by $g_\theta(z) = A_k z + \omega_k$ if z is in the k-th ball. Note that the model can be adapted otherwise (density decaying when moving away from the center, mass proportional to the size of the ellipse) with simple modifications in either the ground density ζ or the pushforward g_θ, but we found this uniform model to be a good fit for this dataset.

4.2.0.1 Numerical Illustration. The ellipse matrices $(A_k)_k$ are all initialized with the identity matrix (which corresponds to the unit ball) and centers $(\omega_k)_k$ are initialized with the K-means algorithm. We fixed a maximal budget of Sinkhorn iterations $L = 5$ to be competitive with MMD time-wise, with a minibatch size $m = 300$ for both algo-

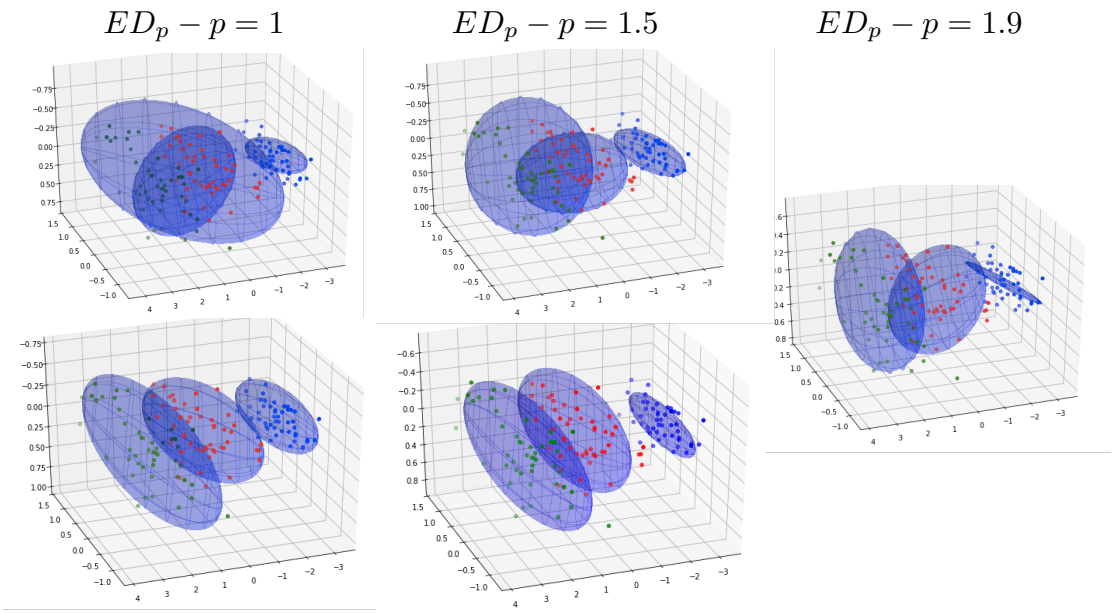


Figure 2.6 – Ellipses after convergence of the stochastic gradient descent, with Energy Distance for varying p. Top row displays cases where the algorithm is stuck in a local minimum which does not correspond to the classes, while the bottom row displays successful cases. When p gets close to 2, the performance decays.

rithms. We display the results of the clustering algorithm for each losses, with varying parameters: MMD with a Gaussian kernel with varying bandwidth σ in Figure 2.5, the Energy Distance with varying distance $\|\cdot\|_2^p$ in Figure 2.6, and Sinkhorn Divergences with varying ε and cost function $c = \|\cdot\|_2^p$ in Figure 2.7.

- The Gaussian kernel yields poor results for all tested σ. When the bandwidth is large, the ellipses only fit a small number of points in the center of the classes, while with a small bandwidth the ellipses get too big, trying to fit all the points. In between, the algorithm gets stuck in local minima, and can not grasp the right class structure even after several re-runs.

- The Energy Distance, on the other hand, performs much better for the right choice of p ($p \in 1, 1.5$ in the plots). However, when getting close to $p = 2$ ($p \in 1.9, 2$ in the plots), the ellipses just capture the centers of the clusters. This suggests that $c = \|\cdot\|_2^p$ for $p < 2$ might be a good ground metric for this problem.

- Based on the performance of different distances from the Energy Distance, we assess the performance of Sinkhorn Divergences for various values of p. When using a medium regularization ($\varepsilon = 1$), the performance is robust to the cost function used. However as seen in the benchmark, using a large ε entails a behavior close to MMD and thus require using cost functions with $p < 2$. On the other hand, in the interest of keeping a low computational budget (which depends on the number of

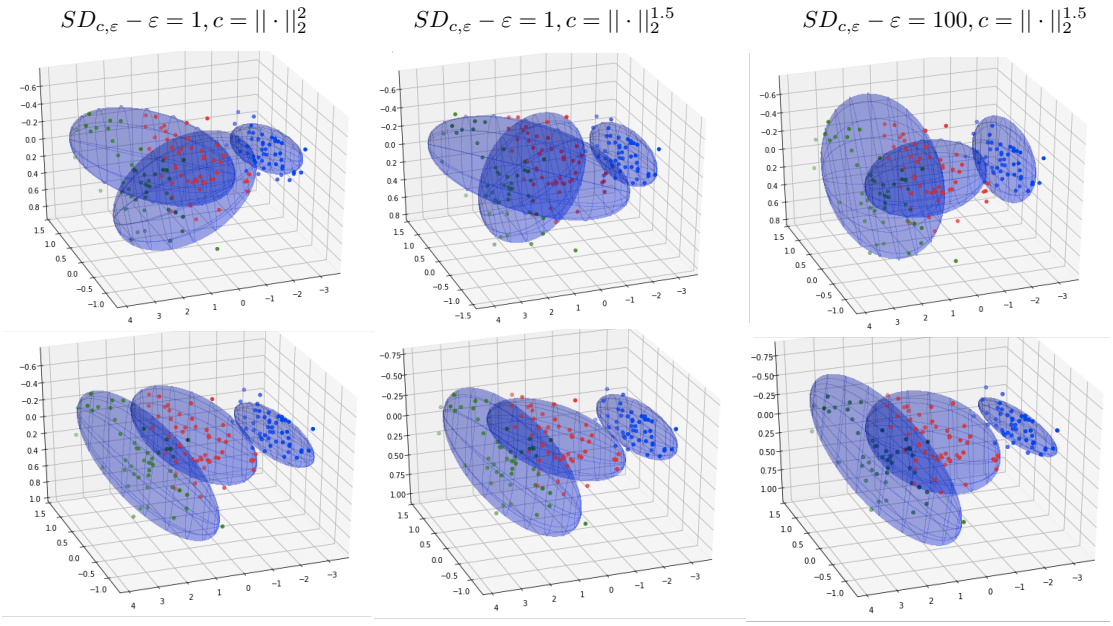


Figure 2.7 – Ellipses after convergence of the stochastic gradient descent, with Sinkhorn Divergence for varying ε and p. Top row displays cases where the algorithm is stuck in a local minimum which does not correspond to the classes, while the bottom row displays successful cases. We give a budget of $L = 5$ Sinkhorn Iterations to compute $SD_{c,\varepsilon}$ to have a fast algorithm, which prevents from using small ε.

Sinkhorn iterations L), smaller regularizations should not be used as they require a lot more Sinkhorn iterations to converge.

Both the Energy Distance and Sinkhorn Divergences can get stuck in local minima. In this experiment, we observe that $p = 1.5$ yields the most reliable results, meaning that the geometry of the dataset is correctly recovered 90% of the time with ED_p and $SD_{c,\varepsilon}$ when ε is large enough. We found that Sinkhorn Divergences with $\varepsilon = 100$ were less prone to fall in local minima than $\varepsilon = 1$ in our setting, and thus gave better performance in average which might be due to the improved sample complexity (see Chapter 3, Theorem 14 where we show how ε affects the sample complexity of Sinkhorn Divergences).

Since the Iris data is labeled, we can asses the fit of the model by checking the class repartition in each ellipse, as summarized in table 2.2. Each entry (i, j) corresponds to the number of points from class j that are inside ellipse i (recall there are 50 points per class). As the Energy Distance is the limit case of Sinkhorn Divergence for $\varepsilon \to +\infty$, it can be used to chose the right exponent for the cost function $c = \|\cdot\|_2^p$, which here is $p = 1.5$. Then the parameter ε gives one more degree of freedom, which can be chosen via cross-validation. In this setting, the advantage of Sinkhorn Divergences over the Energy Distance is not clear: the performance for the blue class is best for $\varepsilon = 1$, with a high coverage and small variance, but it is unstable for the green and red classes, often

loss	ED_p			$SD_{c,\varepsilon}$			$SD_{c,\varepsilon}$			$SD_{c,\varepsilon}$		
p, ε	1.5,-			1.5,10^2			1.5,10			1.5,1		
mean	29.4	0	0	29.9	0	0	30.7	0	0	36.2	0	0
	0	30.8	4.7	0	30.1	3.8	0	31.6	4.4	0	29.5	27.9
	0	4.3	33.6	0	3.1	34.0	0	3.5	34.9	0	18.5	31.5
sd	*9.1*	*0*	*0*	*9.79*	*0*	*0*	*8.55*	*0*	*0*	*3.09*	*0*	*0*
	0	*8.69*	*7.33*	*0*	*8.52*	*6.44*	*0*	*7.07*	*7.94*	*0*	*5.53*	*15.8*
	0	*5.82*	*7.84*	*0*	*4.75*	*7.63*	*0*	*5.21*	*8.35*	*0*	*11*	*5.65*

Table 2.2 – Evaluation of the fit after convergence of the algorithm : entry (i, j) corresponds to the number of points from class j that are inside ellipse i (1= blue class, 2 = red class, 3 = green class). We take the average for each loss over 100 runs and give the standard deviation.

mixing up both classes. When $\varepsilon = 100$, the fit for the green and red classes improves but it degrades for the blue class – the performance is not significantly different from the Energy Distance.

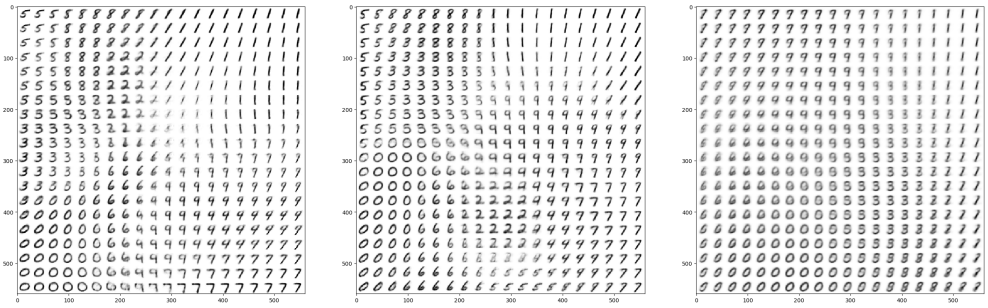
4.3 Tuning a Generative Neural Network

Image generating models such as GAN (Goodfellow et al., 2014) or VAE (Kingma and Welling, 2013) have become popular in recent years. The goal is to train a neural network g_θ which generates images $g_\theta(z)$ that resemble a certain data set $(y_j)_j$, given a random input z in a latent space $\mathcal{Z}$. Both methods require a second network for the training of the generative network (an adversarial network in the case of GANs, an encoding network in the case of VAEs). Depending on the complexity of the data, our method can rely on the generative network alone by directly comparing its output with the data in Wasserstein distance.

4.3.1 With a Fixed Cost c.

This section fits a generative model where the pushforward g_θ is a multilayer perceptron. We begin with experiments on the MNIST dataset, which is a standard benchmark for this type of networks. Since the dataset is relatively simple, learning the cost is superfluous here and we use the ground cost $c(x, y) = \|x - y\|^2$, which is sufficient for these low resolution images and also the baseline in (Kingma and Welling, 2013). We use as g_θ a multilayer perceptron with 2 fully connected layers and the latent space is the unit square $\mathcal{Z} = [0, 1]^2$ over which we put a uniform distribution. Learning is performed in mini-batches over the MNIST dataset, with the Adam optimizer (Kingma and Ba, 2014).

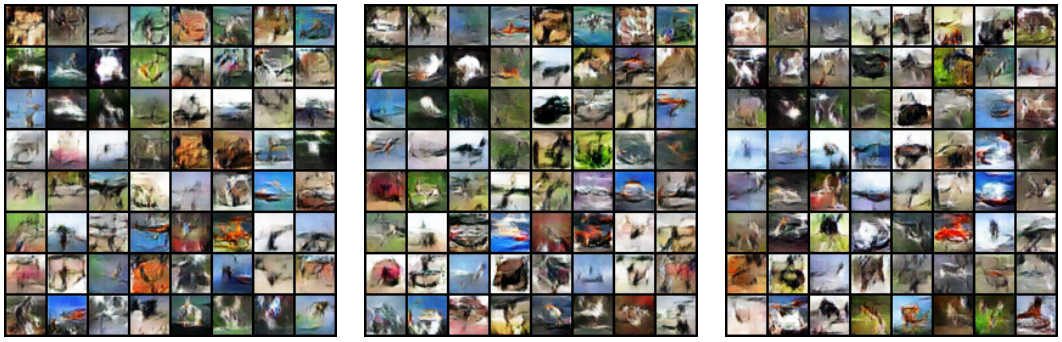
Figure 2.8 displays the manifold of images $g_\theta(z)$ generated by the optimized network after the learning procedure for different values of the hyperparameters (ε, m, L). This manifold is obtained by computing $g_\theta(z_1^i, z_2^j)$ for equi-spaced $(z_1^i, z_2^j) \in [0, 1]^2$, and then



(a) $\varepsilon = 1, m = 200,$ $L = 10$ (b) $\varepsilon = 10^{-1}, m = 200,$ $L = 100$ (c) $\varepsilon = 10^{-1}, m = 10,$ $L = 300$

Figure 2.8 – Influence of the hyperparameters on the manifold of generated digits.

plotting the resulting digit at location (i, j) in the larger picture. This shows that the regularization parameter ε can be chosen quite large, which in turn leads to a fast convergence of Sinkhorn iterations. Indeed, using $\varepsilon = 1$ with only $L = 10$ Sinkhorn iterations (image (a)) yields a result similar to using $\varepsilon = 0.1$ with $L = 100$ iterations (image (b)). Regarding the size m of the mini-batches, a too small m value (e.g. $m = 10$) leads to poor results, and we observe that $m = 200$ is sufficient to learn accurately the manifold.



(a) MMD (b) $\varepsilon = 100$ (c) $\varepsilon = 1$

Figure 2.9 – Samples from the generator trained on CIFAR 10 for MMD and Sinkhorn Divergence

4.3.2 Learning the Cost.

With higher-resolution datasets, such as classical benchmarks CIFAR10 or CelebA, using the ℓ^2 metric between images yields very poor results. It tends to generate images which are basically a blur of similar images. The alternative, already outlined in Algorithm 1 relies on learning another network which encodes meaningful feature vectors for the images, between which the Euclidean distance can be computed.

We compare our loss with different values for the regularization parameter ε to the results obtained with an MMD loss with a Gaussian kernel, as this is the one used

MMD (Gaussian)	$\varepsilon = 100$	$\varepsilon = 10$	$\varepsilon = 1$
4.56 ± 0.07	4.81 ± 0.05	4.79 ± 0.13	4.43 ± 0.07

Table 2.3 – Inception Scores on CIFAR10. We use the experimental setting of (Li et al., 2017), and use their optimized parameters for MMD, which is computed with a Gaussian kernel.

in (Li et al., 2017). We based the experiments on their code, and thus used their optimized parameters for MMD to carry out a fair comparison. Both the generator and discriminator networks follow the DCGAN architecture (Radford et al., 2015). We use small batches of 100 images.

Generative models are very hard to evaluate and there is no consensus on which metric should be used to assess their quality. We use the inception score introduced in (Salimans et al., 2016) as it is widespread, and also the reference in (Dziugaite et al., 2015) against which we compare our approach. The inception score is based on two key aspects of data generation: realism and diversity. Using a pre-trained classifier, the algorithm computed the conditional class probability of the samples: this should have high entropy, meaning the classifier recognizes the object in the generated image. On the other hand, the distribution of the classes should have low entropy, so that all classes are well represented. However, the inception score does not account for the failure cases of mode collapse (generating only one image per class) or overfitting (generating only copies of images from the dataset). The Frechet Inception Distance has since been introduced as an alternative (Heusel et al., 2017), and seems to be preferred by the community. It roughly consists in embedding the true data and the generated data to a feature space (via a pre-trained network) and comparing the resulting distributions in terms of mean and variance. It is robust to mode collapse, but not to overfitting. See the recent survey paper by (Lucic et al., 2018) for a good insight on metrics to evaluate GANs.

Table 2.3 summarizes the inception scores on CIFAR10 for MMD and Sinkhorn Divergence with varying ε. The scores are evaluated on 20000 random images. Figure 2.9 displays a few of the associated samples (generated with the same seed). Although there is no striking difference in visual quality, the model with a Sinkhorn Divergence and a large regularization is the one with the best score. The decaying scores of models which have a loss closer to the true OT loss can be explained by two main factors : (i) the number of iterations required for the convergence of Sinkhorn with such ε might exceed the total iteration budget that we give the algorithm to compute the loss (to ensure reasonable training time of the model), (ii) it reflects the fact that sample complexity worsens when we get closer to OT metrics, and increasing the batch size might be beneficial in that case. We give theoretical grounds for the latter in Chapter 3.

Chapter 3

Sample Complexity of Sinkhorn Divergences

Optimal Transport (OT) and Maximum Mean Discrepancies (MMD) are now routinely used in machine learning to compare probability measures. We focus in this chapter on Sinkhorn Divergences *(SDs), a regularized variant of OT distances which can interpolate, depending on the regularization strength ε, between OT ($\varepsilon = 0$) and MMD ($\varepsilon = \infty$). Although the tradeoff induced by the regularization is now well understood computationally (OT, SDs and MMD require respectively $O(n^3 \log n)$, $O(n^2)$ and $O(n^2)$ operations given a sample size n), much less is known in terms of their* sample complexity*, namely the gap between these quantities, when evaluated using finite samples* vs. *their respective densities. Indeed, while the sample complexity of OT and MMD stand at two extremes, $O(1/n^{1/d})$ for OT in dimension d and $O(1/\sqrt{n})$ for MMD, that for SDs has only been studied empirically. In this chapter, we*

(i) derive a bound on the approximation error made with SDs when approximating OT as a function of the regularizer ε,

(ii) prove that the optimizers of regularized OT are bounded in a Sobolev (RKHS) ball independent of the two measures,

(iii) provide the first sample complexity bound for SDs, obtained by reformulating SDs as a maximization problem in a RKHS. We thus obtain a scaling in $1/\sqrt{n}$ (as in MMD), with a constant that depends however on ε, making the bridge between OT and MMD complete.

This chapter is based on (Genevay et al., 2019 (to appear)).

1 Introduction

OT has been long neglected in data sciences for two main reasons, which could be loosely described as *computational* and *statistical*. Following the seminal work by (Cuturi, 2013), we have showed in Chapters 1 and 2 how entropic regularization of OT alleviates this computational burden. In Chapter 2, we further mentioned that entropy-regularized OT seems to break the curse-of-dimensionality from which OT suffers based on empirical evidence, and the goal of this chapter is to make it more formal through a sample complexity result.

Previous Works. The central theoretical contribution of Chapter 2 (see Theorem 10) states that Sinkhorn Divergences, based on regularized OT, interpolate between OT and MMD. These two metrics, which emerged as popular candidates to compare probability measures, differ on a fundamental aspect: their sample complexity. The definition of sample complexity of a loss function that we choose here is the convergence rate of the loss evaluated on empirical measures to the loss evaluated on the "true" measures, as a function of the number of samples. This notion is crucial in machine learning, as bad sample complexity implies overfitting and high gradient variance when using these divergences for parameter estimation. In that context, it is well known that the sample complexity of MMD is independent of the dimension, scaling as $\frac{1}{\sqrt{n}}$ (Gretton et al., 2006) where n is the number of samples. In contrast, it is well known that standard OT suffers from the curse of dimensionality (Dudley, 1969): Its sample complexity is exponential in the dimension of the ambient space. Although it was recently proved that this result can be refined to consider the implicit dimension of data (Weed and Bach, 2017), the sample complexity of OT appears now to be the major bottleneck for the use of OT in high-dimensional machine learning problems.

A remedy to this problem may lie, again, in regularization. The discrepancies defined through regularized OT, known as Sinkhorn Divergences, seem to be less prone to overfitting. Indeed, a certain amount of regularization tends to improve performance in simple learning tasks (Cuturi, 2013). The interpolation theorem from Chapter 2 also suggests that for large regularizations, Sinkhorn Divergences behave like MMD.

The asymptotic behavior of empirical estimates of the Wasserstein distance has been widely studied, from convergence rates to distributional limits. In particular we refer to (Del Barrio and Loubes, 2017), which recently proved a central limit theorem for empirical OT in general dimension, for a thorough historical review of the subject. However, aside from a recent central limit theorem in the case of measures supported on finite discrete spaces (Bigot et al., 2017), the convergence of empirical Sinkhorn Divergences, and more generally their sample complexity, remains an open question.

Contributions. This chapter provides three main contributions, which all exhibit theoretical properties of Sinkhorn Divergences. Our first result is a bound on the speed of convergence of regularized OT to standard OT as a function of the regularization parameter, in the case of continuous measures. The second theorem proves that the optimizers of the regularized optimal transport problem lie in a Sobolev ball which is independent of the measures. This allows us to rewrite the Sinkhorn Divergence as the maximization of an expectation over a RKHS ball and thus justify the use of kernel-SGD for regularized OT as advocated in Chapter 4, Sec. 5.1. As a consequence of this reformulation, we provide as our third contribution a sample complexity result. We focus on how the sample size and the regularization parameter affect the convergence of the empirical Sinkhorn Divergence (i.e., computed from samples of two continuous measures) to the continuous Sinkhorn Divergence. We show that the Sinkhorn Divergence benefits from the same sample complexity as MMD, scaling in $\frac{1}{\sqrt{n}}$ but with a constant that depends on the inverse of the regularization parameter. Thus sample complexity worsens when getting closer to standard OT, and there is therefore a tradeoff between a good approximation of OT (small regularization parameter) and fast convergence in terms of sample size (larger regularization parameter). We conclude this chapter with a few numerical experiments to asses the dependence of the sample complexity on ε and d in simple cases.

2 Reminders on Sinkhorn Divergences

We consider entropy-regularized optimal transport between two probability measures $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and β on $\mathcal{M}_+^1(\mathcal{Y})$, as introduced in Chapter 1, Sec. 4, where $\mathcal{X}$ and $\mathcal{Y}$ are two bounded subsets of $\mathbb{R}^d$. The optimal transport problem is regularized with the relative entropy of the transport plan with respect to the product measure $\alpha \otimes \beta$ following (Genevay et al., 2016):

$$W_\varepsilon(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon H(\pi \mid \alpha \otimes \beta), \tag{$\mathcal{P}_\varepsilon$}$$

$$\text{where} \quad H(\pi \mid \alpha \otimes \beta) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right) \mathrm{d}\pi(x,y), \tag{2.1}$$

and the feasible set $\Pi(\alpha, \beta)$ is composed of probability distributions over the product space $\mathcal{X} \times \mathcal{Y}$ with fixed marginals α, β. The cost function c, which represents the cost to move a unit of mass from x to y is assumed to be $\mathcal{C}^\infty$ through this chapter (more specifically, we need it to be $\mathcal{C}^{\frac{d}{2}+1}$). Choosing the relative entropy as a regularizer allows to express the dual formulation of regularized OT as the maximization of an expectation

problem (Proposition 4 in Chapter 1):

$$\begin{aligned} W_\varepsilon(\alpha,\beta) &= \max_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\beta(y) \\ &\qquad - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon \\ &= \max_{u\in\mathcal{C}(X),v\in\mathcal{C}(Y)} \mathbb{E}_{\alpha\otimes\beta}\left[f_\varepsilon^{XY}(u,v)\right] + \varepsilon, \end{aligned}$$

where $f_\varepsilon^{xy}(u,v) = u(x) + v(y) - \varepsilon e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}$.

This reformulation as the maximum of an expectation is crucial to obtain sample complexity results. The existence of optimal dual potentials (u, v) is proved in Chapter 1, Sec. 4.1. They are unique $\alpha-$ and $\beta-$a.e. up to an additive constant.

To correct for the fact that $W_\varepsilon(\alpha,\alpha) \neq 0$, we introduced Sinkhorn Divergences in Chapter 2. They are a natural normalization of that quantity defined as

$$SD_\varepsilon(\alpha,\beta) = W_\varepsilon(\alpha,\beta) - \frac{1}{2}(W_\varepsilon(\alpha,\alpha) + W_\varepsilon(\beta,\beta)). \tag{2.2}$$

This normalization ensures that $SD_\varepsilon(\alpha,\alpha) = 0$, but also has a noticeable asymptotic behavior as proved in Theorem 10 of Chapter 2. Indeed, when $\varepsilon \to 0$ one recovers the original (unregularized) OT problem, while choosing $\varepsilon \to +\infty$ yields the maximum mean discrepancy (see Chapter 1, Sec.2.2 for a detailed introduction on the matter) associated to the kernel $k = -c/2$, where MMD is defined by:

$$MMD_k(\alpha,\beta) = \mathbb{E}_{\alpha\otimes\alpha}[k(X,X')] + \mathbb{E}_{\beta\otimes\beta}[k(Y,Y')] - 2\mathbb{E}_{\alpha\otimes\beta}[k(X,Y)].$$

Besides, under some assumptions on the cost function, Sinkhorn Divergences are positive definite and metrize weak convergence of measures (Feydy et al., 2019 (to appear)). In the context of this chapter, we study in detail the sample complexity of $W_\varepsilon(\alpha,\beta)$, which immediately extends to that of $SD_\varepsilon(\alpha,\beta)$ by linearity.

Remark 14. Sinkhorn Divergences can be defined with W_ε being either the primal cost of $(\mathcal{P}_\varepsilon)$ or the primal cost *without* the entropic term (see Definition 9 in Chapter 2 for more details). While the interpolation theorem in Chapter 2 holds for both definitions, the sample complexity theorem that we prove here is only valid when W_ε is the primal cost of $(\mathcal{P}_\varepsilon)$.

3 Approximating Optimal Transport with Sinkhorn Divergences

In the present section, we are interested in bounding the error made when approximating $W(\alpha,\beta)$ with $W_\varepsilon(\alpha,\beta)$.

Theorem 12. *Let α and β be probability measures on $\mathcal{X}$ and $\mathcal{Y}$ subsets of $\mathbb{R}^d$ such that $|\mathcal{X}|$ and $|\mathcal{Y}| \leqslant D$ and assume that c is L-Lipschitz w.r.t. x and y. It holds*

$$0 \leqslant W_\varepsilon(\alpha,\beta) - W(\alpha,\beta) \leqslant 2\varepsilon d \log\left(\tfrac{e^2 \cdot L \cdot D}{\sqrt{d}\cdot\varepsilon}\right) \tag{3.1}$$

$$\sim_{\varepsilon\to 0} 2\varepsilon d \log(1/\varepsilon). \tag{3.2}$$

Proof. For a probability measure π on $\mathcal{X}\times\mathcal{Y}$, we denote by $C(\pi) = \int c \,\mathrm{d}\pi$ the associated transport cost and by $H(\pi)$ its relative entropy with respect to the product measure $\alpha\otimes\beta$ as defined in (2.1). Choosing π_0 a minimizer of $\min_{\pi\in\Pi(\alpha,\beta)} C(\pi)$, we will build our upper bounds using a family of transport plans with finite entropy that approximate π_0. The simplest approach consists in considering block approximation. In contrast to the work of Carlier et al. (2017), who also considered this technique, our focus here is on quantitative bounds.

Definition 10 (Block approximation)**.** *For a resolution $\Delta > 0$, we consider the block partition of $\mathbb{R}^d$ in hypercubes of side Δ defined as*

$$\{Q_k^\Delta = [k_1\cdot\Delta, (k_1+1)\cdot\Delta[\times \ldots [k_d\cdot\Delta, (k_d+1)\cdot\Delta[\,; k=(k_1,\ldots,k_d)\in\mathbb{Z}^d\}.$$

To simplify notations, we introduce $Q_{ij}^\Delta \overset{\text{def.}}{=} Q_i^\Delta \times Q_j^\Delta$, $\boldsymbol{\alpha}_i^\Delta \overset{\text{def.}}{=} \alpha(Q_i^\Delta)$, $\boldsymbol{\beta}_j^\Delta \overset{\text{def.}}{=} \beta(Q_j^\Delta)$. The block approximation of π_0 of resolution Δ is the measure $\pi^\Delta \in \Pi(\alpha,\beta)$ characterized by

$$\pi^\Delta|_{Q_{ij}^\Delta} = \frac{\pi_0(Q_{ij}^\Delta)}{\boldsymbol{\alpha}_i^\Delta\cdot\boldsymbol{\beta}_j^\Delta}(\alpha|_{Q_i^\Delta}\otimes\beta|_{Q_j^\Delta})$$

for all $(i,j)\in(\mathbb{Z}^d)^2$, with the convention $0/0 = 0$.

π^Δ is nonnegative by construction. Observe also that for any Borel set $B\subset\mathbb{R}^d$, one has

$$\pi^\Delta(B\times\mathbb{R}^d) = \sum_{(i,j)\in(\mathbb{Z}^d)^2} \frac{\pi_0(Q_{ij}^\Delta)}{\boldsymbol{\alpha}_i^\Delta\cdot\boldsymbol{\beta}_j^\Delta}\cdot\alpha(B\cap Q_i^\Delta)\cdot\boldsymbol{\beta}_j^\Delta = \sum_{i\in\mathbb{Z}^d}\alpha(B\cap Q_i^\Delta) = \alpha(B),$$

which proves, using the symmetric result in β, that π^Δ belongs to $\Pi(\alpha,\beta)$. As a consequence, for any $\varepsilon > 0$ one has $W_\varepsilon(\alpha,\beta) \leqslant C(\pi^\Delta) + \varepsilon H(\pi^\Delta)$. Recalling also that the relative entropy H is nonnegative over the set of probability measures, we have the bound

$$0 \leqslant W_\varepsilon(\alpha,\beta) - W(\alpha,\beta) \leqslant (C(\pi^\Delta) - C(\pi_0)) + \varepsilon H(\pi^\Delta).$$

We can now bound the terms in the right-hand side, and choose a value for Δ that minimizes these bounds.

The bound on $C(\pi^\Delta) - C(\pi_0)$ relies on the Lipschitz regularity of the cost function.

Using the fact that $\pi^\Delta(Q_{ij}^\Delta) = \pi_0(Q_{ij}^\Delta)$ for all i, j, it holds

$$\begin{aligned} C(\pi^\Delta) - C(\pi_0) &= \sum_{(i,j)\in(\mathbb{Z}^d)^2} \pi_0(Q_{ij}^\Delta)\Big(\sup_{x,y\in Q_{ij}^\Delta} c(x,y) - \inf_{x,y\in Q_{ij}^\Delta} c(x,y)\Big) \\ &\leqslant 2L\Delta\sqrt{d}, \end{aligned}$$

where L is the Lipschitz constant of the cost (separately in x and y) and $\Delta\sqrt{d}$ is the diameter of each set Q_i^Δ.

As for the bound on $H(\pi^\Delta)$, using the fact that $\pi_0(Q_{ij}^\Delta) \leqslant 1$ we get

$$\begin{aligned} H(\pi^\Delta) &= \sum_{(i,j)\in(\mathbb{Z}^d)^2} \log\left(\frac{\pi_0(Q_{ij}^\Delta)}{\boldsymbol{\alpha}_i^\Delta \cdot \boldsymbol{\beta}_j^\Delta}\right) \pi_0(Q_{ij}^\Delta) \\ &\leqslant \sum_{(i,j)\in(\mathbb{Z}^d)^2} \Big(\log(1/\boldsymbol{\alpha}_i^\Delta) + \log(1/\boldsymbol{\beta}_j^\Delta)\Big)\, \pi_0(Q_{ij}^\Delta) \\ &= -H^\Delta(\alpha) - H^\Delta(\beta), \end{aligned}$$

where we have defined $H^\Delta(\alpha) = \sum_{i\in\mathbb{Z}^d} \boldsymbol{\alpha}_i^\Delta \log(\boldsymbol{\alpha}_i^\Delta)$ and similarly for β. Note that in case α is a discrete measure with finite support, $H^\Delta(\alpha)$ is equal to (minus) the discrete entropy of α as long as Δ is smaller than the minimum separation between atoms of α. However, if α is not discrete then $H^\Delta(\alpha)$ blows up to $-\infty$ as Δ goes to 0 and we need to control how fast it does so. Considering α^Δ the block approximation of α with constant density $\boldsymbol{\alpha}_i^\Delta/\Delta^d$ on each block Q_i^Δ and (minus) its differential entropy $H_{\mathcal{L}^d}(\alpha^\Delta) = \int_{\mathbb{R}^d} \alpha^\Delta(x) \log \alpha^\Delta(x) dx$, it holds $H^\Delta(\alpha) = H_{\mathcal{L}^d}(\alpha^\Delta) - d\cdot\log(1/\Delta)$. Moreover, using the convexity of $H_{\mathcal{L}^d}$, this can be compared with the differential entropy of the uniform probability on a hypercube containing $\mathcal{X}$ of size $2D$. Thus it holds $H_{\mathcal{L}^d}(\alpha^\Delta) \geqslant -d\log(2D)$ and thus $H^\Delta(\alpha) \geqslant -d\cdot\log(2D/\Delta)$.

Summing up, we have for all $\Delta > 0$

$$W_\varepsilon(\alpha,\beta) - W(\alpha,\beta) \leqslant 2L\Delta\sqrt{d} + 2\varepsilon d\cdot\log(2D/\Delta).$$

The above bound is convex in Δ, minimized with $\Delta = 2\sqrt{d}\cdot\varepsilon/L$. This yields

$$W_\varepsilon(\alpha,\beta) - W(\alpha,\beta) \leqslant 4\varepsilon d + 2\varepsilon d\log\left(\frac{L\cdot D}{\sqrt{d}\cdot\varepsilon}\right). \qquad \square$$

4 Properties of Sinkhorn Potentials

We prove in this section that Sinkhorn potentials are bounded in the Sobolev space $\mathbf{H}^s(\mathbb{R}^d)$ regardless of the marginals α and β. For $s > \frac{d}{2}$, $\mathbf{H}^s(\mathbb{R}^d)$ is a reproducing kernel Hilbert space (RKHS): This property will be crucial to establish sample complexity results later on, using standard tools from RKHS theory.

Definition 11. *The Sobolev space* $\mathbf{H}^s(\mathcal{X})$*, for* $s \in \mathbb{N}^*$*, is the space of functions* $\varphi : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ *such that for every multi-index* k *with* $|k| \leqslant s$ *the mixed partial derivative* $\varphi^{(k)}$ *exists and belongs to* $L^2(\mathcal{X})$*. It is endowed with the following inner-product*

$$\langle \varphi, \psi \rangle_{\mathbf{H}^s(\mathcal{X})} = \sum_{|k| \leqslant s} \int_{\mathcal{X}} \varphi^{(k)}(x) \psi^{(k)}(x) \mathrm{d}x. \tag{4.1}$$

Theorem 13. *When* $\mathcal{X}$ *and* $\mathcal{Y}$ *are two bounded sets of* $\mathbb{R}^d$ *and the cost* c *is* $\mathcal{C}^\infty$*, then the Sinkhorn potentials* (u, v) *are uniformly bounded in the Sobolev space* $\mathbf{H}^s(\mathbb{R}^d)$ *and their norms satisfy*

$$\|u\|_{\mathbf{H}^s} = O\left(1 + \frac{1}{\varepsilon^{s-1}}\right) \text{ and } \|v\|_{\mathbf{H}^s} = O\left(1 + \frac{1}{\varepsilon^{s-1}}\right),$$

with constants that only depend on $|\mathcal{X}|$ *(or* $|\mathcal{Y}|$ *for* v*),*d*, and* $\left\|c^{(k)}\right\|_\infty$ *for* $k = 0, \dots, s$*. In particular, we get the following asymptotic behavior in* ε*:* $\|u\|_{\mathbf{H}^s} = O(1)$ *as* $\varepsilon \to +\infty$ *and* $\|u\|_{\mathbf{H}^s} = O(\frac{1}{\varepsilon^{s-1}})$ *as* $\varepsilon \to 0$*.*

To prove this theorem, we first need to state some regularity properties of the Sinkhorn potentials.

Proposition 17. *If* $\mathcal{X}$ *and* $\mathcal{Y}$ *are two bounded sets of* $\mathbb{R}^d$ *and the cost* c *is* $\mathcal{C}^\infty$*, then*

- $u(x) \in [\min_y v(y) - c(x, y), \max_y v(y) - c(x, y)]$ *for all* $x \in \mathcal{X}$
- u *is* L*-Lipschitz, where* L *is the Lipschitz constant of* c
- $u \in \mathcal{C}^\infty(\mathcal{X})$ *and* $\left\|u^{(k)}\right\|_\infty = O(1 + \frac{1}{\varepsilon^{k-1}})$

and the same results also stand for v *(inverting* u *and* v *in the first item, and replacing* $\mathcal{X}$ *by* $\mathcal{Y}$*).*

Proof. The proofs of all three claims exploit the optimality condition of the dual problem:

$$\exp\left(\frac{-u(x)}{\varepsilon}\right) = \int \exp\left(\frac{v(y) - c(x, y)}{\varepsilon}\right) \beta(y) \mathrm{d}y. \tag{4.2}$$

Since β is a probability measure, $e^{\frac{-u(x)}{\varepsilon}}$ is a convex combination of $\varphi : x \mapsto e^{\frac{v(x) - c(x,y)}{\varepsilon}}$ and thus $e^{\frac{-u(x)}{\varepsilon}} \in [\min_y \varphi(y), \max_y \varphi(y)]$. We get the desired bounds by taking the logarithm. The two other points use the following lemmas:

Lemma 3. *The derivatives of the potentials are given by the following recurrence*

$$u^{(n)}(x) = \int g_n(x, y) \gamma_\varepsilon(x, y) \beta(y) \mathrm{d}y, \tag{4.3}$$

where

$$g_{n+1}(x, y) = g_n'(x, y) + \frac{u'(x) - c'(x, y)}{\varepsilon} g_n(x, y),$$

$g_1(x,y) = c'(x,y)$ *and* $\gamma_\varepsilon(x,y) = \exp(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})$.

Lemma 4. *The sequence of auxiliary functions* $(g_k)_{k=0\dots}$ *verifies* $\left\|u^{(k)}\right\|_\infty \leqslant \|g_k\|_\infty$. *Besides, for all* $j = 0,\dots,k$, *for all* $k = 0,\dots,n-2$, $\left\|g_{n-k}^{(j)}\right\|_\infty$ *is bounded by a polynomial in* $\frac{1}{\varepsilon}$ *of order* $n-k+j-1$.

From the primal constraint, we have that $\int_{\mathcal{Y}} \gamma_\varepsilon(x,y)\beta(y)\mathrm{d}y = 1$. Thus thanks to Lemma 3 we immediately get that $\left\|u^{(n)}\right\|_\infty \leqslant \|g_n\|_\infty$. For $n = 1$, since $g_1 = c'$ we get that $\|u'\|_\infty = \|c'\|_\infty = L$ and this proves the second point of Proposition 17. The third point is a direct application of Lemma 4, and we prove both lemmas below. □

Proof. (Lemma 3) For better clarity, we carry out the computations in dimension 1 but all the arguments are valid in higher dimension and we will clarify delicate points throughout the proof. Differentiating under the integral is justified with the usual domination theorem, bounding the integrand thanks to the Lipschitz assumption on c, and this bound is integrable thanks to the marginal constraint. Differentiating both sides of the optimality condition (4.2) and rearranging yields

$$u'(x) = \int c'(x,y)\gamma_\varepsilon(x,y)\beta(y)\mathrm{d}y. \tag{4.4}$$

Notice that $\gamma'_\varepsilon(x,y) = \frac{u'(x)-c'(x,y)}{\varepsilon}\gamma_\varepsilon(x,y)$. Thus by immediate recurrence (differentiating both sides of the equality again) we get that

$$u^{(n)}(x) = \int g_n(x,y)\gamma_\varepsilon(x,y)\beta(y)\mathrm{d}y, \tag{4.5}$$

where $g_{n+1}(x,y) = g'_n(x,y) + \frac{u'(x)-c'(x,y)}{\varepsilon}g_n(x,y)$ and $g_1(x,y) = c'(x,y)$

To extend this first lemma to the d-dimensional case, we need to consider the sequence of indexes $\sigma = (\sigma_1,\sigma_2,\dots) \in \{1,\dots,d\}^{\mathbb{N}}$ which corresponds to the axis along which we successively differentiate. Using the same reasoning as above, it is straightforward to check that

$$\frac{\partial^k u}{\partial x_{\sigma_1}\dots\partial x_{\sigma_k}} = \int g_{\sigma,k}\gamma_\varepsilon\beta(y)\mathrm{d}y,$$

where $g_{\sigma,1} = \frac{\partial c}{\partial x_{\sigma_1}}$ and $g_{\sigma,k+1} = \frac{\partial g_{\sigma,k+1}}{\partial x_{\sigma_{k+1}}} + \frac{1}{\varepsilon}\left(\frac{\partial u}{\partial x_{\sigma_{k+1}}} - \frac{\partial c}{\partial x_{\sigma_{k+1}}}\right) g_{\sigma,k+1}$.

□

Proof. (Lemma 4) The proof is made by recurrence on the following property :
P_n *: For all* $j = 0,\dots,k$, *for all* $k = 0,\dots,n-2$, $\left\|g_{n-k}^{(j)}\right\|_\infty$ *is bounded by a polynomial in* $\frac{1}{\varepsilon}$ *of order* $n-k+j-1$.

Let us initialize the recurrence with $n = 2$

$$g_2 = g_1' + \frac{u' - c'}{\varepsilon} g_1 \tag{4.6}$$

$$\|g_2\|_\infty \leqslant \|g_1'\|_\infty + \frac{\|u'\|_\infty + \|c'\|_\infty}{\varepsilon} \|g_1\|_\infty . \tag{4.7}$$

Recall that $\|u'\|_\infty = \|g_1\|_\infty = \|c'\|_\infty$. Let $C = \max_k \left\|c^{(k)}\right\|_\infty$, we get that $\|g_2\|_\infty \leqslant C + \frac{C+C}{\varepsilon} C$ which is of the required form.

Now assume that P_n is true for some $n \geqslant 2$. This means we have bounds on $g_{n-k}^{(i)}$, for $k = 0, \ldots, n-2$ and $i = 0, \ldots, k$. To prove the property at rank $n+1$ we want bounds on $g_{n+1-k}^{(i)}$, for $k = 0, \ldots, n-1$ and $i = 0, \ldots, k$. The only new quantity that we need to bound are $g_{n+1-k}^{(k)}, k = 0, \ldots, n-1$. Let us start by bounding $g_2^{(n-1)}$ which corresponds to $k = n-1$ and we will do a backward recurrence on k. By applying Leibniz formula for the successive derivatives of a product of functions, we get

$$g_2 = g_1' + \frac{u' - c'}{\varepsilon} g_1, \tag{4.8}$$

$$g_2^{(n-1)} = g_1^{(n)} + \sum_{p=0}^{n-1} \binom{n-1}{p} \frac{u^{(p+1)} - c^{(p+1)}}{\varepsilon} g_1^{(n-1-p)}, \tag{4.9}$$

$$\left\|g_2^{(n-1)}\right\|_\infty \leqslant \left\|g_1^{(n)}\right\|_\infty + \sum_{p=0}^{n-1} \binom{n-1}{p} \frac{\left\|u^{(p+1)}\right\|_\infty + \left\|c^{(p+1)}\right\|_\infty}{\varepsilon} \left\|g_1^{(n-1-p)}\right\|_\infty \tag{4.10}$$

$$\leqslant C + \sum_{p=0}^{n-1} \binom{n-1}{p} \frac{\|g_{p+1}\|_\infty + C}{\varepsilon} C. \tag{4.11}$$

Thanks to P_n we have that $\|g_p\|_\infty \leqslant \sum_{i=0}^{p} a_{i,p} \frac{1}{\varepsilon^i}, p = 1, \ldots, n$ so the highest order term in ε in the above inequality is $\frac{1}{\varepsilon^n}$. Thus we get $\left\|g_2^{(n-1)}\right\|_\infty \leqslant \sum_{i=0}^{n+1} a_{i,2,n-1} \frac{1}{\varepsilon^i}$ which is of the expected order

Now assume $g_{n+1-j}^{(j)}$ are bounded with the appropriate polynomials for $j < k \leqslant n-1$. Let us bound $g_{n+1-k}^{(k)}$

$$\left\|g_{n+1-k}^{(k)}\right\|_\infty \leqslant \left\|g_{n-k}^{(k+1)}\right\|_\infty + \sum_{p=0}^{k} \binom{k}{p} \frac{\left\|u^{(p+1)}\right\|_\infty + \left\|c^{(p+1)}\right\|_\infty}{\varepsilon} \left\|g_{n-k}^{(k-p)}\right\|_\infty \tag{4.12}$$

$$\leqslant \left\|g_{n-k}^{(k+1)}\right\|_\infty + \sum_{p=0}^{k} \binom{k}{p} \frac{\|g_{p+1}\|_\infty + C}{\varepsilon} \left\|g_{n-k}^{(k-p)}.\right\|_\infty \tag{4.13}$$

The first term $\left\|g_{n-k}^{(k+1)}\right\|_\infty$ is bounded with a polynomial of order $\frac{1}{\varepsilon^{n+1}}$ by recurrence assumption. Regarding the terms in the sum, they also have all been bounded and

$$\|g_{p+1}\|_\infty \left\|g_{n-k}^{(k-p)}\right\|_\infty \leqslant \left(\sum_{i=0}^{p} a_{i,p+1} \frac{1}{\varepsilon^i}\right) \left(\sum_{i=0}^{n-p} a_{i,n-k,k-p} \frac{1}{\varepsilon^i}\right) \leqslant \sum_{i=0}^{n} \tilde{a}_i \frac{1}{\varepsilon^i},$$

so $\left\|g^{(k)}_{n+1-k}\right\|_\infty \leqslant \sum_{i=0}^{n+1} a_{i,n+1-k,k} \frac{1}{\varepsilon^i}$. To extend the result in $\mathbb{R}^d$, the recurrence is made on the the following property

$$\left\|g^{(j)}_{\sigma,n-k}\right\|_\infty \leqslant \sum_{i=0}^{n-k+|j|-1} a_{i,n-k,j,\sigma} \frac{1}{\varepsilon^i}, \tag{4.14}$$

$\forall j \mid |j| = 0, \dots, k \quad \forall k = 0, \dots, n-2 \quad \forall \sigma \in \{1, \dots, d\}^{\mathbb{N}}$, where j is a multi-index since we are dealing with multi-variate functions, and $g_{\sigma,n-k}$ is defined at the end of the previous proof. The computations can be carried out in the same way as above, using the multivariate version of Leibniz formula in (4.9) since we are now dealing with multi-indexes. □

Combining the bounds of the derivatives of the potentials with the definition of the norm in $\mathbf{H}^s$, is enough to complete the proof of Theorem 13.

Proof. (Theorem 13) The norm of u in $\mathbf{H}^s(\mathcal{X})$ is

$$\|u\|_{\mathbf{H}^s} = \left(\sum_{|k|\leqslant s} \int_{\mathcal{X}} (u^{(k)})^2\right)^{\frac{1}{2}} \leqslant |\mathcal{X}| \left(\sum_{|k|\leqslant s} \left\|u^{(k)}\right\|_\infty^2\right)^{\frac{1}{2}}.$$

From Proposition 17 we have that $\forall k, \left\|u^{(k)}\right\|_\infty = O(1 + \frac{1}{\varepsilon^{k-1}})$ and thus we get that $\|u\|_{\mathbf{H}^s} = O(1 + \frac{1}{\varepsilon^{s-1}})$. We just proved the bound in $\mathbf{H}^s(\mathcal{X})$ but we actually want to have a bound on $\mathbf{H}^s(\mathbb{R}^d)$. This is immediate thanks to the Sobolev extension theorem (Calderón, 1961) which guarantees that $\|u\|_{\mathbf{H}^s(\mathbb{R}^d)} \leqslant C\|u\|_{\mathbf{H}^s(\mathcal{X})}$ under the assumption that $\mathcal{X}$ is a bounded Lipschitz domain. □

This result, aside from proving useful in the next section to obtain sample complexity results on the Sinkhorn Divergence, also proves that kernel-SGD can be used to solve continuous regularized OT. This idea, which we develop in Chapter 4, Sec. 5.1, consists in assuming the potentials are in the ball of a certain RKHS, to write them as a linear combination of kernel functions and then perform stochastic gradient descent on these coefficients. Knowing the potentials are in a ball of a RKHS is enough to guarantee convergence of kernel-SGD (see Theorem 23).

5 Approximating Sinkhorn Divergence from Samples

In practice, measures α and β are only known through a finite number of samples. Thus, what can be actually computed is the Sinkhorn Divergence between the empirical measures $\hat{\alpha}_n \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\beta}_n \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, where $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ are i.i.d. random variables distributed according to α and β respectively. This yields

the empirical Sinkhorn Divergence:

$$W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n) = \max_{u,v} \sum_{i=1}^n u(X_i) + \sum_{i=1}^n v(Y_i) - \varepsilon \sum_{i=1}^n \exp\left(\frac{u(X_i) + v(Y_i) - c(X_i, Y_i)}{\varepsilon}\right) + \varepsilon$$
$$= \max_{u,v} \frac{1}{n} \sum_{i=1}^n f_\varepsilon^{X_i Y_i}(u, v) + \varepsilon,$$

where $(X_i, Y_i)_{i=1}^n$ are i.i.d random variables distributed according to $\alpha \otimes \beta$. On actual samples, these quantities can be computed using Sinkhorn's algorithm (Cuturi, 2013).

Our goal is to quantify the error that is made by approximating α, β by their empirical counterparts $\hat{\alpha}_n, \hat{\beta}_n$, that is bounding the following quantity:

$$|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| = |\mathbb{E} f_\varepsilon^{XY}(u^*, v^*) - \frac{1}{n} \sum_{i=1}^n f_\varepsilon^{X_i Y_i}(\hat{u}, \hat{v})|, \tag{5.1}$$

where (u^*, v^*) are the optimal Sinkhorn potentials associated with (α, β) and $(\hat{u}, \hat{v})$ are their empirical counterparts.

Theorem 14. *Consider the Sinkhorn Divergence between two measures α and β on $\mathcal{X}$ and $\mathcal{Y}$ two bounded subsets of $\mathbb{R}^d$, with a $\mathcal{C}^\infty$, L-Lipschitz cost c. One has*

$$\mathbb{E}|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| = O\left(\frac{e^{\frac{\kappa}{\varepsilon}}}{\sqrt{n}}\left(1 + \frac{1}{\varepsilon^{\lfloor d/2 \rfloor}}\right)\right),$$

where $\kappa = 2L|\mathcal{X}| + \|c\|_\infty$ and constants only depend on $|\mathcal{X}|$,$|\mathcal{Y}|$,d, and $\left\|c^{(k)}\right\|_\infty$ for $k = 0 \dots \lfloor d/2 \rfloor$. In particular, we get the following asymptotic behavior in ε:

$$\mathbb{E}|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| = O\left(\frac{e^{\frac{\kappa}{\varepsilon}}}{\varepsilon^{\lfloor d/2 \rfloor}\sqrt{n}}\right) \qquad \text{as } \varepsilon \to 0,$$
$$\mathbb{E}|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| = O\left(\frac{1}{\sqrt{n}}\right) \qquad \text{as } \varepsilon \to +\infty.$$

An interesting feature from this theorem is the fact when ε is large enough, the convergence rate does not depend on ε anymore. This means that at some point, increasing ε will not substantially improve convergence. However, for small values of ε the dependence is critical.

We prove this result in the rest of this section. The main idea is to exploit standard results from PAC-learning in RKHS. Our theorem is an application of the following result from Bartlett and Mendelson (2002) (combining Theorem 12,4) and Lemma 22 in their paper):

Proposition 18. *(Bartlett-Mendelson '02) Consider α a probability distribution, ℓ a*

B-Lipschitz loss and $\mathcal{G}$ a given class of functions. Then

$$\mathbb{E}_\alpha\left[\sup_{g\in\mathcal{G}}\mathbb{E}_\alpha\ell(g,X)-\frac{1}{n}\sum_{i=1}^n\ell(g,X_i)\right]\leqslant 2B\mathbb{E}_\alpha\mathcal{R}(\mathcal{G}(X_1^n)),$$

where $\mathcal{R}(\mathcal{G}(X_1^n))$ is the Rademacher complexity of class $\mathcal{G}$ defined by $\mathcal{R}(\mathcal{G}(X_1^n)) = \sup_{g\in\mathcal{G}}\mathbb{E}_\sigma\frac{1}{n}\sum_{i=1}^n\sigma_i g(X_i)$ where $(\sigma_i)_i$ are iid Rademacher random variables. Besides, when $\mathcal{G}$ is a ball of radius λ in a RKHS with kernel k the Rademacher complexity is bounded by

$$\mathcal{R}(\mathcal{G}_\lambda(X_1^n))\leqslant\frac{\lambda}{n}\sqrt{\sum_{i=1}^n k(X_i,X_i)}.$$

Our problem falls in this framework thanks to the following lemma:

Lemma 5. *Let $\mathcal{H}_\lambda^s \overset{\text{def.}}{=} \{u\in\mathbf{H}^s(\mathbb{R}^d)\mid\|u\|_{\mathbf{H}^s(\mathbb{R}^d)}\leqslant\lambda\}$, then there exists λ such that:*

$$|W_\varepsilon(\alpha,\beta)-W_\varepsilon(\hat\alpha_n,\hat\beta_n)|\leqslant 3\sup_{(u,v)\in(\mathcal{H}_\lambda^s)^2}|\mathbb{E}f_\varepsilon^{XY}(u,v)-\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u,v)|.$$

Proof. Inserting $\mathbb{E}f_\varepsilon^{XY}(\hat u,\hat v)$ and using the triangle inequality in (5.1) gives

$$\begin{aligned}|W_\varepsilon(\alpha,\beta)-W_\varepsilon(\hat\alpha_n,\hat\beta_n)|\leqslant&|\mathbb{E}f_\varepsilon^{XY}(u^*,v^*)-\mathbb{E}f_\varepsilon^{XY}(\hat u,\hat v)|\\&+|\mathbb{E}f_\varepsilon^{XY}(\hat u,\hat v)-\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat u,\hat v)|.\end{aligned}$$

From Theorem 13, we know that the all the dual potentials are bounded in $\mathbf{H}^s(\mathbb{R}^d)$ by a constant λ which doesn't depend on the measures. Thus the second term is bounded by $\sup_{(u,v)\in(\mathcal{H}_\lambda^s)^2}|\mathbb{E}f_\varepsilon(u,v)-\frac{1}{n}\sum_{i=1}^n f_\varepsilon(u,v)|$.
The first quantity needs to be broken down further. Notice that it is non-negative since (u^*,v^*) is the maximizer of $\mathbb{E}f_\varepsilon(\cdot,\cdot)$ so we can leave out the absolute value. We have:

$$\begin{aligned}\mathbb{E}f_\varepsilon^{XY}(u^*,v^*)-\mathbb{E}f_\varepsilon^{XY}(\hat u,\hat v)\leqslant&\mathbb{E}f_\varepsilon^{XY}(u^*,v^*)-\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u^*,v^*)&(5.2)\\&+\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u^*,v^*)-\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat u,\hat v)&(5.3)\\&+\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat u,\hat v)-\mathbb{E}f_\varepsilon^{XY}(\hat u,\hat v).&(5.4)\end{aligned}$$

Both (5.2) and (5.4) can be bounded by $\sup_{(u,v)\in(\mathcal{H}_\lambda^s)^2}|\mathbb{E}f_\varepsilon^{XY}(u,v)-\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u,v)|$ while (5.3) is non-positive since $(\hat u,\hat v)$ is the maximizer of $\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\cdot,\cdot)$. □

To apply Proposition 18 to Sinkhorn Divergences we need to prove that *(a)* the optimal potentials are in a RKHS and *(b)* our loss function f^ε is Lipschitz in the potentials.

The first point has already been proved in the previous section. The RKHS we are considering is $\mathbf{H}^s(\mathbb{R}^d)$ with $s = \lfloor \frac{d}{2} \rfloor + 1$. It remains to prove that f^ε is Lipschitz in (u, v) on a certain subspace that contains the optimal potentials.

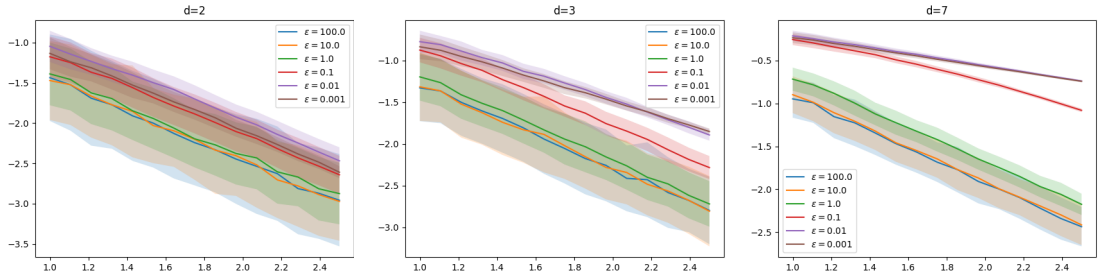


Figure 3.1 – $\bar{W}_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n)$ as a function of n in log-log space : Influence of ε for fixed d on two uniform distributions on the hypercube with quadratic cost.

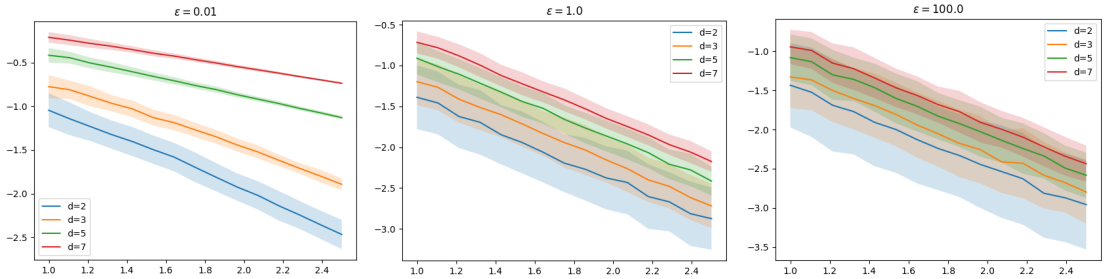


Figure 3.2 – $\bar{W}_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n)$ as a function of n in log-log space : Influence of d for fixed ε on two uniform distributions on the hypercube with quadratic cost.

Lemma 6. *Let* $\mathcal{A} = \{(u, v) \mid u \oplus v \leqslant 2L|\mathcal{X}| + \|c\|_\infty\}$. *We have:*

(i) the pairs of optimal potentials (u^*, v^*) *such that* $u^*(0) = 0$ *belong to* $\mathcal{A}$,

(ii) f^ε *is B-Lipschitz in* (u, v) *on* $\mathcal{A}$ *with* $B \leqslant 1 + \exp(2\frac{L|\mathcal{X}|+\|c\|_\infty}{\varepsilon})$.

Proof. Let us prove that we can restrict ourselves to a subspace on which f^ε is Lipschitz in (u, v),

$$f^\varepsilon(u, v, x, y) = u(x) + v(y) - \varepsilon \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right),$$

$$\nabla f^\varepsilon(u, v) = 1 - \exp\left(\frac{u + v - c}{\varepsilon}\right).$$

To ensure that f^ε is Lipschitz, we simply need to ensure that the quantity inside the exponential is upper bounded at optimality and then restrict the function to all (u, v) that satisfy that bound.

Recall the bounds on the optimal potentials from Proposition 17. We have that $\forall x \in \mathcal{X}, y \in \mathcal{Y}$,

$$u(x) \leqslant L|x| \quad \text{and} \quad v(y) \leqslant \max_x u(x) - c(x, y).$$

Since we assumed $\mathcal{X}$ to be a bounded set, denoting by $|\mathcal{X}|$ the diameter of the space we get that at optimality $\forall x \in \mathcal{X}, y \in \mathcal{Y}$

$$u(x) + v(y) \leqslant 2L|\mathcal{X}| + \|c\|_\infty .$$

Let us denote $\mathcal{A} = \{(u, v) \in (\mathbf{H}^s(\mathbb{R}^d))^2 \mid u \oplus v \leqslant 2L|\mathcal{X}| + \|c\|_\infty\}$, we have that $\forall (u, v) \in \mathcal{A}$,

$$|\nabla f^\varepsilon(u, v)| \leqslant 1 + \exp(2\frac{L|\mathcal{X}| + \|c\|_\infty}{\varepsilon}). \qquad \square$$

We now have all the required elements to prove our sample complexity result on the Sinkhorn loss, by applying Proposition 18.

Proof. (Theorem 14) Since f_ε is Lipschitz and we are optimizing over $\mathbf{H}^s(\mathbb{R}^d)$ which is a RKHS, we can apply Proposition 18 to bound the sup in Lemma 5. We get:

$$\mathbb{E}|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| \leqslant 3\frac{2B\lambda}{n}\mathbb{E}\sqrt{\sum_{i=1}^n k(X_i, X_i)},$$

where $B \leqslant 1 + \exp(2\frac{L|\mathcal{X}|+\|c\|_\infty}{\varepsilon})$ (Lemma 6), $\lambda = O(\max(1, \frac{1}{\varepsilon^{d/2}}))$ (Theorem 13). We can further bound $\sqrt{\sum_{i=1}^n k(X_i, X_i)}$ by $\sqrt{n \max_{x\in\mathcal{X}} k(x, x)}$ where k is the kernel associated to $H^s(\mathbb{R}^d)$ (usually called Matern or Sobolev kernel) and thus $\max_{x\in\mathcal{X}} k(x, x) = k(0, 0) \overset{\text{def.}}{=} K$ which does not depend on n or ε. Combining all these bounds, we get the convergence rate in $\frac{1}{\sqrt{n}}$ with different asymptotic behaviors in ε when it is large or small. $\square$

Using similar arguments, we can also derive a concentration result:

Corollary 2. *With probability at least* $1 - \delta$,

$$|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| \leqslant 6B\frac{\lambda K}{\sqrt{n}} + C\sqrt{\frac{2\log\frac{1}{\delta}}{n}},$$

where B, λ, K *are defined in the proof above, and* $C = \kappa + \varepsilon\exp(\frac{\kappa}{\varepsilon})$ *with* $\kappa = 2L|\mathcal{X}| + \|c\|_\infty$.

Proof. We apply the bounded differences inequality (McDiarmid, 1989) to

$$g : (x_1, \ldots, x_n) \mapsto \sup_{u,v\in\mathcal{H}^s_\lambda} (\mathbb{E}f^{XY}_\varepsilon - \frac{1}{n}f^{X_i,Y_i}_\varepsilon).$$

From Lemma 6 we get that $\forall x, y$, $f^{xy}_\varepsilon(u, v) \leqslant \kappa + \varepsilon e^{\kappa/\varepsilon} \overset{\text{def.}}{=} C$, and thus, changing one of the variables in g changes the value of the function by at most $2C/n$. Thus the bounded

differences inequality gives

$$\mathbb{P}\left(|g(X_1,\dots,X_n) - \mathbb{E}g(X_1,\dots,X_n)| > t\right) \leqslant 2\exp\left(\frac{t^2 n}{2C^2}\right).$$

Choosing $t = C\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$ yields that with probability at least $1-\delta$

$$g(X_1,\dots,X_n) \leqslant \mathbb{E}g(X_1,\dots,X_n) + C\sqrt{\frac{2\log\frac{1}{\delta}}{n}},$$

and from Theorem 14 we already have

$$\mathbb{E}g(X_1,\dots,X_n) = \mathbb{E}\sup_{u,v\in\mathcal{H}^s_\lambda}(\mathbb{E}f_\varepsilon^{XY} - \frac{1}{n}f_\varepsilon^{X_i,Y_i}) \leqslant \frac{2B\lambda K}{\sqrt{n}}.$$

□

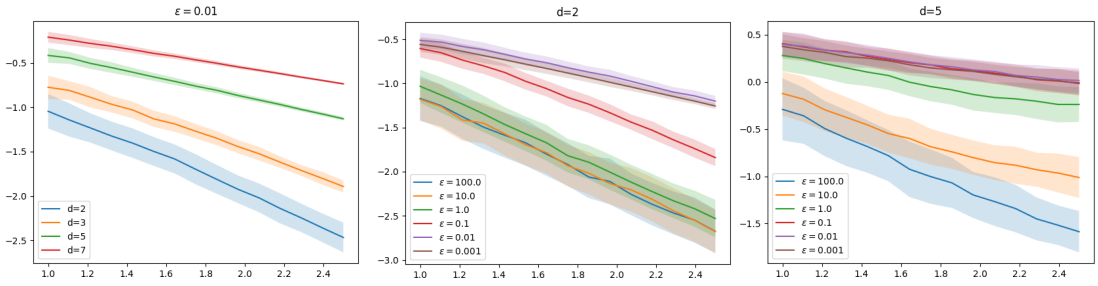


Figure 3.3 – $\bar{W}_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n)$ as a function of n in log-log space - cost $c(x,y) = \|x-y\|_1$ with uniform distributions (two leftmost figures) and quadratic cost $c(x,y) = \|x-y\|_2^2$ with standard normal distributions (right figure).

6 Experiments

We conclude with some numerical experiments on the sample complexity of Sinkhorn Divergences. As there are no explicit formulas for W_ε in general, we consider $\bar{W}_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n)$ where $\hat{\alpha}_n \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{Xi}$, $\hat{\alpha}'_n \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{Xi'}$ and $(X_1,\dots,X_n)$ and $(X'_1,\dots,X'_n)$ are two independent n-samples from α. Note that we use in this section the normalized Sinkhorn Divergence as defined in (2.2), since we know that $\bar{W}_\varepsilon(\alpha,\alpha) = 0$ and thus $\bar{W}_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n) \to 0$ as $n \to +\infty$.

Each of the experiments is run 300 times, and we plot the average of $\bar{W}_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n)$ as a function of n in log-log space, with shaded standard deviation bars.

First, we consider the uniform distribution over a hypercube with the standard quadratic cost $c(x,y) = \|x-y\|_2^2$, which falls within our framework, as we are dealing with a $\mathcal{C}^\infty$ cost on a bounded domain. Figure 3.1 shows the influence of the dimension

d on the convergence, while Figure 3.2 shows the influence of the regularization ε on the convergence for a given dimension. The influence of ε on the convergence rate increases with the dimension: the curves are almost parallel for all values of ε in dimension 2 but they get further apart as dimension increases. As expected from our bound, there is a cutoff which happens here at $\varepsilon = 1$. All values of $\varepsilon \geqslant 1$ have similar convergence rates, and the dependence on $\frac{1}{\varepsilon}$ becomes clear for smaller values. The same cutoff appears when looking at the influence of the dimension on the convergence rate for a fixed ε. The curves are parallel for all dimensions for $\varepsilon \geqslant 1$ but they have very different slopes for smaller ε.

We relax next some of the assumptions needed in our theorem to see how the Sinkhorn Divergence behaves empirically. First we relax the regularity assumption on the cost, using $c(x, y) = \|x - y\|_1$. As seen on the two left images in figure 3.3 the behavior is very similar to the quadratic cost but with a more pronounced influence of ε, even for small dimensions. The fact that the convergence rate gets slower as ε gets smaller is already very clear in dimension 2, which wasn't the case for the quadratic cost. The influence of the dimension for a given value of ε is not any different however.

We also relax the bounded domain assumption, considering a standard normal distribution over $\mathbb{R}^d$ with a quadratic cost. While the influence of ε on the convergence rate is still obvious, the influence of the dimension is less clear. There is also a higher variance, which can be expected as the concentration bound from Corollary 2 depends on the diameter of the domain.

For all curves, we observe that d and ε impact variance, with much smaller variance for small values of ε and high dimensions. From the concentration bound, the dependency on ε coming from the uniform bound on f_ε is of the form $\varepsilon \exp(\kappa/\varepsilon)$, suggesting higher variance for small values of ε. This could indicate that our uniform bound on f_ε is not tight, and we should consider other methods to get tighter bounds in further work.

Chapter 4

Stochastic Optimization for Large-Scale Optimal Transport

Entropy-regularized Optimal Transport (OT) has alleviated the computational burden of OT, for many applications. However, its state of the art solver, Sinkhorn's algorithm, only copes with discrete measures and its iteration complexity scales as $O(n^2)$ (where n is the number of points in the discrete measures). We thus propose a new class of online stochastic optimization algorithms to cope with large-scale OT problems. They can handle arbitrary distributions (discrete or continuous) as long as one is able to draw samples from them. This alleviates the need to discretize these densities, while giving access to provably convergent methods without discretization error. These algorithms rely on one key idea which is that the dual OT problem can be re-cast as the maximization of an expectation.

We exploit this formulation in three different setups: (i) *when comparing a discrete distribution to another, we show that incremental stochastic optimization schemes can beat Sinkhorn's algorithm, the current state-of-the-art finite dimensional OT solver;* (ii) *when comparing a discrete distribution to a continuous density, a semi-discrete reformulation of the dual program is amenable to averaged stochastic gradient descent (SGD), leading to better performance than approximately solving the problem by discretization ;* (iii) *when dealing with two continuous densities, we propose a stochastic gradient descent over a reproducing kernel Hilbert space (RKHS) and introduce an approximate feature approach (via incomplete Cholesky decomposition or Random Fourier Features) to significantly alleviate computational time. This is currently the only known method to solve this problem, apart from computing OT on finite samples. We backup these claims on a set of discrete, semi-discrete and continuous benchmark problems.*

Most of the content in this chapter comes from (Genevay et al., 2016), but the section on continuous transport is revised in the light of recent results from Genevay et al. (2019 (to appear). Knowing that dual potentials are uniformly bounded in a RKHS, we can state a stronger version of the convergence theorem for kernel-SGD. The numerical experiments are also improved, in particular thanks to the addition of the approximate features approach.

1 Introduction

Throughout this thesis, we have already advocated that OT is the natural choice to solve a large variety of problems which require comparing probability measures or data in the form of histograms, in particular because it takes into account the underlying geometry of the problem. However, this comes at the price of an enormous computational overhead, compared to geometrically-oblivious distances such as the Euclidean or χ^2 distances or the Kullback-Leibler divergence. This is especially true because current OT solvers require to sample beforehand the distributions on a pre-defined set of points, or on a grid. This is both inefficient (in term of storage and speed) and counter-intuitive. Indeed, most high-dimensional computational scenarios naturally represent distributions as objects from which one can *sample*, not as density functions to be discretized. Our goal is to alleviate these shortcomings. We propose a class of provably convergent stochastic optimization schemes that can handle both discrete and continuous distributions through sampling.

Previous works. The prevalent way to compute OT distances is by solving the so-called Kantorovitch problem (Kantorovich, 1942) (see Chapter 1, Sec. 2.3 for a short primer on the basics of OT formulations), which boils down to a large-scale linear program when dealing with discrete distributions (i.e., finite weighted sums of Dirac masses). This linear program can be solved using network flow solvers, which can be further refined to assignment problems when comparing measures of the same size with uniform weights (Burkard et al., 2009). Recently, regularized approaches that solve the OT with an entropic penalization (Cuturi, 2013) have been shown to be extremely efficient to approximate OT solutions at a very low computational cost. These regularized approaches have supported recent applications of OT to computer graphics (Solomon et al., 2015) and machine learning (Frogner et al., 2015). These methods apply the celebrated Sinkhorn algorithm (Sinkhorn, 1964), and can be extended to solve a variety of transportation-related problems such as the computation of barycenters for the optimal transport metric or multimarginal optimal transport (Benamou et al., 2015). Their chief computational advantage over competing solvers is that each iteration boils down to matrix-vector multiplications, which can be easily parallelized, streams extremely well on GPU, and enjoys linear-time implementation on regular grids or triangulated domains (Solomon et al., 2015).

These methods are however purely discrete and cannot cope with continuous densities. The only known class of methods that can overcome this limitation are so-called semi-discrete solvers (Aurenhammer et al., 1998), that can be implemented efficiently using computational geometry primitives (Mérigot, 2011). They can compute distance between a discrete distribution and a continuous density. Nonetheless, they are restricted to the Euclidean squared cost, and can only be implemented in low dimensions

(2-D and 3-D). Solving these semi-discrete problems efficiently could have a significant impact for applications to density fitting with an OT loss (Bassetti et al., 2006) for machine learning applications, see (Montavon et al., 2016). Lastly, let us point out that there is currently no method that can compute OT distances between two continuous densities, which is thus an open problem we tackle in this chapter.

Contributions. This chapter introduces stochastic optimization methods to compute large-scale optimal transport in all three possible settings: *discrete* OT, to compare a discrete *vs.* another discrete measure; *semi-discrete* OT, to compare a discrete *vs.* a continuous measure; and *continuous* OT, to compare a continuous *vs.* another continuous measure. These methods can be used to solve classical OT problems, but they enjoy faster convergence properties when considering their entropic-regularized versions. We show that the discrete regularized OT problem can be tackled using incremental algorithms, and we consider in particular the stochastic averaged gradient (SAG) method (Schmidt et al., 2016). Each iteration of that algorithm requires n operations (n being the size of the supports of the input distributions), which makes it scale better in large-scale problems than the state-of-the-art Sinkhorn algorithm, while still enjoying a convergence rate of $O(1/k)$, k being the number of iterations. We show that the semi-discrete OT problem can be solved using averaged stochastic gradient descent (SGD), whose convergence rate is $O(1/\sqrt{k})$. This approach is numerically advantageous over the brute force approach consisting in sampling first the continuous density to solve next a discrete OT problem. Following the publication of this work, this online semi-discrete algorithm has been successfully applied to texture synthesis in image processing (Galerne et al., 2018), and to the computation of Wasserstein Barycenters (Staib et al., 2017). Lastly, for continuous optimal transport, we propose a novel method which makes use of an expansion of the dual variables in a reproducing kernel Hilbert space (RKHS). This allows us for the first time to compute with a converging algorithm OT distances between two arbitrary densities, thanks to the fact that the dual potentials are known to be in a RKHS ball(see Chapter 3 for details). We also provide an approach using approximate features (via incomplete Cholesky decomposition (Wu et al., 2006) or Random Fourier features (Rahimi and Recht, 2007)) to significantly alleviate computational time (going from quadratic to linear in the number of iterations). More recently, (Seguy et al., 2017) exploited our dual formulation as an expectation with a neural network parametrization of the dual variables rather than a RKHS expansion. This gives interesting results on the tasks they consider, although they do not derive convergence rates for their approach.

2 Optimal Transport: Primal, Dual and Semi-dual Formulations

We consider the optimal transport problem between two measures $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, defined on metric spaces $\mathcal{X}$ and $\mathcal{Y}$. No particular assumption is made on the form of α and β, we only assume that they both can be sampled from to be able to apply our algorithms.

2.1 Primal, Dual and Semi-dual Formulations.

In this section, we recall the different formulations of entropy-regularized optimal transport which will be exploited in the remainder of the chapter. We refer the reader to Chapter 1 for more details on their derivation and specific properties.

As previously, we consider the Kantorovich formulation (Kantorovich, 1942) of OT with entropic regularization (Cuturi, 2013) between two probability measures $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$:

$$W_\varepsilon(\alpha,\beta) \stackrel{\text{def.}}{=} \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\pi(x,y), \quad (\mathcal{P}_\varepsilon)$$

where the constraint set $\Pi(\alpha,\beta)$ is the set of couplings on $\mathcal{X}\times\mathcal{Y}$ with marginals α and β.

When $\varepsilon > 0$, problem $(\mathcal{P}_\varepsilon)$ is strictly convex, so that the optimal π is unique, and algebraic properties of the entropy H result in computations that can be tackled using Sinkhorn's algorithm which is extensively described in Chapter 1, Sec. 4.2.

Recall from Chapter 1, Proposition 7, that entropy-regularized OT between two probability measures α and β has an equivalent dual formulation:

$$\begin{aligned} W_\varepsilon(\alpha,\beta) = &\max_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\beta(y) \\ &- \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon, \end{aligned} \quad (\mathcal{D}_\varepsilon)$$

and the primal-dual relationship is given by

$$\mathrm{d}\pi(x,y) = \exp(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})\mathrm{d}\alpha(x)\mathrm{d}\beta(y).$$

A nice feature of entropy-regularized OT, which we already highlighted, is the fact that it yields an unconstrained dual problem, contrarily to standard OT. The third term in the dual $\varepsilon\int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$ is a smooth approximation of the indicator of the constraint set U_c set that appears in the dual of standard OT:

$$U_c \stackrel{\text{def.}}{=} \{(u,v)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})\ ;\ \forall(x,y)\in\mathcal{X}\times\mathcal{Y}, u(x)+v(y)\leqslant c(x,y)\}\,.$$

For any $v \in \mathcal{C}(\mathcal{Y})$, recall the definition of c-transform and its "smoothed" approximation, already introduced in Chapter 1, Sec. 4.3:

$$\forall x \in \mathcal{X}, \quad v^{c,\varepsilon}(x) \stackrel{\text{def.}}{=} \begin{cases} \min_{y \in \mathcal{Y}} c(x,y) - v(y) & \text{if} \quad \varepsilon = 0, \\ -\varepsilon \log\left(\int_{\mathcal{Y}} \exp(\frac{v(y)-c(x,y)}{\varepsilon}) \mathrm{d}\beta(y) \right) & \text{if} \quad \varepsilon > 0. \end{cases} \tag{2.1}$$

This allows us to introduce another equivalent formulation for entropy-regularized OT, as done in Proposition 12 in Chapter 1, which we call the *semi-dual*:

$$W_\varepsilon^c(\alpha,\beta) = \max_{v \in \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} v^{c,\varepsilon}(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y) \mathrm{d}\beta(y). \tag{$\mathcal{S}_\varepsilon$}$$

The other dual potential u solving ($\mathcal{D}_\varepsilon$) is recovered from an optimal v solving ($\mathcal{S}_\varepsilon$) as $u = v^{c,\varepsilon}$.

We refer to ($\mathcal{D}_\varepsilon$) as the "semi-dual" problem, because in the special case $\varepsilon = 0$, ($\mathcal{S}_\varepsilon$) boils down to the so-called semi-discrete OT problem (Aurenhammer et al., 1998). Both dual problems are concave maximization problems. The optimal dual variables (u, v) – known as Kantorovitch potentials – are not unique, since for any solution (u, v) of ($\mathcal{D}_\varepsilon$), $(u + \lambda, v - \lambda)$ is also a solution for any $\lambda \in \mathbb{R}$. When $\varepsilon > 0$, they can be shown to be unique up to this scalar translation. The proof is given in Section 4.1 of Chapter 1. We also refer to Chapter 1, Sec. 4.4 for a discussion (and proofs) of the convergence of the solutions of ($\mathcal{P}_\varepsilon$), ($\mathcal{D}_\varepsilon$) and ($\mathcal{S}_\varepsilon$) towards those of ($\mathcal{P}_0$), ($\mathcal{D}_0$) and ($\mathcal{S}_0$) as $\varepsilon \to 0$.

A key advantage of ($\mathcal{S}_\varepsilon$) over ($\mathcal{D}_\varepsilon$) is that, when β is a discrete density (but not necessarily α), then ($\mathcal{S}_\varepsilon$) is a finite-dimensional concave maximization problem, which can thus be solved using stochastic programming techniques, as highlighted in Section 4. By contrast, when both α and β are continuous densities, these dual problems are intrinsically infinite dimensional, and we propose in Section 5.1 more advanced techniques based on RKHSs.

2.2 Stochastic Optimization Formulations

The fundamental property needed to apply stochastic programming is that both dual problems ($\mathcal{D}_\varepsilon$) and ($\mathcal{S}_\varepsilon$) can be rephrased as maximizing expectations:

Proposition 19. *The dual of entropy-regularized OT between two probability measures α and β can be rewritten as the maximization of an expectation over $\alpha \otimes \beta$:*

$$W_\varepsilon^c(\alpha,\beta) = \max_{u,v \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})} \mathbb{E}_{\alpha\otimes\beta}[f_\varepsilon^{XY}(u,v)] + \varepsilon,$$

where

$$f_\varepsilon^{xy} \stackrel{\text{def.}}{=} u(x) + v(y) - \varepsilon \exp \frac{u(x)+v(y)-c(x,y)}{\varepsilon} \quad \text{for } \varepsilon > 0. \tag{2.2}$$

and when $\beta \stackrel{\text{def.}}{=} \sum_{j=1}^m \boldsymbol{\beta}_j \delta_{y_j}$ is discrete, the potential v is a m-dimensional vector $(\mathbf{v}_j)_j$

and the semi-dual is the maximization of an expectation over α:

$$W_\varepsilon^c(\alpha,\beta) = \max_{\mathbf{v}\in\mathbb{R}^m} \mathbb{E}_\alpha[g_\varepsilon^X(v)],$$

where

$$g_\varepsilon^x(\mathbf{v}) = \sum_{j=1}^m \mathbf{v}_j\boldsymbol{\beta}_j + \begin{cases} -\varepsilon\log(\sum_{j=1}^m \exp(\frac{\mathbf{v}_j - c(x,y_j)}{\varepsilon})\boldsymbol{\beta}_j) & \text{if } \varepsilon > 0, \\ \min_j\left(c(x,y_j) - \mathbf{v}_j\right) & \text{if } \varepsilon = 0, \end{cases} \tag{2.3}$$

This reformulation is at the heart of the methods detailed in the remainder of this article. Note that the dual problem ($\mathcal{D}_\varepsilon$) cannot be cast as an unconstrained expectation maximization problem when $\varepsilon = 0$, because of the constraint on the potentials which arises in that case.

When β is discrete, since the potential v is a m-dimensional vector $(\mathbf{v}_j)_{j=\{1\dots m\}}$ we can compute the gradient and Hessian of g_ε^x. This was already done in Proposition 13 in Chapter 1 but we rewrite their expressions here for convenience.

Proposition 20. *Consider the semi-dual functional g_ε^x defined in (2.3).*
When $\varepsilon > 0$ its gradient is defined by

$$\nabla_v g_\varepsilon^x(\mathbf{v}) = \boldsymbol{\beta} - \chi_\varepsilon(x,\mathbf{v}), \tag{2.4}$$

and the Hessian is given by

$$\partial_v^2 g_\varepsilon^x(\mathbf{v}) = \frac{1}{\varepsilon}\left(\chi_\varepsilon(x)\chi_\varepsilon(x)^T - \mathrm{diag}(\chi_\varepsilon(x;\mathbf{v}))\right), \quad \textit{where} \quad \chi_\varepsilon(x,\mathbf{v})_i = \frac{\exp(\frac{\mathbf{v}_i - c(x,y_i)}{\varepsilon})}{\sum_{j=1}^m \exp(\frac{\mathbf{v}_j - c(x,y_j)}{\varepsilon})}.$$

Besides, $0 \preceq \partial_v^2 g_\varepsilon^x(\mathbf{v}) \preceq \frac{1}{\varepsilon}$ and thus g_ε^x is a convex function with a Lipschitz gradient.

When $\varepsilon = 0$ (standard OT) g_0 is not smooth and a subgradient is given by

$$\nabla_v g_0^x(\mathbf{v}) = \boldsymbol{\beta} - \chi(x,\mathbf{v}), \tag{2.5}$$

$$\textit{where} \quad \chi(x,\mathbf{v})_i = \mathbb{1}_{i=j^*(x)} \quad \textit{with } j^*(x) = \mathrm{argmin}_{i\in\{1\dots m\}} c(x,y_i) - \mathbf{v}_i.$$

Note that since the lower bound on the eigenvalues of the Hessian is 0 the semi-dual functional is convex but not strongly convex as strong convexity requires a strictly positive lower-bound on eigenvalues of the Hessian. We insist on the lack of strong convexity of the semi-dual problem, as it impacts the convergence properties of the stochastic algorithms (stochastic averaged gradient and stochastic gradient descent) used below.

3 Discrete Optimal Transport

We assume in this section that both α and β are discrete measures, i.e. finite sums of Diracs, of the form $\alpha = \sum_{i=1}^n \boldsymbol{\alpha}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^m \boldsymbol{\beta}_j \delta_{y_j}$, where $(x_i)_i \subset \mathcal{X}$ and $(y_j)_j \subset \mathcal{Y}$, and the histogram vector weights are $\boldsymbol{\alpha} \in \Sigma_n$ and $\boldsymbol{\beta} \in \Sigma_m$ where Σ_n denotes the simplex in $\mathbb{R}^n$. These discrete measures may come from the evaluation of continuous densities on a grid, counting features in a structured object, or be empirical measures based on samples. This setting is relevant for several applications, including all known applications of the earth mover's distance. We show in this section that our stochastic formulation can prove extremely efficient to compare measures with a large number of points.

3.1 Discrete Optimization and Sinkhorn

In this setup, the primal ($\mathcal{P}_\varepsilon$), dual ($\mathcal{D}_\varepsilon$) and semi-dual ($\mathcal{S}_\varepsilon$) problems can be rewritten as finite-dimensional optimization problems involving the cost matrix $\mathbf{c} \in \mathbb{R}_+^{n \times m}$ defined by $\mathbf{c}_{i,j} = c(x_i, y_j)$:

$$
\begin{aligned}
& W_\varepsilon(\alpha, \beta) \\
&= \min_{\boldsymbol{\pi} \in \mathbb{R}_+^{n \times m}} \Big\{ \textstyle\sum_{i,j} \mathbf{c}_{i,j} \boldsymbol{\pi}_{i,j} + \varepsilon \sum_{i,j} \Big(\log \frac{\boldsymbol{\pi}_{i,j}}{\boldsymbol{\alpha}_i \boldsymbol{\beta}_j} - 1 \Big) \boldsymbol{\pi}_{i,j} \;\; | \;\; \boldsymbol{\pi} \mathbb{1}_m = \boldsymbol{\alpha}, \boldsymbol{\pi}^\top \mathbb{1}_n = \boldsymbol{\beta} \Big\}, && (\bar{\mathcal{P}}_\varepsilon) \\
&= \max_{\mathbf{u} \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^m} \textstyle\sum_{i=1}^n \mathbf{u}_i \boldsymbol{\alpha}_i + \sum_{j=1}^m \mathbf{v}_j \boldsymbol{\beta}_j - \varepsilon \sum_{i,j} \exp\Big(\frac{\mathbf{u}_i + \mathbf{v}_j - \mathbf{c}_{i,j}}{\varepsilon} \Big) \boldsymbol{\alpha}_i \boldsymbol{\beta}_j, \text{ (for } \varepsilon > 0) && (\bar{\mathcal{D}}_\varepsilon) \\
&= \max_{\mathbf{v} \in \mathbb{R}^m} \bar{G}_\varepsilon(\mathbf{v}) \quad \text{where} \quad \bar{G}_\varepsilon(\mathbf{v}) \overset{\text{def.}}{=} \textstyle\sum_{i=1}^n g_\varepsilon^{x_i}(\mathbf{v}) \boldsymbol{\alpha}_i, && (\bar{\mathcal{S}}_\varepsilon)
\end{aligned}
$$

and g_ε^x is defined in (2.3).

The state-of-the-art method to solve the discrete regularized OT (i.e. when $\varepsilon > 0$) is Sinkhorn's algorithm (Cuturi, 2013, Alg.1), which has linear convergence rate (Franklin and Lorenz, 1989). It corresponds to a block coordinate maximization, successively optimizing ($\bar{\mathcal{D}}_\varepsilon$) with respect to either $\mathbf{u}$ or $\mathbf{v}$ (see Sec. 4.2, Chapter 1 for a thorough presentation). Each iteration of this algorithm is however costly, because it requires a matrix-vector multiplication. Indeed, this corresponds to a "batch" method where all the samples $(x_i)_i$ and $(y_j)_j$ are used at each iteration, which has thus complexity $O(N^2)$ where $N = \max(n, m)$. The prohibitive cost of iterations is a common drawback of batch methods, which thus scale poorly with the size of the problem. Online methods are often preferred when provided with a large number of samples, which is why we resort to stochastic optimization in this context.

3.2 Incremental Discrete Optimization with SAG when $\varepsilon > 0$.

Stochastic gradient descent (SGD) can be used to minimize the finite sum that appears in in $\bar{\mathcal{S}}_\varepsilon$. An index k is drawn from distribution $\boldsymbol{\alpha}$ at each iteration, and he

gradient of that term $g_\varepsilon^{x_k}(\cdot)$ can be used as a proxy for the full gradient in a standard gradient ascent step to maximize $\bar{G}_\varepsilon$.

Algorithm 3 SAG for Discrete OT

Input: step size $C \in \mathbb{R}_+$

Output: dual potential $\mathbf{v} \in \mathbb{R}^m$

$\mathbf{v} \leftarrow \mathbb{0}_m$ (dual potential)
$\mathbf{DG} \leftarrow \mathbb{0}_m$ (proxy of the full gradient $\nabla \bar{G}_\varepsilon$)
$\forall i, \mathbf{z}_i \leftarrow \mathbb{0}_m$ (vector of partial gradients $\nabla g_\varepsilon^{x_k}$)
for $k = 1, 2, \ldots$ **do**
  Sample $i \in \{1, 2, \ldots, n\}$ uniform.
  $\mathbf{DG} \leftarrow \mathbf{DG} - \mathbf{z}_i$ (remove contribution of sample x_i from proxy of $\nabla \bar{G}_\varepsilon$)
  $\mathbf{z}_i \leftarrow \boldsymbol{\alpha}_i \nabla_v g_\varepsilon^{x_i}(\mathbf{v})$ (update gradient of sample x_i)
  $\mathbf{DG} \leftarrow \mathbf{d} + \mathbf{z}_i$ (update proxy of $\nabla \bar{G}_\varepsilon$ with contribution of sample x_i)
  $\mathbf{v} \leftarrow \mathbf{v} + C\mathbf{d}$ (gradient ascent step)
end for

When $\varepsilon > 0$, the finite sum appearing in ($\bar{\mathcal{S}}_\varepsilon$) suggests to use incremental gradient methods – rather than purely stochastic ones – which are known to converge faster than SGD. We propose to use the stochastic averaged gradient (SAG) (Schmidt et al., 2016). The iterates of SAG can be summarized by the following formula

$$\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + \frac{C}{n} \sum_{i=1}^{n} z_i^{(k)},$$

where an index $i^{(k)}$ is selected at random in $\{1 \ldots n\}$ and

$$z_i^{(k)} = \begin{cases} \nabla g_\varepsilon^{x_i}(\mathbf{v}^{(k)}) & \text{if } i = i^{(k)}, \\ z_i^{(k-1)} & \text{otherwise.} \end{cases}$$

At each iteration an index i_k is selected at random in $\{1 \ldots n\}$ to compute $\nabla g_\varepsilon^{x_{i_k}}(\mathbf{v}^{(k)})$, the gradient corresponding to the sample $x_{i^{(k)}}$ at the current estimate $\mathbf{v}^{(k)}$. However, SAG doesn't use this as a proxy for the full gradient $\nabla \bar{G}_\varepsilon$, but rather keeps in memory a copy of that gradient and computes an *average* of all gradients stored so far which provides a better proxy of the gradient corresponding to the entire sum. Another difference is that SAG applies a *fixed* length update, which gives a better convergence rate than SGD:

Proposition 21. *Consider* $\mathbf{v}_\varepsilon^*$ *a minimizer of* $\bar{G}_\varepsilon$*, and* $v^{(k)}$ *that* $k-$*th iterate of SAG defined in* (3.2)*. Then:*

$$|\bar{G}_\varepsilon(\mathbf{v}_\varepsilon^*) - \bar{G}_\varepsilon(\mathbf{v}_k)| = O(1/k).$$

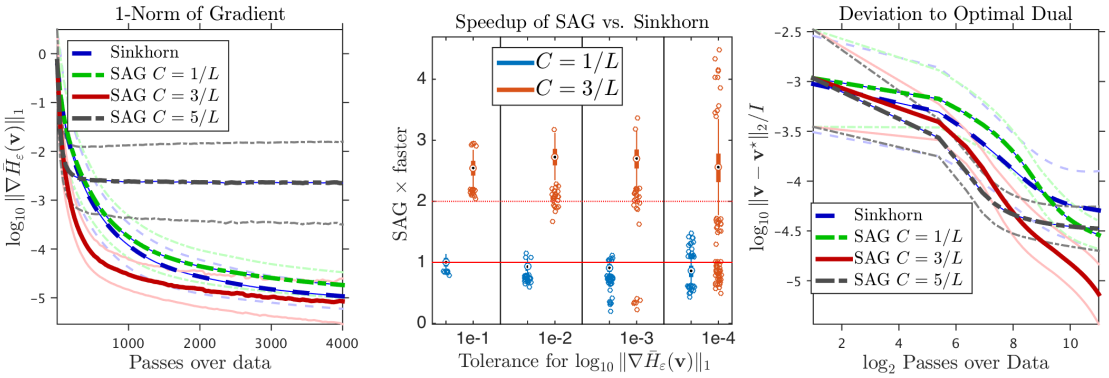


Figure 4.1 – We compute all 595 pairwise word mover's distances (Kusner et al., 2015) between 35 very large corpora of text, each represented as a cloud of $n = 20,000$ word embeddings. We compare the Sinkhorn algorithm with SAG, tuned with different stepsizes. Each pass corresponds to a $n \times n$ matrix-vector product. We used minibatches of size 200 for SAG. *Left plot*: convergence of the gradient ℓ_1 norm (average and $\pm$ standard deviation error bars). A stepsize of $3/L$ achieves a substantial speed-up of ≈ 2.5, as illustrated in the boxplots in the *center plot*. Convergence to $\mathbf{v}^*$ (the best dual variable across all variables after $4,000$ passes) in ℓ_2 norm is given in the *right plot*, up to $2,000 \approx 2^{11}$ steps.

This proposition is a direct application of the convergence rate of SAG for non-strongly convex functions. However, this improvement is made at the expense of storing the gradient for each of the n points. This expense can be mitigated by considering mini-batches instead of individual points. Note that the SAG algorithm is adaptive to strong-convexity and will be linearly convergent around the optimum. The pseudo-code for SAG is provided in Algorithm 3, and we defer more details on SGD for Section 4, in which it will be shown to play a crucial role. Note that the choice of the step-size (C in the algorithm) depends on the Lipschitz constant of all these terms, which is upper bounded by $L = \max_i \boldsymbol{\alpha}_i / \varepsilon$. We discuss this in the following section.

3.3 Numerical Illustrations on Bags of Word-Embeddings.

Comparing texts using the Wasserstein distance on their representations as clouds of word embeddings has been recently shown to yield state-of-the-art accuracy for text classification (Kusner et al., 2015). The authors of the latter have however highlighted that this accuracy comes at a large computational cost. We test our stochastic approach to discrete OT in this scenario, using the complete works of 35 authors [1]. We use Glove word embeddings (Pennington et al., 2014) to represent words, namely $\mathcal{X} = \mathcal{Y} = \mathbb{R}^{300}$. We discard all most frequent $1,000$ words that appear at the top of the

[1] The list of authors we consider is: Keats, Cervantes, Shelley, Woolf, Nietzsche, Plutarch, Franklin, Coleridge, Maupassant, Napoleon, Austen, Bible, Lincoln, Paine, Delafontaine, Dante, Voltaire, Moore, Hume, Burroughs, Jefferson, Dickens, Kant, Aristotle, Doyle, Hawthorne, Plato, Stevenson, Twain, Irving, Emerson, Poe, Wilde, Milton, Shakespeare.

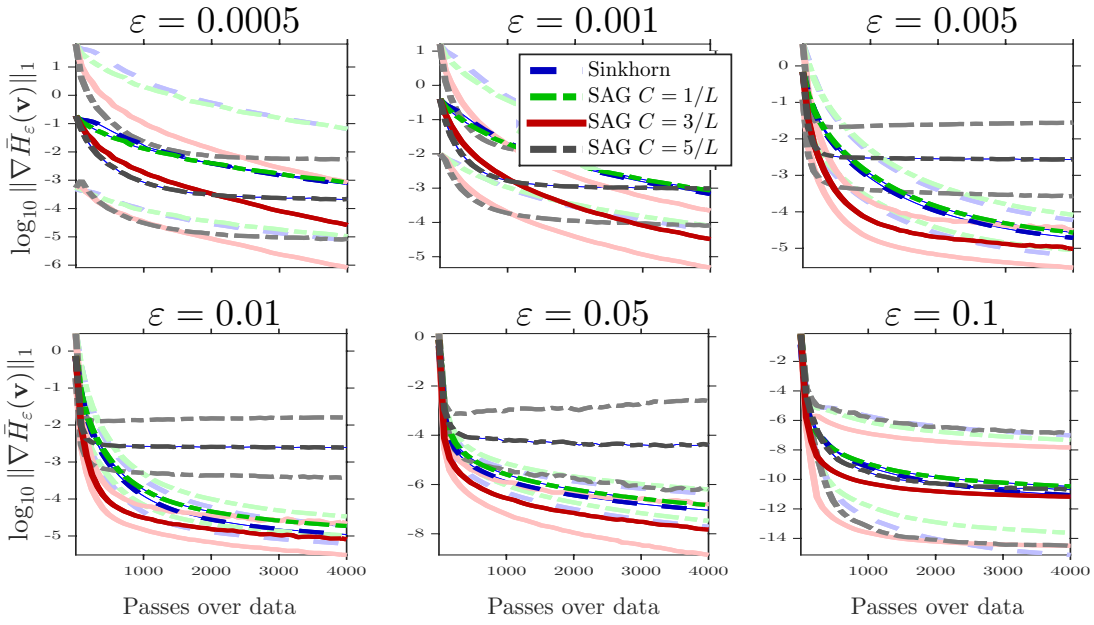


Figure 4.2 – Comparisons between the Sinkhorn algorithm and SAG, tuned with different stepsizes, using different regularization strengths. The setting is identical to that used in Figure 1. Note that to prevent numerical overflow when using very small regularizations, the metric is thresholded such that rescaled costs $c(x, y_j)/\varepsilon$ are not bigger than $\log(10^{200})$.

file `glove.840B.300d` provided on the authors' website. We sample $N = 20,000$ words (found within the remaining huge dictionary of relatively rare words) from each authors' complete work. Each author is thus represented as a cloud of $20,000$ points in $\mathbb{R}^{300}$. The cost function c between the word embeddings is the squared-Euclidean distance, rescaled so that it has a unit empirical median on $2,000$ points sampled randomly among all vector embeddings. We set ε to 0.01 (other values are considered in Figure 4.2). We compute all $(35 \times 34/2 = 595)$ pairwise regularized Wasserstein distances using both the Sinkhorn algorithm and SAG. Following the recommendations in (Schmidt et al., 2016), SAG's stepsize is tested for 3 different settings, $1/L, 3/L$ and $5/L$. The convergence of each algorithm is measured by computing the ℓ_1 norm of the gradient of the full sum (which also corresponds to the marginal violation of the primal transport solution that can be recovered with these dual variables(Cuturi, 2013)), as well as the ℓ_2 norm of the deviation to the optimal scaling found after $4,000$ passes for any of the three methods. Results are presented in Fig. 4.1 and suggest that SAG can be more than twice faster than Sinkhorn on average for all tolerance thresholds. Note that SAG retains exactly the same parallel properties as Sinkhorn: all of these computations can be streamlined on GPUs. We used 4 Tesla K80 cards to compute both SAG and Sinkhorn results. For each computation, all $4,000$ passes take less than 3 minutes (far less are needed if the goal is only to approximate the Wasserstein distance itself, as proposed in (Kusner et al.,

2015)).

4 Semi-Discrete Optimal Transport

In this section, we assume that α is an arbitrary measure (in particular, it needs not to be discrete) and that $\beta = \sum_{j=1}^{m} \boldsymbol{\beta}_j \delta_{y_j}$ is a discrete measure. This corresponds to the semi-discrete OT problem (Aurenhammer et al., 1998; Mérigot, 2011). The semi-dual problem ($\mathcal{S}_\varepsilon$) is then a finite-dimensional maximization problem, written in expectation form as

$$W_\varepsilon(\alpha, \beta) = \max_{\mathbf{v} \in \mathbb{R}^m} G_\varepsilon(\mathbf{v}) \quad \text{where} \quad G_\varepsilon(\mathbf{v}) \overset{\text{def.}}{=} \mathbb{E}_\alpha \left[g_\varepsilon^X(\mathbf{v}) \right],$$

and g_ε^x is defined in (2.3).

4.1 Stochastic Semi-discrete Optimization with SGD

Since the expectation is taken over an arbitrary measure, neither Sinkhorn algorithm nor incremental algorithms such as SAG can be used directly. An alternative is to approximate α by an empirical measure $\hat{\alpha}_N \overset{\text{def.}}{=} \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}$ where $(x_i)_{i=1,\dots,N}$ are i.i.d samples from α, and computing $W_\varepsilon(\hat{\alpha}_N, \beta)$ using the discrete methods (Sinkhorn or SAG) detailed in Section 3. However this introduces a discretization noise in the solution as the discrete problem is now different from the original one and thus has a different solution. SGD on the other hand does not require α to be discrete and is thus perfectly adapted to this semi-discrete setting. The idea of SGD is fairly intuitive : at each iteration, a sample x_k is drawn from α and the gradient $\nabla g_\varepsilon^{x_k}$ is computed at the current iterate $\mathbf{v}^{(k)}$ to serve as a proxy for the full gradient ∇G_ε. The iterates are given by:

$$\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + \frac{C}{\sqrt{k}} \nabla_v g_\varepsilon^{x_k}(\mathbf{v}^{(k+1)}) \quad \text{where} \quad x_k \sim \alpha. \tag{4.1}$$

The convergence rate is given for the average of the iterates, as it is known to converge faster (Polyak and Juditsky, 1992):

Proposition 22. *Consider $\mathbf{v}_\varepsilon^*$ a minimizer of G_ε, and $\mathbf{v}^{(k)}$ the iterates of SGD defined in (4.1). Let $\bar{\mathbf{v}}^{(k)} \overset{\text{def.}}{=} \frac{1}{k} \sum_{i=1}^{k} \mathbf{v}^{(k)}$ the average of these iterates. Then*

$$|G_\varepsilon(\mathbf{v}_\varepsilon^*) - G_\varepsilon(\bar{\mathbf{v}}^{(k)})| = O(1/\sqrt{k}).$$

The algorithm, including the averaging step, is detailed in Algorithm 4.

Recall from Proposition 20 that the gradient of g_ε^x (or subgradient, when $\varepsilon = 0$) is given by

$$\nabla_v g_\varepsilon^x(\mathbf{v}) = \boldsymbol{\beta} - \chi_\varepsilon(x, \mathbf{v}), \quad \text{where } \chi_\varepsilon(x, \mathbf{v})_i = \begin{cases} \frac{\exp(\frac{\mathbf{v}_i - c(x, y_i)}{\varepsilon})}{\sum_{j=1}^{m} \exp(\frac{\mathbf{v}_j - c(x, y_j)}{\varepsilon})} & \text{if } \varepsilon > 0, \\ \mathbb{1}_{i = j^*(x)} & \text{if } \varepsilon = 0, \end{cases}$$

and $j^*(x) = \text{argmin}_{i\in\{1\dots n\}} c(x, y_i) - \mathbf{v}_i$. The function in the gradient, $\chi_\varepsilon(x, \mathbf{v})$ is a smoothed version of the indicator of Laguerre cells with weight vector $\mathbf{v}$ which naturally appear in semi-discrete Optimal Transport (see rem. 11 in sec 4.3.3 of Chapter 1 for a detailed explanation and some illustrations). In particular, (Mérigot, 2011) considers the unregularized dual problem, $\max_{\mathbf{v}\in\mathbb{R}^m} G_0(\mathbf{v})$, where

$$G_0(\mathbf{v}) \overset{\text{def.}}{=} \sum_{j=1}^m \mathbf{v}_j\boldsymbol{\beta}_j + \int_{\mathcal{X}} (\min_k c(x, y_k) - \mathbf{v}_k)\mathrm{d}\alpha(x) = \sum_{j=1}^m \left(\mathbf{v}_j\boldsymbol{\beta}_j + \int_{Lag_j(\mathbf{v})} c(x, y_j) - \mathbf{v}_j \mathrm{d}\alpha(x)\right),$$

and $Lag_j(\mathbf{v})$ is the cell with center y_j in the Laguerre diagram with weights $\mathbf{v}$. The problem is solved using gradient descent, where the gradient is given by

$$(\nabla G_0(\mathbf{v}))_j = \boldsymbol{\beta}_j - \int_{Lag_j(\mathbf{v})} \mathrm{d}\alpha(x).$$

In our stochastic gradient descent approach, for the unregularized case, we are thus replacing the integral over the Laguerre cell, which is very costly to compute, by a simple max search.

Algorithm 4 Averaged SGD for Semi-Discrete OT

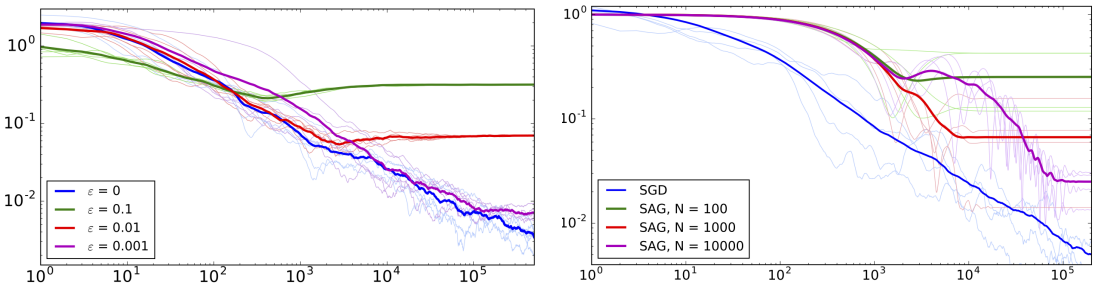
Input: step size $C \in \mathbb{R}_+$
Output: dual potential $\bar{\mathbf{v}} \in \mathbb{R}^m$

$\mathbf{v} \leftarrow \mathbb{0}_m$ (iterates for SGD)
$\bar{\mathbf{v}} \leftarrow \mathbf{v}$ (dual potential obtained by averaging)
for $k = 1, 2, \dots$ **do**
  Sample x_k from α
  $\mathbf{v} \leftarrow \mathbf{v} + \frac{C}{\sqrt{k}} \nabla_v g_\varepsilon^{x_k}(\mathbf{v})$ (gradient ascent step using $\mathbf{v}$)
  $\bar{\mathbf{v}} \leftarrow \frac{1}{k}\mathbf{v} + \frac{k-1}{k}\bar{\mathbf{v}}$ (averaging step to get faster convergence of $\mathbf{v}$)
end for

4.2 Numerical Illustrations on Synthetic Data

Simulations are performed in $\mathcal{X} = \mathcal{Y} = \mathbb{R}^3$. Here α is a Gaussian mixture (continuous density) and $\beta = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$ with $m = 10$ and $(x_j)_j$ are i.i.d. samples from another Gaussian mixture. Each mixture is composed of three Gaussians whose means are drawn randomly in $[0, 1]^3$, and their correlation matrices are constructed as $\Sigma = 0.01(R^T + R) + 3I_3$ where R is 3×3 with random entries in $[0, 1]$. In the following, we denote $\mathbf{v}_\varepsilon^*$ a solution of $(\mathcal{S}_\varepsilon)$, which is approximated by running SGD for 10^7 iterations, 100 times more than those plotted, to ensure reliable convergence curves. Both plots are averaged over 50 runs, lighter lines show the variability in a single run.

Figure 4.3 (a) shows the evolution of $\|\mathbf{v}_k - \mathbf{v}_0^*\|_2 / \|\mathbf{v}_0^*\|_2$ as a function of k. It highlights the influence of the regularization parameters ε on the iterates of SGD. While the



(a) SGD (b) SGD vs. SAG

Figure 4.3 – (a) Plot of $\|\mathbf{v}_k - \mathbf{v}_0^*\|_2 / \|\mathbf{v}_0^*\|_2$ as a function of k, for SGD and different values of ε ($\varepsilon = 0$ being un-regularized). (b) Plot of $\|\mathbf{v}_k - \mathbf{v}_\varepsilon^*\|_2 / \|\mathbf{v}_\varepsilon^*\|_2$ as a function of k, for SGD and SAG with different number N of samples, for regularized OT using $\varepsilon = 10^{-2}$.

regularized iterates converge faster, they do not converge to the correct unregularized solution. This figure also illustrates the convergence theorem of solution of $(\mathcal{S}_\varepsilon)$ toward those $(\mathcal{S}_0)$ when $\varepsilon \to 0$.. Figure 4.3 (b) shows the evolution of $\|\mathbf{v}_k - \mathbf{v}_\varepsilon^*\|_2 / \|\mathbf{v}_\varepsilon^*\|_2$ as a function of k, for a fixed regularization parameter value $\varepsilon = 10^{-2}$. It compares SGD to SAG using different numbers N of samples for the empirical measures $\hat{\alpha}_N$. While SGD converges to the true solution of the semi-discrete problem, the solution computed by SAG is biased because of the approximation error which comes from the discretization of α. This error decreases when the sample size N is increased, as the approximation of α by $\hat{\alpha}_N$ becomes more accurate.

5 Continuous Optimal Transport Using RKHS

In the case where neither α nor β are discrete, problem $(\mathcal{S}_\varepsilon)$ is infinite-dimensional, so it cannot be solved directly using SGD. We propose in this section to solve the initial dual problem $(\mathcal{D}_\varepsilon)$, using expansions of the dual variables in a reproducing kernel Hilbert spaces (RKHS). Comparing two probability distributions thanks to a maximization problem over a RKHS reminds of the definition of Maximum Mean Discrepancy (MMD)(Sriperumbudur et al., 2012), which is described in details in Chapter 1, Sec 2.2. However, unlike the MMD, problem $(\mathcal{D}_\varepsilon)$ involves two different dual functions u and v, one for each measure. Contrarily to the semi-discrete setting, we can only solve the regularized problem here (i.e. $\varepsilon > 0$), since $(\mathcal{D}_\varepsilon)$ cannot be cast as an expectation maximization problem when $\varepsilon = 0$.

5.1 Kernel SGD

We consider two RKHS $\mathcal{H}$ and $\mathcal{G}$ defined on $\mathcal{X}$ and on $\mathcal{Y}$, with kernels κ associated with norms $\|\cdot\|_\mathcal{H}$ and $\|\cdot\|_\mathcal{G}$. Note that we could consider two distinct kernels κ and ℓ for

each RKHS but since we know from Chapter 3, Sec. 4 that both potentials are in similar RKHS (they might be defined on different spaces, but have the same regularity) it is more natural to use the same kernel function κ. Recall the two fundamental properties of RKHS:

(a) if $u \in \mathcal{H}$, then $u(x) = \langle u, \kappa(\cdot, x)\rangle_{\mathcal{H}}$,

(b) $\kappa(x, x') = \langle \kappa(\cdot, x), \kappa(\cdot, x')\rangle_{\mathcal{H}}$.

The dual problem $(\mathcal{D}_\varepsilon)$ is conveniently re-written in Proposition (19) as the maximization of the expectation of $f_\varepsilon^{XY}(u, v)$ with respect to the random variables $(X, Y) \sim \alpha \otimes \beta$, where

$$f_\varepsilon^{xy}(u, v) \stackrel{\text{def.}}{=} u(x) + v(y) - \varepsilon \exp^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}. \tag{5.1}$$

The SGD algorithm applied to this infinite-dimensional problem reads, starting with $u_0 = 0$ and $v_0 = 0$,

$$\begin{cases} u^{(k)} & \stackrel{\text{def.}}{=} u^{(k-1)} + \frac{C}{\sqrt{k}} \nabla_u f_\varepsilon^{x_k, y_k}(u^{(k-1)}, v^{(k-1)}) \\ v^{(k)} & \stackrel{\text{def.}}{=} v^{(k-1)} + \frac{C}{\sqrt{k}} \nabla_v f_\varepsilon^{x_k, y_k}(u^{(k-1)}, v^{(k-1)}), \end{cases} \tag{5.2}$$

where (x_k, y_k) are i.i.d. samples from $\alpha \otimes \beta$ and u and v are functions over $\mathcal{X}$ and $\mathcal{Y}$ respectively. Following Kivinen et al. (2002), we solve this problem with stochastic gradient descent over a RKHS. This amounts to restricting the minimization space to functions that are expansions of kernel functions (property (b) of RKHS stated above). We shows that these $(u^{(k)}, v^{(k)})$ iterates can be expressed as finite sums of kernel functions, with a simple recursion formula.

Algorithm 5 Kernel SGD for continuous OT

Input: step size C, kernel κ

Output: $(w^{(k)}, x_k, y_k)_{k=1,\dots}$

for $k = 1, 2, \dots$ **do**

Sample x_k from α

Sample y_k from β

$u^{(k-1)}(x_k) \stackrel{\text{def.}}{=} \sum_{i=1}^{k-1} w^{(i)} \kappa(x_k, x_i)$

$v^{(k-1)}(y_k) \stackrel{\text{def.}}{=} \sum_{i=1}^{k-1} w^{(i)} \kappa(y_k, y_i)$

$w^{(k)} \stackrel{\text{def.}}{=} \frac{C}{\sqrt{k}} \Big(1 - \exp\Big(\frac{u^{(k-1)}(x_k) + v^{(k-1)}(y_k) - c(x_k, y_k)}{\varepsilon}\Big)\Big)$

end for

Proposition 23. *The iterates of kernel-SGD in a RKHS $\mathcal{H}$ with kernel κ are given by*

$$\begin{cases} u^{(k)} & \stackrel{\text{def.}}{=} \sum_{i=1}^{k} w^{(i)} \kappa(\cdot, x_i) \\ v^{(k)} & \stackrel{\text{def.}}{=} \sum_{i=1}^{k} w^{(i)} \kappa(\cdot, y_i)), \end{cases} \tag{5.3}$$

where

$$w^{(i)} \overset{\text{def.}}{=} \frac{C}{\sqrt{i}}\left(1 - \exp\left(\frac{u^{(i-1)}(x_i) + v^{(i-1)}(y_i) - c(x_i, y_i)}{\varepsilon}\right)\right),$$

and $(x_i, y_i)_{i=1\dots k}$ *are i.i.d samples from* $\alpha \otimes \beta$.

Proof. Replacing $u(x)$ and $v(y)$ by their scalar product formulation in their respective RKHS, $f_\varepsilon^{xy}(u, v)$ can be rewritten

$$f_\varepsilon^{xy}(u, v) = \langle u,\, \kappa(x, \cdot)\rangle_{\mathcal{H}} + \langle v,\, \kappa(y, \cdot)\rangle_{\mathcal{G}} - \varepsilon \exp\Big(\frac{\langle u, \kappa(x, \cdot)\rangle_{\mathcal{H}} + \langle v, \kappa(y, \cdot)\rangle_{\mathcal{G}} - c(x, y)}{\varepsilon}\Big).$$

The partial derivatives with respect to u and v are thus given by

$$\frac{\partial f_\varepsilon^{xy}}{\partial u}(u, v) = \kappa(x, \cdot)\left(1 - \exp\left(\frac{\langle u, \kappa(x, \cdot)\rangle + \langle v, \kappa(y, \cdot)\rangle - c(x, y)}{\varepsilon}\right)\right),$$
$$\frac{\partial f_\varepsilon^{xy}}{\partial v}(u, v) = \kappa(y, \cdot)\left(1 - \exp\left(\frac{\langle u, \kappa(x, \cdot)\rangle + \langle v, \kappa(y, \cdot)\rangle - c(x, y)}{\varepsilon}\right)\right).$$

Plugging this formula in the SGD iteration (5.2) yields : $u^{(k)} = u^{(k-1)} + w^{(k)}\kappa(\cdot, x_k)$ and $v^{(k)} = v^{(k-1)} + w^{(k)}\kappa(\cdot, y_k)$, where $w^{(k)} \overset{\text{def.}}{=} \Big(1 - \exp(\frac{u^{(k-1)}(x_k) + v^{(k-1)}(y_k) - c(x_k, y_k)}{\varepsilon})\Big)$. As we start from $(u^{(0)}, v^{(0)}) = (0, 0)$, the parameters $(w^{(i)})_{i<k}$ are not updated at iteration k, we get the announced formula. □

Algorithm 5 describes our kernel SGD approach, in which both potentials u and v are approximated by a linear combination of kernel functions. After n_{it} iterations, the algorithm returns the samples $(x_k, y_k)_{k=1\dots n_{it}}$ and the iterates $(w^{(k)})_{k=1\dots n_{it}}$ which are stored at each iteration. The dual potentials (u, v) can then be evaluated at any point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ with the following formula

$$u(x) = \sum_{i=1}^{n_{it}} w^{(i)}\kappa(x, x_i) \quad \text{and} \quad v(y) = \sum_{i=1}^{n_{it}} w^{(i)}\kappa(y, y_i).$$

The main cost at each iteration k lies in the computation of the terms $u^{(k-1)}(x_k) \overset{\text{def.}}{=} \sum_{i=1}^{k-1} w^{(i)}\kappa(x_k, x_i)$ and $v^{(k-1)}(y_k) \overset{\text{def.}}{=} \sum_{i=1}^{k-1} w^{(i)}\kappa(y_k, y_i)$ which imply a quadratic complexity $O(k^2)$. Thus the complexity of each iteration increases over time. Several methods exist to alleviate the running time complexity of kernel algorithms, e.g. random Fourier features (Rahimi and Recht, 2007) or incremental incomplete Cholesky decomposition (Wu et al., 2006) whose implementation we detail below.

Proposition 24. ***(Convergence of Kernel SGD)*** *When* α *and* β *are supported on bounded subspaces of* $\mathbb{R}^d$*, then if* κ *is the Matern kernel, or any universal kernel, the iterates* $(u^{(k)}, v^{(k)})$ *defined in proposition 23 converge to a solution of* $(\mathcal{D}_\varepsilon)$.

Proof. To obtain convergence of kernel SGD, we need to make sure that the poten-

tials can be approximated by a linear combination of kernel functions. Theorem 13 in Chapter 3 tells us that if the cost function is smooth enough, the dual variables are also smooth and to belong to a ball with radius independent of α and β in the Sobolev space $\mathbf{H}^s(\mathbb{R}^d)$. Since $\mathbf{H}^s(\mathbb{R}^d)$ is a RKHS for $s > d/2$, its functions can be expressed as a linear combination of the associated kernel, which is called Matérn kernel. Otherwise, universal kernels can by definition approximate any smooth function (Steinwart and Christmann, 2008). □

The choice of the kernel function is instrumental in kernel methods to obtain good performance. Since the dual potentials (u, v) are in $\mathbf{H}^s(\mathbb{R}^d)$ (under smoothness assumptions on the cost) which is a RKHS for $s > d/2$, its associated kernel - called Matérn kernel - is a natural choice. However, their complex definition in dimension larger than 1 makes them impractical. We thus resort to universal kernels, which can by definition approximate any smooth function. In Euclidean spaces $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, where $d > 0$, a natural choice of universal kernel is the Gaussian kernel $\kappa(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$. Tuning its bandwidth σ is crucial to obtain a good convergence of the algorithm, as we will point out in the numerical experiments below.

Finally, let us note that, while entropic regularization of the primal problem $(\mathcal{P}_\varepsilon)$ was necessary to be able to apply semi-discrete methods in Sections 3 and 4, this is not the case here. Indeed, since the kernel SGD algorithm is applied to the dual $(\mathcal{D}_\varepsilon)$, it is possible to replace $KL(\pi|\alpha\otimes\beta)$ appearing in $(\mathcal{P}_\varepsilon)$ by other regularizing divergences. An example of another regularizer would be a χ^2 divergence $\int_{\mathcal{X}\times\mathcal{Y}}(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha\mathrm{d}\beta}(x,y))^2\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$ (with positivity constraints on π). See Chapter 1, Sec. 3 for details on regularizing OT with φ-divergences. However, note that convergence of the iterates is only proved for entropic regularization, as a result of the boundedness of the potentials in Sobolev norm proved in Chapter 3.

5.2 Speeding up Iterations with Kernel Approximation

The main drawback of kernel-SGD is the fact that as the computational time grows quadratically with the number of samples (or equivalently, the number of iterations). We explore here approximate feature expansion methods, which replace the kernel function by the scalar product between two approximate feature functions in low dimension.

5.2.1 Incomplete Cholesky Decomposition

An fundamental property of RKHS is the fact that the kernel function κ can be rewritten as a scalar product of feature maps $\varphi : \mathcal{X} \to F$ where F is the (possibly infinite dimensional) feature space. The idea behind incomplete Cholesky decomposition is to introduce an approximate feature function $\tilde{\varphi}$ that maps data points to a finite dimensional vector, through a kernel matrix computed on a small number of samples.

Algorithm 6 Kernel SGD for continuous OT with incomplete Cholesky decomposition

Input: step size C, kernel κ, feature space dimension I
Output: $(\mathbf{w}_u, \mathbf{w}_v)_{k=1,\dots}$

Sample $(X_I, Y_I) \overset{\text{def.}}{=} (x_i, y_i)_{i=1\dots I}$ from $\alpha \otimes \beta$
$KX \overset{\text{def.}}{=} \kappa(X_I, X_I)$; $KY \overset{\text{def.}}{=} \kappa(Y_I, Y_I)$
Compute $(KX)^{-\frac{1}{2}} \overset{\text{def.}}{=} \texttt{pinv}(\texttt{Chol}(KX))$ (pseudo-inverse of Cholesky root of KX)
$(KY)^{-\frac{1}{2}} \overset{\text{def.}}{=} \texttt{pinv}(\texttt{Chol}(KY))$.
for $k = 1, 2, \dots$ **do**
  Sample (x_k, y_k) from $\alpha \otimes \beta$
  $\tilde{\varphi}_k^x \overset{\text{def.}}{=} (KX)^{-\frac{1}{2}} \kappa(X_I, x_k)$; $\tilde{\psi}_k^y \overset{\text{def.}}{=} (KY)^{-\frac{1}{2}} \kappa(Y_I, y_k)$ (approximate features)
  $\lambda^{(k)} \overset{\text{def.}}{=} \exp\left(\frac{(\tilde{\varphi}_k^x)^T W_{k-1}^u + (\tilde{\psi}_k^y)^T W_{k-1}^v - c(x_k, y_k)}{\varepsilon}\right)$
  $\mathbf{w}_u^{(k)} = \mathbf{w}_u^{(k-1)} + \frac{C_u}{\sqrt{k}}(1 - \lambda^{(k)})\tilde{\varphi}_k^x$; $\mathbf{w}_v^{(k)} = \mathbf{w}_v^{(k-1)} + \frac{C_v}{\sqrt{k}}(1 - \lambda^{(k)})\tilde{\psi}_k^y$
end for

Consider $K_{\mathcal{I}}$ the kernel matrix computed on a sample $\mathcal{I} = (x_1, \dots, x_I)$ of i.i.d. realizations of α, such that $(K_{\mathcal{I}})_i j = \kappa(x_i, x_j)$. Following (Bach, 2013), we use this sample to compute the following approximate feature function:

$$\tilde{\varphi}(x) \overset{\text{def.}}{=} K_{\mathcal{I}}^{-\frac{1}{2}} (\kappa(x_i, x))_{i \in \mathcal{I}} \in \mathbb{R}^I, \tag{5.4}$$

where $K_{\mathcal{I}}^{-\frac{1}{2}}$ is the inverse of the Cholesky decomposition of K_I. One can easily check that for any pair (x_i, x_j) in the dataset $\mathcal{I}$, $\tilde{\varphi}(x_i)^T \tilde{\varphi}(x_j) = \kappa(x_i, x_j)$ Thus, for any pair of points (x, x') we can approximate the kernel by $\kappa(x, x') \simeq \tilde{\varphi}(x)^T \tilde{\varphi}(x')$. The functional whose expectation has to be maximized over (u, v) reads

$$f_\varepsilon^{xy}(u, v) = u(x) + v(y) - \varepsilon \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right).$$

In the RKHS, considering a sample $(x_i, y_i)_{i=1\dots n}$, we can express u and v as linear combinations of kernel functions $u(x) = \sum_{i=1}^n \mathbf{a}_i \kappa(x_i, x)$ and $v(y) = \sum_{i=1}^n \mathbf{b}_i \kappa(y_i, y)$. Replacing the kernel by the scalar product of approximate features, we can rewrite $u(x) = \sum_{i=1}^n \mathbf{a}_i \tilde{\varphi}(x_i)^T \tilde{\varphi}(x)$ and $v(y) = \sum_{i=1}^n \mathbf{b}_i \tilde{\psi}(y_i)^T \tilde{\psi}(y)$ where $\tilde{\varphi}$ (resp. $\tilde{\psi}$) is constructed from the kernel matrix of i.i.d. samples $(x_1, \dots, x_n)$ (resp.$(y_1, \dots, y_n)$) from distribution α (resp. β). Plugging these expressions back in $f_\varepsilon^{xy}(u, v)$, the problem boils down to optimizing over $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^n$

$$\begin{aligned}\tilde{f}_\varepsilon^{xy}(a, b) = &\sum_{i=1}^n \mathbf{a}_i \tilde{\varphi}(x_i)^T \tilde{\varphi}(x) + \sum_{i=1}^n \mathbf{b}_i \tilde{\psi}(y_i)^T \tilde{\psi}(y) \\ &- \varepsilon \exp\left(\frac{\sum_{i=1}^n \mathbf{a}_i \tilde{\varphi}(x_i)^T \varphi(x) + \sum_{i=1}^n \mathbf{b}_i \tilde{\psi}(y_i)^T \tilde{\psi}(y) - c(x, y)}{\varepsilon}\right).\end{aligned}$$

We introduce a change of variables $(\mathbf{w}_u, \mathbf{w}_v) \overset{\text{def.}}{=} (\sum_{i=1}^n \mathbf{a}_i \tilde{\varphi}(x_i), \sum_{i=1}^n \mathbf{b}_i \tilde{\psi}(y_i))$ which yields

$$\tilde{f}_\varepsilon^{xy}(\mathbf{w}_u, \mathbf{w}_v) = \tilde{\varphi}(x)^T \mathbf{w}_u + \tilde{\psi}(y)^T \mathbf{w}_v + \exp\left(\frac{\tilde{\varphi}(x)^T \mathbf{w}_u + \tilde{\psi}(y)^T \mathbf{w}_v - c(x,y)}{\varepsilon}\right).$$

SGD can now be used to compute iterates of $(\mathbf{w}_u, \mathbf{w}_v)$ which are two vectors of size I, whereas $(\mathbf{a}, \mathbf{b})$ were vectors of size n, the size of the sample growing with each iteration of the algorithm. The algorithm for kernel SGD with incomplete Cholesky decomposition is outlined below. The algorithm outputs the pair of vectors $(\mathbf{w}_u, \mathbf{w}_v)$ from which we recover the dual variables via

$$u(x) = \mathbf{w}_u^T \tilde{\varphi}(y) \quad \text{and} \quad u(x) = \mathbf{w}_v^T \tilde{\psi}(y).$$

5.2.2 Random Fourier Features

Random Fourier Features (RFF) are another popular approximation of the feature map in the case where the kernel function is translation invariant i.e. $\kappa(x,y) = \kappa(y-x)$.

Proposition 25. *(Rahimi and Recht, 2007) Consider a translation invariant kernel κ, and let p denote its Fourier transform. Let $(\omega_1, \dots, \omega_D)$ a $D-$sample from p and $(b_1, \dots, b_D)$ a $D-$sample from $\mathcal{U}[0, 2\pi]$. We define the approximate feature map $z : \mathcal{X} \mapsto \mathbb{R}^D$ by*

$$z(x) = \sqrt{\frac{2}{D}}[\cos(\omega_1^T x + b_1), \dots, \cos(\omega_D^T x + b_D)],$$

Then $z(x)^T z(y)$ is an good approximation of $k(x-y)$ with high probability :

$$\mathbb{P}[\sup_{x,y} |z(x)^T z(y) - k(x-y)| \geqslant \varepsilon] = O(\exp \frac{-D\varepsilon}{4(d+2)}).$$

Results from (Rahimi and Recht, 2009) imply that $O(n)$ random features are needed to obtain a $O(1/\sqrt{n})$ bound on the error when learning with RFF in a general setting. However, these bounds are refined in (Rudi and Rosasco, 2017) and (Carratino et al., 2018) to $O(\sqrt{n})$ random features for kernel ridge regression and supervised learning with a squared loss, respectively. Contrarily to the incomplete Cholesky decomposition, which could be used for any positive definite kernel κ, Random Fourier Features require to have an explicit formula for the Fourier transform of κ which restricts the possibilities. For a Gaussian kernel with bandwidth σ, its Fourier transform is a Gaussian with bandwidth $1/\sigma^2$. Thus the frequencies ω are drawn according to a $\mathcal{N}(0, 1/\sigma^2)$. The details of the implementation of kernel SGD with RFF are given in algorithm 7. The procedure is the same as the one used for kernel SGD with incomplete Cholesky, namely using the kernel expansion for u and v and then making a change of variable to reduce the dimensionality of the problem. The approximate feature functions $\tilde{\varphi}$ and $\tilde{\psi}$ from

Algorithm 7 Kernel SGD for continuous OT with Random Fourier Features

Input: C, kernel κ, Fourier transform of the kernel p, dimension of feature space D
Output: $(\mathbf{w}_u, \mathbf{w}_v)_{k=1,\dots}$

Sample $(\omega_1, \dots, \omega_D)$ from p
Sample $(b_1, \dots, b_D)$ from $\mathcal{U}[0, 2\pi]$
Def $z(x)$:
    return $\sqrt{\frac{2}{D}}[\cos(\omega_1^T x + b_1), \dots, \cos(\omega_D^T x + b_D)]$
for $k = 1, 2, \dots$ **do**
    Sample (x_k, y_k) from $\alpha \otimes \beta$
    $z_x^{(k)} = z(x_k) \quad ; \quad z_y^{(k)} = z(y_k)$
    $\lambda^{(k)} \overset{\text{def.}}{=} \exp\left(\frac{(z_x^{(k)})^T \mathbf{w}_u^{(k-1)} + (z_y^{(k)})^T \mathbf{w}_v^{(k-1)} - c(x_k, y_k)}{\varepsilon}\right)$
    $\mathbf{w}_u^{(k)} = \mathbf{w}_u^{(k-1)} + \frac{C_u}{\sqrt{k}}(1 - \lambda^{(k)}) z_x^{(k)} \quad ; \quad \mathbf{w}_v^{(k)} = \mathbf{w}_v^{(k-1)} + \frac{C_v}{\sqrt{k}}(1 - \lambda^{(k)}) z_y^{(k)}$
end for

Cholesky decomposition are replaced by z, the feature map obtained with RFF (note that contrarily to the Cholesky method, we use the same feature map for expansions of u and v). The algorithm outputs the pair of vectors $(\mathbf{w}_u, \mathbf{w}_v)$ from which we recover the dual variables via

$$u(x) = \mathbf{w}_u^T z(x) \quad \text{and} \quad v(y) = \mathbf{w}_v^T z(y).$$

5.3 Comparison of the Three Algorithms on Synthetic Data

We consider optimal transport in 1D between a Gaussian α and a Gaussian mixture β whose densities are represented in Figure 4.4 (a). Since there is no existing benchmark for continuous transport, we use as a proxy for β an empirical distribution $\hat{\beta}_N \overset{\text{def.}}{=} \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$ with $N = 10^3$ and we compute the solution of the semi-discrete problem $W_\varepsilon(\alpha, \hat{\beta}_N)$ with SGD. SGD yields a $N-$dimensional vector $\mathbf{v}$ from which we can compute u at any point of the space thanks to the optimality condition $u(x) = -\varepsilon(\log \frac{1}{N}\sum_{i=1}^N e^{\frac{\mathbf{v}_i - c(x, y_i)}{\varepsilon}})$.

We first exhibit the convergence of the classic method (without speedup by approximate features) by studying the convergence of the potential u. The iterates $u^{(k)}$ are plotted on a grid for different values of k in Figure 4.4 (c), to emphasize the convergence to the proxy $\hat{u}^*$. We can see that the iterates computed with the RKHS converge faster where α has more mass. This makes sense, since convergence estimates are for $\mathbb{E}[f_\varepsilon^{XY}(u^{(k)}, v^{(k)})]$ and thus the value of $u^{(k)}$ has more influence where α has more mass. Figure 4.4 (b) represents the plot of $\|\mathbf{u}^{(k)} - \hat{\mathbf{u}}^*\|_2 / \|\hat{\mathbf{u}}^*\|_2$ where $\mathbf{u}^{(k)}$ (resp. $\hat{\mathbf{u}}^*$) is the evaluation of $u^{(k)}$ (resp. $\hat{u}^*$) on a sample $(x_i)_{i=1\dots N'}$ drawn from α. This gives more emphasis to the norm on points where α has more mass, for the reason given before.

We then compare the classic method to both speedup methods in terms of CPU time. We choose a Gaussian kernel as it is simple to implement in all three cases although it is

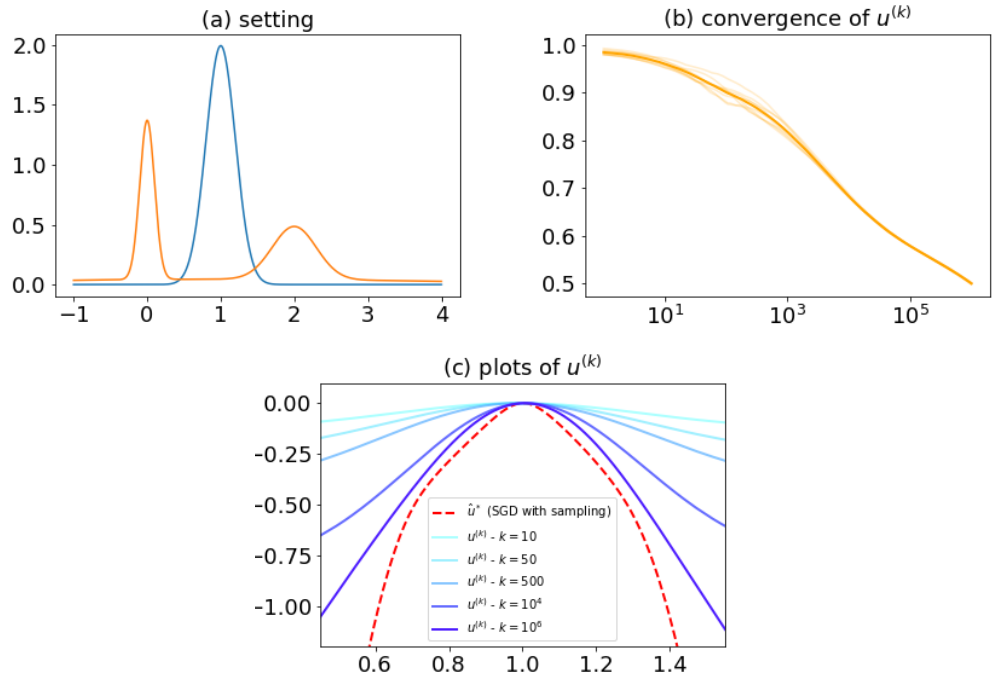


Figure 4.4 – Numerical illustration of the performance of classic kernel-SGD (without features approximation) (a) Plot of $\frac{\mathrm{d}\alpha}{\mathrm{d}x}$ and $\frac{\mathrm{d}\beta}{\mathrm{d}x}$. (b) Plot of $\|\mathbf{u}^{(k)} - \hat{\mathbf{u}}^*\|_2 / \|\hat{\mathbf{u}}^*\|_2$ as a function of k with SGD in the RKHS, for regularized OT using $\varepsilon = 10^{-1}$. (c) Plot of the iterates $u^{(k)}$ for $k = 10^3, 10^4, 10^5$ and the proxy for the true potential $\hat{\mathbf{u}}^*$, evaluated on a grid where α has non negligible mass.

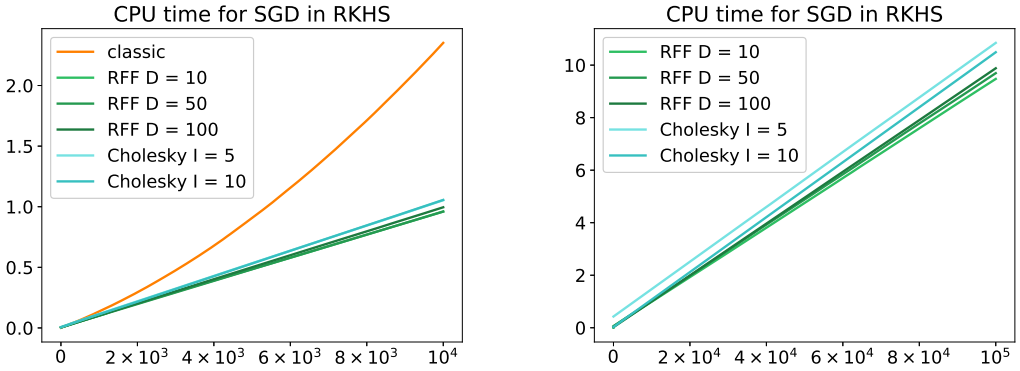


Figure 4.5 – Comparison of the three kernel-SGD algorithms (without speedup, with incomplete Cholesky decomposition, and with Random Fourier Features) for the Gaussian kernel. Computational time is quadratic in the number of iterations for classic kernel-SGD, but becomes linear with an approximate features approach. Increasing the quality of feature approximation (parameter I for Cholesky, D for RFF) does not significantly impact computational time.

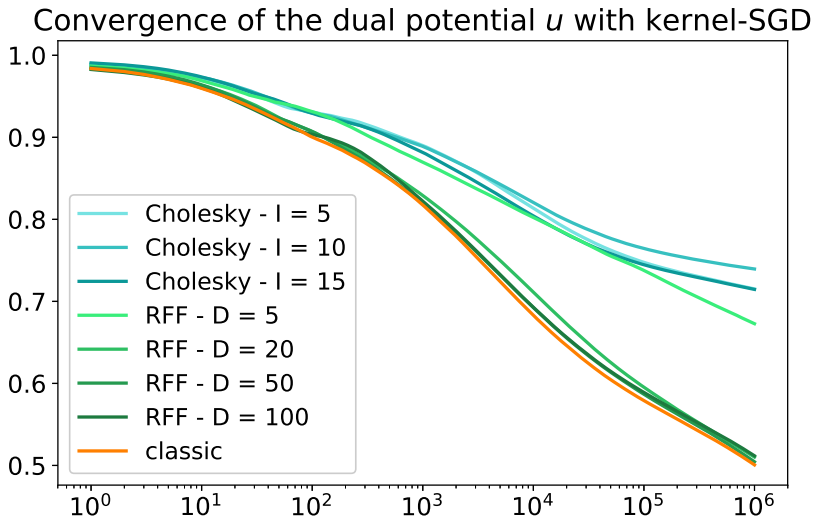


Figure 4.6 – Comparison of convergence of kernel-SGD with different methods, to solve regularized OT using $\varepsilon = 10^{-1}$for different methods. The curves represent $\|\mathbf{u}_k - \hat{\mathbf{u}}^*\|_2/\|\hat{\mathbf{u}}^*\|_2$ as a function of k. Random Fourier Features with $D > 20$ give similar performance to the classic kernel-SGD method in under 3 minutes against over 6 hours for the classic method (for 10^6 iterations).

fairly sensitive to the bandwidth parameter σ. The computation time as a function of the iteration number is given in Figure 4.5, As mentioned previously, the main cost of classic kernel-SGD lies in the computation of the iterates $u^{(k-1)}(x_k) \overset{\text{def.}}{=} \sum_{i=1}^{k-1} w^{(i)}\kappa(x_k, x_i)$ and $v^{(k-1)}(y_k) \overset{\text{def.}}{=} \sum_{i=1}^{k-1} w^{(i)}\kappa(y_k, y_i)$. Thus, the iterations become more costly over time, making this algorithm impractical for applications. On the other hand, for both speedup methods, the computation time of $u^{(k-1)}(x_k)$ is the same for each iteration. The incomplete Cholesky decomposition requires some preprocessing to compute the inverse of the Cholesky root of the kernel matrix on a sample denoted by $(KX)^{-\frac{1}{2}}$, after which the main cost of each iteration is computing the feature vector of x_k, denoted $\tilde{\varphi}_k^x \overset{\text{def.}}{=} (KX)^{-\frac{1}{2}}\kappa(X_I, x_k)$ and then its scalar product with $\mathbf{w}_u^{(k-1)}$ to get $u^{(k-1)}(x_k)$. For the Random Fourier Features, the preprocessing is minimal, as is simply consists in drawing a D-sample from the probability distribution p corresponding to the Fourier transform of the kernel, and another D-sample from the uniform on $[0, 2\pi]$ to define the approximate feature function z. Then the main cost of each iteration also resides in the computation of the feature vector $z(x_k) \overset{\text{def.}}{=} \sqrt{\frac{2}{D}}[\cos(\omega_1^T x + b_1), \ldots, \cos(\omega_D^T x + b_D)]$, before computing its scalar product with $\mathbf{w}_u^{(k-1)}$ to get $u^{(k-1)}(x_k)$. Thus, aside from preprocessing, the difference in iteration time between Cholesky decomposition and RFF lies in the computation of the feature vector. In our implementation, RFF are slightly more efficient, even for larger feature size D.

We are now interested in how the major speedup gained with approximate features

impacts the quality of the solution. The curves in Figure 4.6 plot the convergence of $\mathbf{u}^{(k)}$ to $\hat{\mathbf{u}}^*$ for each of the methods, with the same learning rate C and bandwidth σ. For the Cholesky decomposition, the parameter I controls the quality of the approximation. Note that numerically, we are limited to small values of I (no more than 15) because the eigenvalues of the kernel matrix decay exponentially fast with its dimension and thus its Cholesky root quickly become non-invertible. For the Random Fourier Features, the quality of the approximation is controlled by D, for which we have no particular restriction. From the CPU-time experiment, we see that taking larger values of I and D doesn't impact the computation time very much, but it clearly improves convergence up to a certain point after which there is no more improvement. We can see that Random Fourier Features yield a better approximation than incomplete Cholesky decomposition even for small D. In terms of performance, RFF with feature vectors of size $D = 20, 50, 100$ give similar results, while $D = 5$ is too small to get a good approximation. In terms of computational time, RFF with $D = 20$ takes less than 3 minutes to perform 10^6 iterations, while $D = 100$ takes around 5 minutes, without any significant improvement. In comparison, to reach the same level of precision (which requires the same number of iterations), classic kernel-SGD takes over 6 hours!

Conclusion

Entropy-regularized OT was historically introduced in (Cuturi, 2013) as a computational tool to solve discrete OT efficiently thanks to Sinkhorn's algorithm, and it opened the door to a rich line of research which aims at better understanding its computational and theoretical scope. The contributions of this thesis to this topic can be organized in two main axes. The first one consists in exploiting the properties of entropic regularization to make OT-based losses efficient in machine-learning problems. The second one concerns the interpolation property of entropy-regularized OT, bridging the gap between standard OT and MMD.

Making OT-based Losses Tractable for Machine Learning. Using the entropic regularization with respect to the product measure of the marginals (Genevay et al., 2016) enables us to address both the computational and the statistical issues from which standard OT suffers, by reformulating entropy-regularized OT as the maximization of an expectation (Chapter 1, Sec. 3).

The Statistical Issue: We prove that the dual optimizers of entropy-regularized OT lie in a ball of a Reproducing Kernel Hilbert Space (Chapter 3, Sec. 4). Combined with the formulation as an expectation, this enables us to use techniques from error bounding in learning theory to get a sample complexity result for entropy-regularized OT. We prove that for a large enough regularization, entropy-regularized OT does not suffer from a curse of dimensionality (Chapter 3, Sec. 5).

The Computational Issue: Sinkhorn's algorithm was a major breakthrough for computational OT, but it is limited to discrete measures, and does not scale well when these measures have a very large number of points. The formulation as an expectation allows us to use stochastic optimization solvers, which only require samples from the measures and operate in an online manner (Chapter 4). These algorithms can tackle cases where Sinkhorn is not a suitable choice to solve entropy-regularized OT: discrete problems with a very large number of points, or problems involving continuous measures. For problems involving two continuous measures, we can exploit the fact that the dual optimizers of entropy-regularized OT lie in a ball of a Reproducing Kernel Hilbert Space to derive a provably convergent kernel-SGD solver (Chapter 4, Sec. 5.1).

Minimizing OT-based Losses: We make use of the GPU-friendly structure of Sinkhorn's

algorithm to propose a minimization scheme for OT based-losses (Chapter 2, Sec. 3). We use stochastic gradient over an approximate loss computed with Sinkhorn's algorithm and compute the gradient with automatic differentiation. We use this method to learn a parametric distribution from samples, and prove that it scales well to high-dimensional problems such as generative models of images (Chapter 2, Sec. 4.3) where it can improve on state-of-the-art methods.

Interpolating Between OT and MMD with Sinkhorn Divergences. When comparing one measure to itself, the loss induced by entropy-regularized OT is not equal to zero. To solve this issue, we introduced Sinkhorn Divergences, which are based on entropy-regularized OT with corrective terms. This new family of losses interpolates between OT when the regularization parameter goes to zero and MMD when the regularization parameter goes to infinity (Chapter 2, Sec. 2.4). The interpolation property is also true in terms of sample complexity, which gives theoretical grounds to empirical evidence suggesting that using a regularizer that is not too small is better in practice. Indeed, when the regularization parameter is large enough we recover sample complexity rates from MMD, thus breaking the curse of dimensionality from OT. However, when taking a small regularization, sample complexity degrades quickly in high dimension (Chapter 3, Sec. 5). This theoretical result further advocates for the use of Sinkhorn Divergences with regularization parameters that are not too small. Aside from yielding better performance for machine learning tasks, a large regularization parameter ensures a faster convergence of Sinkhorn's algorithm and is thus also beneficial in terms of computational time. In practice, the regularization parameter in Sinkhorn Divergences gives an additional degree of freedom to the loss function which can be cross-validated to get the best of both OT and MMD in learning tasks (Chapter 2, Sec. 4).

Perspectives for Further Work.

Let us start by mentioning direct extensions of results from this thesis. The first idea to explore is the extension of the sample complexity result from Chapter 3, Theorem 14, to non-smooth cost functions and metric spaces that are not bounded subsets of $\mathbb{R}^d$. Another improvement would consist in tightening the upper-bound on the convergence rate in this theorem to combine it with Theorem 12, which gives a convergence rate on the approximation of OT with regularized OT. Thus we would get a heuristic on the choice of the regularization parameter depending on the number of available samples when one wants to approximate standard OT with regularized OT computed on samples. Another issue linked to this chapter, but which extends beyond the techniques that we used, is to derive convergence rates with respect to the number of samples for the regularized transport plan, i.e. the optimizer of the primal problem, as it is a crucial feature for some machine learning problems (e.g. domain adaptation (Courty et al.,

2016)).

In Chapter 4, Sec. 5.1 we develop an algorithm to compute regularized OT between any two arbitrary measures using kernel-SGD on the dual problem. We only apply our algorithm to a simple 1D problem as a proof of concept, as we did not have any baseline to assess convergence of entropy-regularized OT for continuous measures in high dimension. However, following (Seguy et al., 2017) which uses our scheme with a neural-network parametrization of the dual variables instead of a RKHS expansion, we could use the results directly to perform learning tasks that involve OT between continuous measures. The applications they consider are domain adaptation and image generation. Besides, all stochastic algorithms presented in Chapter 4 could be extended to the case of regularized unbalanced OT (Chizat et al., 2018), as it can also be cast as the maximization of an expectation (see Chapter 1, Remark 6).

The introduction of Sinkhorn Divergences opens the door to several extensions or generalizations. For instance, one might consider unbalanced Sinkhorn Divergences defined with regularized unbalanced OT and see if the interpolation property, positive definiteness and sample complexity results still hold. Another thing would be to extend the sample complexity and positive definiteness results to Sinkhorn Divergences defined without the added entropy in the cost function (see Definition 9 and Remarks 13 and 14), and more generally to understand the potential benefits and drawbacks of using the entropy or not when considering the cost in Sinkhorn Divergences. Eventually, using regularizers other than the relative entropy as introduced in Chapter 1, Sec. 3 is a track worth exploring, although the entropy is central to most of our analysis and only the online algorithm from Chapter 4, Sec. 4 for continuous measures directly applies.

Wasserstein barycenters, which are used to represent the mean of a set of empirical probability measures (Agueh and Carlier, 2011) represent a rich line of research in OT although we did not explore it in this thesis. They can be computed efficiently with entropic regularization (Cuturi and Doucet, 2014), and the online semi-discrete solver developed in Chapter 4, Sec. 4 is well suited to aggregate streaming data (Staib et al., 2017). Given the good numerical results obtained for the Wasserstein barycenter problem with entropy, they could also benefit from the corrective terms in Sinkhorn Divergences.

Taking a step back from entropy-regularized OT, an open issue which is central to Chapter 2 is the evaluation of generative models (see Sec. 4.3). Comparing the outcomes of learning procedures with various losses is a burning issue and the lack of good evaluation metrics makes it impossible to make a definitive ranking of the different losses appearing in the literature. The recent paper by (Lucic et al., 2018) introduces some evaluation metrics suggesting that losses are all equivalent for high dimensional problems such as image generation, making the architecture of the network the most important factor. It is then crucial to understand how the architecture influences smoothness, generalization properties or interpolation in the latent space for instance. Besides,

even though different losses yield similar results for the inference of generative models, entropy-regularized OT can still act as robust metrics to evaluate models *after* the inference procedure.

Eventually, we conclude this thesis with a final question : *are there ways to break the curse of dimensionality for the Wasserstein distance?* Sinkhorn Divergences provide a robust loss for a large enough regularization as seen in Chapter 3, but they do not solve the curse of dimensionality when one wants to compute standard OT from samples. The existence of robust empirical estimators of the Wasserstein distance in high dimension still remains an open question.

Bibliography

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

D. Alvarez-Melis, T. S. Jaakkola, and S. Jegelka. Structured optimal transport. *arXiv preprint arXiv:1712.06199*, 2017.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Springer, 2006.

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

F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302, 2006.

M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

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, 2015.

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Inference in generative models using the Wasserstein distance. *arXiv preprint arXiv:1701.05146*, 2017.

J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for Sinkhorn divergence between probability distributions on finite spaces and statistical applications. *arXiv preprint arXiv:1711.08947*, 2017.

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018.

O. Bousquet, S. Gelly, I. Tolstikhin, C.-J. Simon-Gabriel, and B. Schoelkopf. From optimal transport to generative modeling: the VEGAN cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems*. SIAM, 2009.

P. J. Bushell. Hilbert's metric and positive contraction mappings in a Banach space. *Archive for Rational Mechanics and Analysis*, 52(4):330–338, 1973.

A. Calderón. Lebesgue spaces of differentiable functions. In *Proc. Sympos. Pure Math*, volume 4, pages 33–49, 1961.

G. Canas and L. Rosasco. Learning probability measures with respect to optimal transport metrics. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2492–2500. Curran Associates, Inc., 2012.

G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49 (2):1385–1418, 2017.

L. Carratino, A. Rudi, and L. Rosasco. Learning with sgd and random features. In *Advances in Neural Information Processing Systems*, pages 10213–10224, 2018.

Y. Chen, T. Georgiou, and M. Pavon. Entropic and displacement interpolation: a computational approach using the hilbert metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396, 2016.

L. Chizat. Unbalanced optimal transport: Models, numerical methods, applications. *PhD thesis, PSL Research University*, 2017.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

R. Cominetti and J. S. Martin. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1-3):169–187, 1994.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP (99):1–1, 2016.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975.

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

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

E. Del Barrio and J.-M. Loubes. Central limit theorem for empirical transportation cost in general dimension. *arXiv preprint arXiv:1705.01299*, 2017.

R. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

G. Dziugaite, D. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence-Proceedings of the 31st Conference, UAI 2015*, pages 258–267, 2015.

J. Feydy and A. Trouvé. Global divergences between measures: from Hausdorff distance to Optimal Transport. In *International Workshop on Shape in Medical Imaging*, pages 102–115. Springer, 2018.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2019 (to appear).

J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. Poggio. Learning with a Wasserstein loss. In *Adv. in Neural Information Processing Systems*, pages 2044–2052, 2015.

K. Fukumizu, A. Gretton, G. R. Lanckriet, B. Schölkopf, and B. K. Sriperumbudur. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in neural information processing systems*, pages 1750–1758, 2009.

B. Galerne, A. Leclaire, and J. Rabin. A texture synthesis model based on semi-discrete optimal transport in patch space. *SIAM Journal on Imaging Sciences*, 11(4):2456–2493, 2018.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Proc. NIPS'16*, pages 3432–3440. Curran Associates, Inc., 2016.

A. Genevay, G. Peyre, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2019 (to appear).

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *Adv. in Neural Information Processing Systems*, pages 513–520, 2006.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*, 2017.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

G. Huang, C. Guo, M. J. Kusner, Y. Sun, F. Sha, and K. Q. Weinberger. Supervised word mover's distance. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4862–4870. Curran Associates, Inc., 2016.

L. Kantorovich. On the transfer of masses (in Russian). *Doklady Akademii Nauk*, 37 (2):227–229, 1942.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

J. Kivinen, A. J. Smola, and R. C. Williamson. Online learning with kernels. In *Advances in neural information processing systems*, pages 785–792, 2002.

M. Kusner, Y. Sun, N. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In *Proc. of the 32nd Intern. Conf. on Machine Learning*, pages 957–966, 2015.

A. B. L. Larsen, S. K. Sonderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, New York, New York, USA, 20–22 Jun 2016. PMLR.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards deeper understanding of moment matching network. *arXiv preprint arXiv:1705.08584*, 2017.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are GANs created equal? a large-scale study. In *Advances in neural information processing systems*, pages 697–706, 2018.

G. Luise, A. Rudi, M. Pontil, and C. Ciliberto. Differential properties of Sinkhorn approximation for learning with Wasserstein distance. *arXiv preprint arXiv:1805.11897*, 2018.

C. McDiarmid. *On the method of bounded differences*, page 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

Q. Mérigot. A multiscale approach to optimal transport. *Comput. Graph. Forum*, 30 (5):1583–1592, 2011.

G. Montavon, K.-R. Müller, and M. Cuturi. Wasserstein training of restricted Boltzmann machines. In *Adv. in Neural Information Processing Systems*, 2016.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. *Proc. of the Empirical Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543, 2014.

G. Peyré, M. Cuturi, et al. Computational optimal transport. Technical report, 2017.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Adv. in Neural Information Processing Systems*, pages 1177–1184, 2007.

A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320, 2009.

A. Ramdas, N. G. Trillos, and M. Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2), 2017.

A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 630–638, Cadiz, Spain, 09–11 May 2016. PMLR.

Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, Nov. 2000.

A. Rudi and L. Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

F. Santambrogio. *Optimal Transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their applications*. Springer, 2015.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 2016.

B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *arXiv preprint arXiv:1610.06519*, 2016.

V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.

R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

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 (Proc. SIGGRAPH 2015)*, 2015.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

M. Staib, S. Claici, J. M. Solomon, and S. Jegelka. Parallel streaming Wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pages 2647–2658, 2017.

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

G. Wu, E. Chang, Y. K. Chen, and C. Hughes. Incremental approximate matrix factorization for speeding up support vector machines. In *Proc. of the 12th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*, pages 760–766, 2006.

Résumé

Le Transport Optimal régularisé par l'Entropie (TOE) permet de définir les Divergences de Sinkhorn (DS), une nouvelle classe de distance entre mesures de probabilités basées sur le TOE. Celles-ci permettent d'interpoler entre deux autres distances connues: le Transport Optimal (TO) et l'Ecart Moyen Maximal (EMM). Les DS peuvent être utilisées pour apprendre des modèles probabilistes avec de meilleures performances que les algorithmes existants pour une régularisation adéquate. Ceci est justifié par un théorème sur l'approximation des SD par des échantillons, prouvant qu'une régularisation suffisante permet de se débarrasser de la malédiction de la dimension du TO, et l'on retrouve à l'infini le taux de convergence des EMM. Enfin, nous présentons de nouveaux algorithmes de résolution pour le TOE basés sur l'optimisation stochastique 'en-ligne' qui, contrairement à l'état de l'art, ne se restreignent pas aux mesures discrètes et s'adaptent bien aux problèmes de grande dimension.

Mots Clés

Apprentissage Statistique, Transport Optimal

Abstract

This thesis proposes theoretical and numerical contributions to use Entropy-regularized Optimal Transport (EOT) for machine learning. We introduce Sinkhorn Divergences (SD), a class of discrepancies between probability measures based on EOT which interpolates between two other well-known discrepancies: Optimal Transport (OT) and Maximum Mean Discrepancies (MMD). We develop an efficient numerical method to use SD for density fitting tasks, showing that a suitable choice of regularization can improve performance over existing methods. We derive a sample complexity theorem for SD which proves that choosing a large enough regularization parameter allows to break the curse of dimensionality from OT, and recover asymptotic rates similar to MMD. We propose and analyze stochastic optimization solvers for EOT, which yield online methods that can cope with arbitrary measures and are well suited to large scale problems, contrarily to existing discrete batch solvers.

Keywords

Machine Learning, Optimal Transport