Learning with minibatch Wasserstein : asymptotic and gradient properties

Kilian Fatras[*] **Younes Zine**[†] **Rémi Flamary**[‡] **Rémi Gribonval**[†§] **Nicolas Courty**[*]

[*]Univ Bretagne Sud, Inria, CNRS, IRISA, France
[†]Univ Rennes, Inria, CNRS, IRISA, France
[‡]Univ Côte d'Azur, OCA, UMR 7293, CNRS, Laboratoire Lagrange, France
[§]Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, France

Abstract

Optimal transport distances are powerful tools to compare probability distributions and have found many applications in machine learning. Yet their algorithmic complexity prevents their direct use on large scale datasets. To overcome this challenge, practitioners compute these distances on minibatches *i.e.* they average the outcome of several smaller optimal transport problems. We propose in this paper an analysis of this practice, which effects are not well understood so far. We notably argue that it is equivalent to an implicit regularization of the original problem, with appealing properties such as unbiased estimators, gradients and a concentration bound around the expectation, but also with defects such as loss of distance property. Along with this theoretical analysis, we also conduct empirical experiments on gradient flows, GANs or color transfer that highlight the practical interest of this strategy.

1 Introduction

Measuring distances between probability distributions is a key problem in machine learning. Considering the space of probability distributions $\mathcal{M}_1^+(\mathcal{X})$ over a space $\mathcal{X}$, and given an empirical probability distribution $\alpha \in \mathcal{M}_1^+(\mathcal{X})$, we want to find a parametrized distribution β_λ which approximates the distribution α. Measuring the distance between the distributions requires a function $L : \mathcal{M}_1^+(\mathcal{X}) \times \mathcal{M}_1^+(\mathcal{X}) \to \mathbb{R}$. The distribution β is parametrized by a vector λ and the goal is to find the best λ which minimizes the distance L between β_λ and α, i.e $L(\alpha, \beta_\lambda)$. As the distributions are empirical, we need a distance L with good statistical performance and which have optimization guarantees with modern optimization techniques. Optimal transport (OT) losses as distances have emerged recently as a competitive tool on this problem [Genevay et al., 2018, Arjovsky et al., 2017]. The corresponding estimator is usually found in the literature under the name of *Minimum Kantorovich Estimator* [Bassetti et al., 2006, Peyré and Cuturi, 2019]. Furthermore, OT losses have been widely used to transport samples from a source domain to a target domain using barycentric mappings [Ferradans et al., 2013, Courty et al., 2017, Seguy et al., 2018].

Several previous works challenged the heavy computational cost of optimal transport, as the Wasserstein distance comes with a complexity of $\mathcal{O}(n^3 log(n))$, where n is the size of the probability distribution supports. Variants of optimal transport have been proposed to reduce its complexity. [Cuturi, 2013] used an entropic regularization term to get a strongly convex problem which is solvable using the Sinkhorn algorithm with a computational cost of $\mathcal{O}(n^2)$, both in time and space. However, despite some scalable solvers based on stochastic optimization [Genevay et al., 2016, Seguy et al., 2018], in the big data setting n is very large and still leads to bottleneck computation problems especially when trying to minimize the OT loss. That is why [Genevay et al., 2018, Damodaran et al., 2018] use a minibatch strategy in their implementations to reduce the cost per iteration. They propose to compute the averaged of several optimal transport terms between minibatches from the source and the target distributions. However, using this strategy leads to a different optimization problem that results in a "non optimal" transportation plan between the full original distributions. Recently, [Bernton et al., 2017] worked on minimizers and [Sommerfeld et al., 2019] on a bound between the true optimal transport and the minibatch optimal transport. However they did not study the asymptotic convergence, the loss properties and behavior of the minibatch loss.

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

In this paper we propose to study minibatch optimal transport by reviewing its relevance as a loss function. After defining the minibatch formalism, we will show which properties are inherited and which ones are lost. We describe the asymptotic behavior of the estimator and show that we can derive a concentration bound without dependence on the data space dimension. Then, we prove that the gradients of the minibatch OT losses are unbiased, which justifies its use with SGD in [Genevay et al., 2018]. Finally, we demonstrate the effectiveness of minibatches in large scale setting and show how to alleviate the memory issues for barycentric mapping. The paper is structured as follows: in Section 2, we propose a brief review of the different optimal transport losses. In Section 3, we give formal definitions of the minibatch strategy and illustrate their impacts on OT plans. Basic properties, asymptotic behaviors of the estimator and differentiability are then described. Finally in Section 4, we highlight the behavior of the minibatch OT losses on a number of experiments: gradient flows, generative networks and color transfer.

2 Wasserstein distance and regularization

Wasserstein distance The Optimal Transport metric measures a distance between two probability distributions $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{X})$ by considering a ground metric c on the space $\mathcal{X}$ [Peyré and Cuturi, 2019]. Formally, the Wasserstein distance between two distributions can be expressed as

$$W_c(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\pi \in \boldsymbol{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})} \int_{\mathcal{X} \times \mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) d\pi(\boldsymbol{x}, \boldsymbol{y}), \quad (1)$$

where $\boldsymbol{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the set of joint probability distribution with marginals $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\boldsymbol{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left\{\pi \in \mathcal{M}_+^1(\mathcal{X}, \mathcal{Y}) : \boldsymbol{P}_{\mathcal{X}} \# \pi = \boldsymbol{\alpha}, \boldsymbol{P}_{\mathcal{Y}} \# \pi = \boldsymbol{\beta}\right\}$. $\boldsymbol{P}_{\mathcal{X}} \# \pi$ (resp. $\boldsymbol{P}_{\mathcal{Y}} \# \pi$) is the marginalization of π over $\mathcal{X}$ (resp. $\mathcal{Y}$). The ground cost $c(\boldsymbol{x}, \boldsymbol{y})$ is usually chosen as as the Euclidean or squared Euclidean distance on $\mathbb{R}^d$, in this case W_c is a metric as well. Note that the optimization problem above is called the Kantorovitch formulation of OT and the optimal π is called an optimal transport plan. When the distributions are discrete, the problem becomes a discrete linear program that can be solved with a cubic complexity in the size of the distributions support. Also the convergence in population of the Wasserstein distance is known to be slow with a rate $O(n^{-1/d})$ depending on the dimensionality d of the space $\mathcal{X}$ and the size of the population n [Weed and Bach, 2019]. [Gerber and Maggioni, 2017] used a multi-scale strategy in order to compute a fast approximation of the Wasserstein distance.

Entropic regularization Regularized entropic OT was proposed in [Cuturi, 2013] and leads to a more efficient $\mathcal{O}(n^2)$ solver. We define the entropic loss as:
$W_c^\varepsilon(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\pi \in \boldsymbol{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})} \int_{\mathcal{X} \times \mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) d\pi(\boldsymbol{x}, \boldsymbol{y}) + \varepsilon H(\pi | \xi)$, with $H(\pi|\xi) = \int_{\mathcal{X} \times \mathcal{Y}} \log(\frac{d\pi(\boldsymbol{x}, \boldsymbol{y})}{d\boldsymbol{\alpha}(\boldsymbol{x}) d\boldsymbol{\beta}(\boldsymbol{y})}(\boldsymbol{x}, \boldsymbol{y})) d\pi(\boldsymbol{x}, \boldsymbol{y})$ where $\xi = \boldsymbol{\alpha} \otimes \boldsymbol{\beta}$ and ε is the regularization coefficient. We call this function, the entropic OT loss. As we will see later, this entropic regularization also makes the problem strongly convex and differentiable with respect to the cost or the input distributions.

It is well known that adding an entropic regularization leads to sub-optimal solutions π on the original problem, and it is not a metric since $W_c^\varepsilon(\boldsymbol{\beta}, \boldsymbol{\beta}) \neq 0$. This motivated [Genevay et al., 2018] to introduce an unbiased loss which uses the entropic regularization and called it the Sinkhorn divergence. It is defined as:
$S_c^\varepsilon(\boldsymbol{\alpha}, \boldsymbol{\beta}) = W_c^\varepsilon(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \frac{1}{2}(W_c^\varepsilon(\boldsymbol{\alpha}, \boldsymbol{\alpha}) + W_c^\varepsilon(\boldsymbol{\beta}, \boldsymbol{\beta}))$

It can still be computed with the same order of complexity as the entropic loss and has been proven to interpolate between OT and maximum mean discrepancy (MMD) [Feydy et al., 2019] with respect to the regularization coefficient. MMD are integral probability metrics over a reproducing kernel Hilbert space [Gretton et al.,]. When ε tends to 0, we get the OT solution back and when ε tends to ∞, we get a solution closer to the MMD solution. Second, as proved by [Feydy et al., 2019], if the cost c is Lipschitz, then S_c^ε is a convex, symmetric, positive definite loss function. Hence the use of the Sinkhorn divergence instead of the regularized OT. The sample complexity of the Sinkhorn divergence, that is the convergence rate of a metric between a probability distribution and its empirical counterpart as a function of the number of samples, was proven in [Genevay et al., 2019] to be: $O\left(\frac{e^{\frac{\kappa}{\varepsilon}}}{\sqrt{n}}\left(1 + \frac{1}{\varepsilon^{\lfloor d/2 \rfloor}}\right)\right)$ where d is the dimension of $\mathcal{X}$. We see an interpolation between MMD and OT sample complexity depending on ε.

Minibatch Wasserstein While the entropic loss has better computational complexity than the original Wasserstein distance, it is still challenging to compute it for a large dataset. To overcome this issue, several papers rely on a minibatch computation [Genevay et al., 2018, Damodaran et al., 2018, Liutkus et al., 2019, Kolouri et al., 2016]. Instead of computing the OT problem between the full distributions, they compute an averaged of OT problems between batches of the source and the target domains. It differs from [Gerber and Maggioni, 2017] as the size of the minibatch remains constant. Several work came out to justify the minibatch paradigm. [Bernton et al., 2017] showed that for generative models, the minimizers of the minibatch loss converge to the true minimizer when the minibatch size increases. [Sommerfeld et al., 2019] considered another approach, where they approximate OT with the minibatch strategy and exhibit a deviation bound between the two quantities. We follow a different approach from the two previous work. We are interested in the behavior of using the minibatch strategy as a loss function. We study the asymptotic behavior of using minibatch, the

optimization procedure, the resulting transportation plan and the behavior of such a loss for data fitting problems.

3 Minibatch Wasserstein

In this section we first define the Minibatch Wasserstein and illustrate it on simple examples. Next we study its asymptotic properties and optimization behavior.

3.1 Notations and Definitions

Notations Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ (resp. $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$) be samples of n *iid* random variables drawn from a distribution α (resp. β) on the source (resp. target) domain. We denote by α_n and β_n the empirical distributions of support $\{X_1, \cdots, X_n\}$ and $\{Y_1, \cdots, Y_n\}$ respectively. The weights of X_i (resp. Y_i) are uniform, i.e equal to $1/n$. We further suppose that α and β have compact support, the ground cost is then bounded by a constant M. $\alpha^{\otimes m}$ denotes a sample of m random variables following α. In the rest of the paper, we will not make a difference between a batch A of cardinality m and its associated (uniform probability) distribution $\hat{A} := \frac{1}{m}\sum_{a\in A}\delta_a$. The number of possible mini-batches of size m on n distinct samples is the binomial coefficient $\binom{n}{m} = \frac{n!}{m!(n-m)!}$. For $1 \leqslant m \leqslant n$, we write $\mathcal{P}_m(\alpha_n)$ (resp. $\mathcal{P}_m(\beta_n)$) the collection of subsets of cardinality m of α_n (resp. of β_n). We will denote the integer part of the ratio n/m as $\lfloor n/m \rfloor$.

Definitions We will first give formal definitions of the different quantities that we will use in this paper. We start with minibatch Wasserstein losses for continuous, semi-discrete and discrete distributions.

Definition 1 (Minibatch Wasserstein definitions)**.** *Given an OT loss h and an integer $m \leq n$, we define the following quantities:*

The continuous loss:

$$U_h(\alpha, \beta) := \mathbb{E}_{(X,Y)\sim\alpha^{\otimes m}\otimes\beta^{\otimes m}}[h(X, Y)] \tag{2}$$

The semi-discrete loss:

$$U_h(\alpha_n, \beta) := \binom{n}{m}^{-1} \sum_{A\in\mathcal{P}_m(\alpha_n)} \mathbb{E}_{Y\sim\beta^{\otimes m}}[h(A, Y)] \tag{3}$$

The discrete-discrete loss:

$$U_h(\alpha_n, \beta_n) := \binom{n}{m}^{-2} \sum_{A\in\mathcal{P}_m(\alpha_n)} \sum_{B\in\mathcal{P}_m(\beta_n)} h(A, B) \tag{4}$$

where h can be the Wasserstein distance W, the entropic loss W_ε or the sinkhorn divergence S_ε for a cost $c(\boldsymbol{x}, \boldsymbol{y})$.

Note that h is a U-statistic kernel. Note also that the mini-batches elements are drawn without replacement. These quantities represent an average of Wasserstein distance over minibatches of size m. Note that samples in A have uniform weights $1/m$ and that the ground cost can be computed between all pair of batches A and B. It is easy to see that (4) is an empirical estimator of (2). In real world applications, computing the average over all batches is too costly as we have a combinatorial number of batches, that is why we will rely on a subsampled quantity.

Definition 2 (Minibatch subsampling)**.** *Pick an integer $k > 0$. We define:*

$$\widetilde{U}_h^k(\alpha_n, \beta_n) := k^{-1} \sum_{(A,B)\in D_k} h(A, B) \tag{5}$$

where D_k is a set of cardinality k whose elements are drawn at random from the uniform distribution on $\Gamma := \mathcal{P}_m(\{X_1, \cdots, X_n\}) \times \mathcal{P}_m(\{Y_1, \cdots, Y_n\})$.

As the transportation plan might be of interest, let us now review the minibatch definition for the OT plan which can be built for all OT variants which have an OT plan. Formal definitions are provided in appendix.

Definition 3 (Mini-batch transport plan)**.** *Consider α_n and β_n two discrete probability distributions. For each $A = \{a_1, \ldots, a_m\} \in \mathcal{P}_m(\alpha_n)$ and $B = \{b_1, \ldots, b_m\} \in \mathcal{P}_m(\beta_n)$ we denote by $\Pi_{A,B}$ the optimal plan between the random variables, considered as a $n \times n$ matrix where all entries are zero except those indexed in $A \times B$. We define the averaged mini-batch transport matrix:*

$$\Pi_m(\alpha_n, \beta_n) := \binom{n}{m}^{-2} \sum_{A\in\mathcal{P}_m(\alpha_n)} \sum_{B\in\mathcal{P}_m(\beta_n)} \Pi_{A,B}. \tag{6}$$

Following the subsampling idea, we define the subsampled minibatch transportation matrix for A and B:

$$\Pi_k(\alpha_n, \beta_n) := k^{-1} \sum_{(A,B)\in D_k} \Pi_{A,B} \tag{7}$$

where D_k is drawn as in Definition 2.

It is well known that the Wasserstein distance suffers from biased gradients [Bellemare et al., 2017]. We study if $U_h(\alpha_n, \beta_n)$ has a bias *wrt* $U_h(\alpha, \beta)$, and then the bias in $U_h(\alpha_n, \beta_n)$ gradients for first order optimization methods.

3.2 Illustration on simple examples

To illustrate the effect of the minibatch, we compute Π_m (6) on two simple examples.

Distributions in 1D The 1D case is an interesting problem because we have access to a closed-form of the optimal transport solution which allows us to calculate the closed-form of a minibatch paradigm. It is the foundation of the sliced Wasserstein distance [Bonnotte, 2013] which

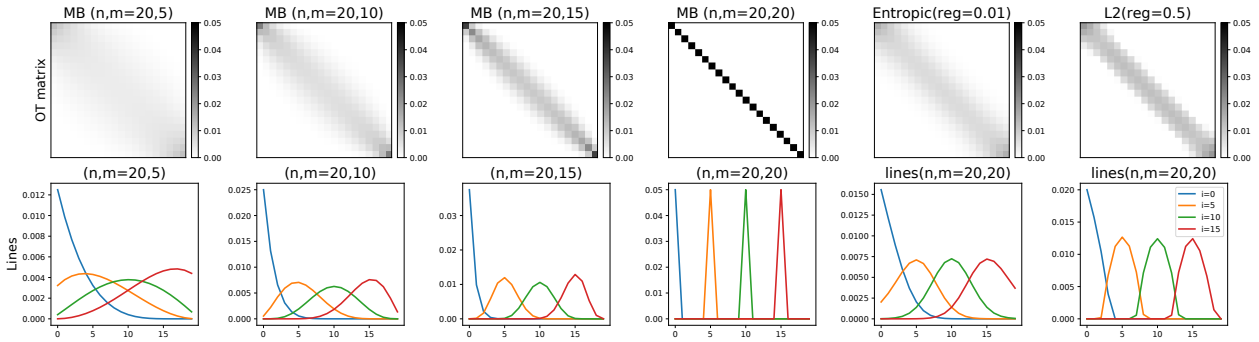

Figure 1: Several OT matrices between distributions with $n = 20$ samples in 1D. The first row shows the minibatch OT matrices Π_m for different values of m, the second row provides the shape of the distributions on the rows of Π_m. The two last columns correspond to classical entropic and quadratic regularized OT.

is widely used as an alternative to the Wasserstein distance [Liutkus et al., 2019, Kolouri et al., 2016].

We suppose that we have uniform empirical distributions α_n and β_n. We assume (without loss of generality) that the points are ordered in their own distribution. In such a case, we can compute the 1D Wasserstein 1 distance with cost $c(x, y) = |x - y|$ as: $W(\alpha_n, \beta_n) = \frac{1}{n}\sum_{i=1}^{n} |x_i - y_j|$ and the OT matrix is simply an identity matrix scaled by $\frac{1}{n}$ (see [Peyré and Cuturi, 2019] for more details). After a short combinatorial calculus (given in appendix A.5), the 1D minibatch transportation matrix coefficient $\pi_{j,k}$ can be computed as $\pi_{j,k} =$:

$$\frac{1}{m}\binom{n}{m}^{-2} \sum_{i=i_{\min}}^{i_{\max}} \binom{j-1}{i-1}\binom{k-1}{i-1}\binom{n-j}{m-i}\binom{n-k}{m-i}$$

where $i_{\min} = \max(0, m - n + j, m - n + k)$ and $i_{\max} = \min(j, k)$. $i_{\min}$ and $i_{\max}$ represent the sorting constraints.

We show on the first row Figure 1 the minibatch OT matrices Π_m with $n = 20$ samples for different value of the minibatch size m. We also provide on the second row of the figure a plot of the distributions in several rows of Π_m. We give the matrices for entropic and quadratic regularized OT for comparison purpose. It is clear from the figure that the OT matrix densifies when m decreases, which has a similar effect as entropic regularization. Note the more localized spread of mass of quadratic regularization that preserve sparsity as discussed in [Blondel et al., 2018]. While the entropic regularization spreads the mass in a similar manner for all samples, minibatch OT spreads less the mass on samples at the extremities. Note that the minibatch OT matrices solution is for ordered samples and do not depend on the position of the samples once ordered, as opposed to the regularized OT methods. This will be better illustrated in the next example.

Minibatch Wasserstein in 2D We illustrate the OT matrix between two empirical distributions of 10 samples each in 2D in Figure 2. We use two 2D empirical distributions (point cloud) where the samples have a cluster structure and the samples are sorted *w.r.t.* their cluster. We can see from the OT matrices in the first row of the figure that the cluster structure is more or less recovered with the regularization effect of the minibatches (and also regularized OT). On the second row one can see the effect of the geometry of the samples on the spread of mass. Similarly to 1D, for Minibatch OT, samples on the border of the simplex cannot spread as much mass as those in the center and have darker rows. This effect is less visible on regularized OT.

3.3 Basic properties

We now state some basic properties for minibatch Wasserstein losses. All properties are proved in the appendix. The first property is about the transportation plan Π_m between the two initial distributions, defined in (6).

Proposition 1. *The transportation plan $\Pi_m(\alpha_n, \beta_n)$ is an admissible transportation plan between the full input distributions α_n, β_n, and we have : $U_h(\alpha_n, \beta_n) \geq W(\alpha_n, \beta_n)$.*

The fact that Π_m is an admissible transportation plan means that even though it is not optimal, we still do transportation similarly to regularized OT. Note that Π_k is not a transportation plan, in general, for a finite k but we study its asymptotic convergence to marginals in the next section. Regarding our empirical estimator, when we have *iid* data, it enjoys the following property:

Proposition 2 (Unbiased estimator)**.** *$U_h(\alpha_n, \beta_n)$ is an unbiased estimator of $U_h(\alpha, \beta)$ for the continuous setting and of $U_h(\alpha_n, \beta)$ for the semi-discrete setting.*

As we use minibatch OT for loss function, it is of interest to see if it is still a distance on the distribution space such as the Wasserstein distance or the Sinkhorn divergence.

Proposition 3 (Positivity and symmetry)**.** *The minibatch Wasserstein losses are positive and symmetric losses. However, they are not metrics since $U_h(\alpha, \alpha) > 0$.*

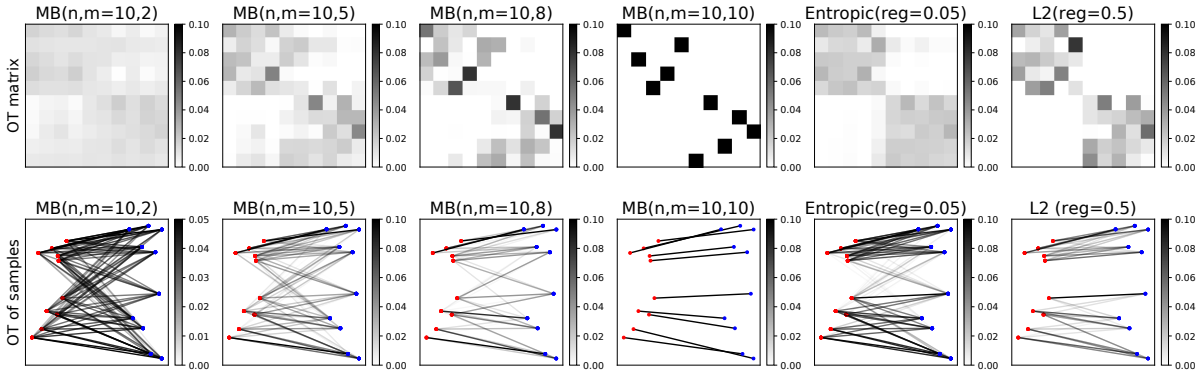

Figure 2: Several OT matrices between 2D distributions with $n = 10$ samples. The first row shows the minibatch OT matrices Π_m for different values of m, the second row provide the shape of the distributions on the rows of the OT matrices. The second row provide a 2D visualization of where the mass is transported between the 2D positions of the sample.

The minibatch Wasserstein losses inherits some properties from the Wasserstein distance but the minibatch procedure leads to a strictly positive loss even when starting from unbiased losses such as Sinkhorn divergence or Wasserstein distance. Remarkably, the Sinkhorn divergence was introduced in the literature to correct the bias from the entropic regularization, and interestingly it was performed in practice on GANs experiments with a minibatch strategy which reintroduced a bias. Whether removing the bias by following the same idea than the Sinkhorn divergence leads to a positive loss is an open question left to future work. Furthermore, given the definition of the minibatch losses it is natural to conjecture that they are convex. Informal ingredients towards a proof of this fact are given in the supplementary material.

An important parameter is the value of the minibatch size m. We remark that the minibatch procedure allows us to interpolate between OT, when $m = n$ and averaged pairwise distance, when $m = 1$. The value of m will also be important for the convergence of our estimator as we will see in the next section.

3.4 Asymptotic convergence

We are now interested in the asymptotic behavior of our estimator $\widetilde{U}_h^k(\alpha_n, \beta_n)$ and its deviation to $U_h(\alpha, \beta)$. We will give a deviation bound between our subsampled estimator and the expectation (taken on both drawn minibatches and drawn empirical data) of our estimator. This result is given in the continuous setting but a similar result holds for the semi-discrete setting and it follows the same proof. We will give a bound with respect to both k and n.

Theorem 1 (Maximal deviation bound)**.** *Let $\delta \in (0, 1)$, $k \geqslant 1$ and m be fixed, and consider two distributions α, β with bounded support and an OT loss $h \in \{W, W_\epsilon, S_\epsilon\}$. We have a deviation bound between $\widetilde{U}_h^k(\alpha_n, \beta_n)$ and $U_h(\alpha, \beta)$ depending on the number of empirical data n and the number of batches k, with probability at least $1 - \delta$ on the draw of α_n, β_n and D_k we have:*

$$|\widetilde{U}_h^k(\alpha_n, \beta_n) - U_h(\alpha, \beta)| \leq M_h\left(\sqrt{\frac{\log(\frac{2}{\delta})}{2\lfloor \frac{n}{m} \rfloor}} + \sqrt{\frac{2\log(\frac{2}{\delta})}{k}}\right)$$

where M_h depends on h and scales at most as $\mathcal{O}(\log(m))$.

This result can be extended with a Bernstein bound (see appendix). The proof is based on two quantities gotten from the triangle inequality. The first quantity is the difference between $U_h(\alpha_n, \beta_n)$ and its expectation $U_h(\alpha, \beta)$. $U_h(\alpha_n, \beta_n)$ is a two-sample U-statistic and we can prove a bound between itself and its expectation in probability [Hoeffding, 1963]. The second quantity is the difference between $U_h(\alpha_n, \beta_n)$ and the expectation of $\widetilde{U}_h^k(\alpha_n, \beta_n)$. We use the difference between the two quantities to obtain a new random variable quantity. From this new random variable, we use the Hoeffing inequality to obtain a dependence with respect to k.

This deviation bound shows that if we increase the number of data n and batches k while keeping the minibatch size m fixed, we get closer to the expectation. We will investigate the dependence on k and m in different scenarios in the numerical experiments. Remarkably, the bound does not depend on the dimension of $\mathcal{X}$, which is an appealing property when optimizing in high dimension.

As discussed before, an interesting output of Minibatch Wasserstein is the minibatch OT matrix Π_m. Since it is hard to compute in practice, we investigate the error on the marginal constraint of Π_k. In what follows, we denote by $\Pi_{(i)}$ the i-th row of matrix Π and by $\mathbf{1} \in \mathbb{R}^n$ the vector whose entries are all equal to 1.

Theorem 2 (Distance to marginals)**.** *Let $\delta \in (0, 1)$, and consider two distributions α_n, β_n. For all $k \geqslant 1$, all $1 \leqslant$*

$i \leqslant n$, with probability at least $1-\delta$ on the draw of α_n, β_n and D_k we have:

$$|\Pi_k(\alpha_n, \beta_n)_{(i)}\mathbf{1} - \frac{1}{n}| \leqslant \sqrt{\frac{2\log(2/\delta)}{k}}. \tag{8}$$

The proof uses the convergence of Π_k to Π_m and the fact that Π_m is a transportation plan and respects the marginals.

3.5 Gradient and optimization

In this section we review the optimization properties of the minibatch OT losses to ensure the convergence of our loss functions with modern optimization frameworks. We study a standard parametric data fitting problem. Given some discrete samples $(x_i)_{i=1}^n \subset \mathcal{X}$ from some unknown distribution α , we want to fit a parametric model $\lambda \mapsto \beta_\lambda \in \mathcal{M}(\mathcal{X})$ to α using the mini-batch Wasserstein distance for a set Λ in an Euclidian space.

$$\min_{\lambda \in \Lambda} \quad U_h(\alpha_n, \beta_\lambda) \tag{9}$$

Such problems are written as semi discrete OT problems because one of the distributions is continuous while the other one is discrete. For instance, generative models fall under the scope of such problems [Genevay et al., 2018] also known as minimal Wasserstein estimation. As we have an expectation over one of the distributions, we would like to use a stochastic gradient descent strategy to minimize the problem. By using SGD for their method, [Genevay et al., 2018] observed that it worked well in practice and they got meaningful results with minibatches. However it is well known that the empiricial Wasserstein distance is a biased estimator of the Wasserstein distance over the true distributions and leads to biased gradients as discussed in [Bellemare et al., 2017], hence SGD might fail. The goal of this section is to prove that unlike the full Wasserstein distance, the minibatch strategy does not suffer from biased gradients.

As stated in Proposition 2, we enjoy an unbiased estimator. However, the original Wasserstein distance is not differentiable, hence we will, further on, only consider the entropic loss and the Sinkhorn divergence which are differentiable.

Theorem 3 (Exchange of Gradient and expectation)**.** *Consider two distributions α and β on two bounded subsets $\mathcal{X}$ and $\mathcal{Y}$, a $\mathcal{C}^1$ cost. Assume $\lambda \mapsto Y_\lambda$ is differentiable. Then we are allowed to exchange gradients and expectation when h is the entropic loss or the Sinkhorn divergence:*

$$\nabla_\lambda \mathop{\mathbb{E}}_{Y_\lambda \sim \beta_\lambda^{\otimes m}} h(A, Y_\lambda) = \mathop{\mathbb{E}}_{Y_\lambda \sim \beta_\lambda^{\otimes m}} \nabla_\lambda h(A, Y_\lambda) \tag{10}$$

The proof relies on the differentiation lemma. Contrary to the full Wasserstein distance, we proved that the minibatch OT losses do not suffer from biased gradients and this justifies the use of SGD to optimize the problem.

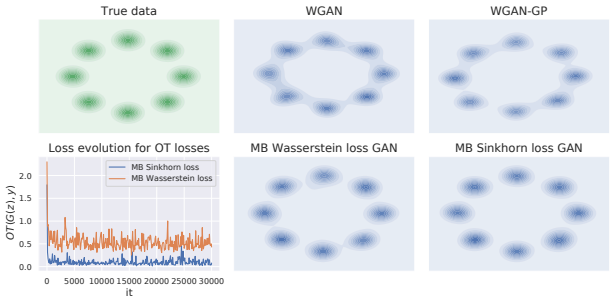

Figure 3: Generated data in 2D for gaussian modes for several generative models.

4 Experiments

In this section, we illustrate the behavior of minibatch Wasserstein. We use it as a loss function for generative models, use it for gradient flow and color transfer experiments. For our experiments, we relied on the POT package [Flamary and Courty, 2017] to compute the exact OT solver or the entropic OT loss and the Geomloss package [Feydy et al., 2019] for the Sinkhorn divergence. The generative model and gradient flow experiments were designed in PyTorch [Paszke et al., 2017] and all the code is released here *.

4.1 Minibatch Wasserstein generative networks

We illustrate the use of minibatch Wasserstein loss for generative modeling [Goodfellow et al., 2014]. The goal is to learn a generative model to generate data close to the target data. We draw 8000 points which follow 8 different gaussian modes (1000 points per mode) in 2D where the modes form a circle. After generating the data, we use a minibatch Wasserstein distance and minibatch Sinkhorn divergence as loss functions with a squared euclidian cost and compared them to WGAN [Arjovsky et al., 2017] and its variant with gradient penalty WGAN-GP [Gulrajani et al., 2017]. We give implementation details in supplementary.

We show the estimated 2D distributions in Figure 3. For the same architecture it seems that MB Wasserstein trains better generators than WGAN and WGAN-GP. This could come from the fact that MB Wasserstein minimize a complex but well posed objective function (with the squared euclidian cost) while WGAN still need to solve the minmax problem making convergence more difficult especially on this 2D problem.

4.2 Minibatch Wasserstein gradient flow

For a given target distribution α, the purpose of gradient flows is to model a distribution $\beta(t)$ which at each iter-

*https://github.com/kilianFatras/minibatch_Wasserstein

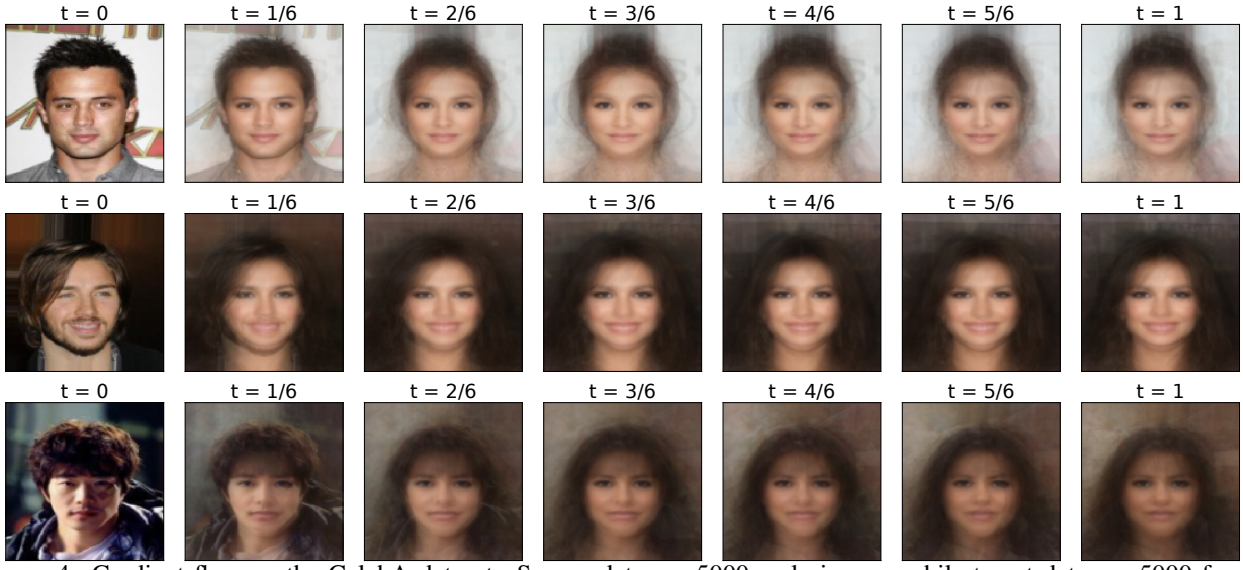

Figure 4: Gradient flow on the CelebA dataset. Source data are 5000 male images while target data are 5000 female images. The batch size m is set to 500 and the number of minibatch k is set to 10. The results were computed with the minibatch Wasserstein distance.

ation follows the gradient direction to minimize the loss $\beta_t \mapsto h(\alpha, \beta_t)$ [Peyré, 2015, Liutkus et al., 2019]. The gradient flow simulate the non parametric setting of data fitting problem. In this setting, the modeled distribution β is parametrized by a vector λ which is the vector position $\boldsymbol{x}$ that encodes its support.

We follow the same procedure as in [Feydy et al., 2019]. The original gradient flow algorithm uses an Euler scheme. Formally, starting from an initial distribution at time $t = 0$, it means that at each iteration we integrate the ODE

$$\dot{\boldsymbol{x}}(t) = -\nabla_{\boldsymbol{x}} F\left(\boldsymbol{x}(t)\right).$$

In our case, we cannot compute the gradient directly from our minibatch OT losses. As the OT loss inputs are distributions, we have an inherent bias when we calculate the gradient from the weights $\frac{1}{m}$ of samples. To correct this bias, we multiply the gradient by the inverse weight m. Finally, for each data $\boldsymbol{x}$ we integrate:

$$\dot{\boldsymbol{x}}(t) = -m\nabla_{\boldsymbol{x}} \left[\widetilde{U}_h^k(\alpha_n, \beta_n)\right](\boldsymbol{x}(t)) \tag{11}$$

We recall that the inherent bias from minibatch makes that the final solution can not be the target distribution.

The considered data are from the CelebA dataset [Liu et al., 2015]. We use 5000 male images as source data and 5000 female images as target data. We show the evolution of 3 samples in the source data in Figure 4. We use a squared euclidean cost, a batch size of 500, a learning rate of 0.05 and make 750 iterations. k did not need to be large and was set to 10 in order to stabilize the gradient flow. We see a natural evolution in the images along the gradient flow similar to results obtained in [Liutkus et al., 2019]. Interestingly the gradient flow with MB Wasserstein in Figure 4 leads to possibly more detailed backgrounds than with MB Sinkhorn (provided in supplementary) probably due to the two layers of regularization in the latter.

4.3 Large scale barycentric mapping for color transfer

The purpose of color transfer is to transform the color of a source image so that it follows the color of a target image. Optimal Transport is a well known method to solve this problem and has been studied before in [Ferradans et al., 2013, Blondel et al., 2018]. Images are represented by point clouds in the RGB color space identified with [0, 1]. Then by calculating the transportation plan between the two point clouds, we get a transfer color mapping by using a barycentric projection. As the number of pixels might be huge, previous work selected a subset of pixels using k-means clusters for each point cloud. This strategy allows to make the problem memory tractable but looses some information. With MB optimal transport, we can compute a barycentric mapping for all pixels in the image by incrementally updating the mapping at each minibtach. When one selects a source batch A and a target batch B, she just needs to update the transformed vector between the considered batches as $Y_s\big|_A = \sum_{B \in \mathcal{P}_m(\beta_n)} \Pi_{A,B} X_t\big|_B$. Indeed, to perform the color transfer when we have the full Π_k matrix, we compute the matrix product:

$$Y_s = n_s \Pi_k(\alpha_n, \beta_n) X_t \tag{12}$$

that can be computed incrementally by considering restriction to batches (the full algorithm is given in appendix). To the best of our knowledge, it is the first time that a barycentric mapping algorithm has been scaled up to 1M pixel

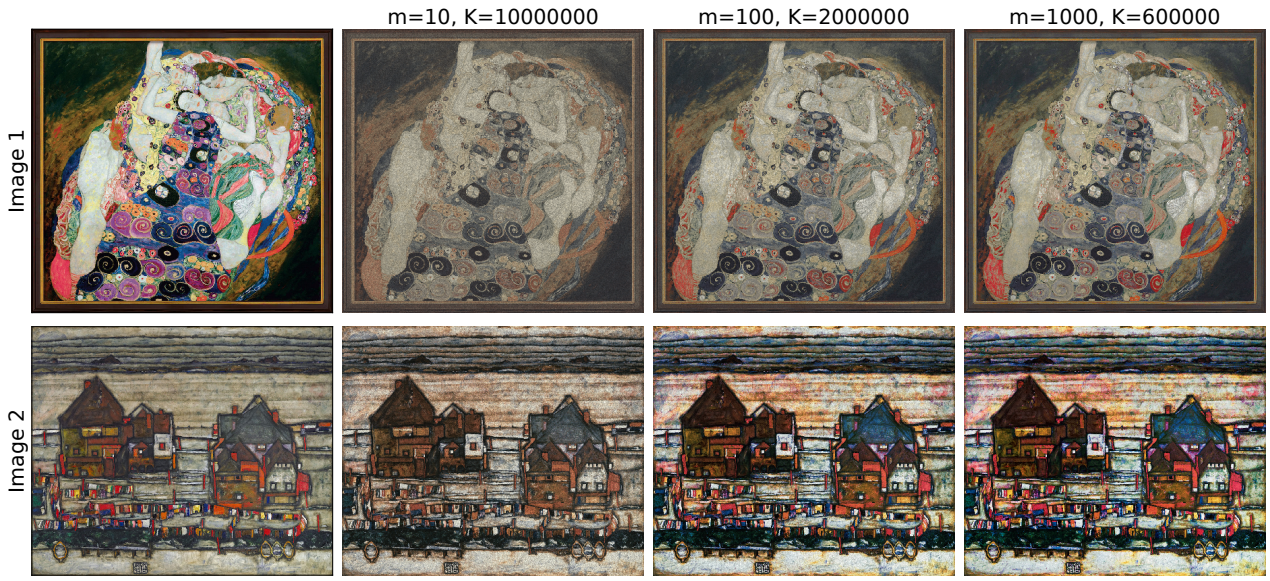

Figure 5: Color transfert between full images for different batch size and number of batches. (Top) color transfert from image 1 to image 2. (Bottom) color transfert from image 2 to image 1.

images. About the required memory for experiments, the memory cost to store data is $O(n)$. The minibatch OT calculus requires $O(m^2)$ because we need to store the ground cost and the OT plan. The marginal experiment requires $O(n)$, as we just need to average the marginals of the plan. Finally, the memory cost is $O(n)$ while OT is $O(n^2)$.

The source image has (943000, 3) RGB dimension and the target image has RGB dimension (933314, 3). For this experiments, we used the minibatch Wasserstein distance with squared euclidean ground cost for several m and k.We used batch of size 10, 100 and 1000. We selected k so as to obtain a good visual quality and observed that a smaller k was needed when using large minibatches. Further experiments which show the dependence on k can be found in appendix. Also note that performing MB optimal transport can be done in parallel and can be greatly speed-up on multi-CPU architectures. One can see in 5 the color transfer (in both directions) provided with our method. We can see that the diversity of colors falls when the batch size is too small as the entropic solver would do for a large regularization parameter. However, even for 1M pixels, a batch size of 1000 is enough to keep a good diversity of colors.

We also studied empirically the results of theorem 2, as shown in Figure 6 we recover the $O(k^{-1/2})$ convergence rate on the marginal with a constant depending on the batch size m. Furthermore, we also empirically studied the computational time and showed that our method is not affected by the number of points with a fixed complexity when an algorithm like Sinkhorn still has a $O(n^2)$ complexity. These experiments show that the minibatch Wasserstein losses are well suited for large scale problems where both memory and computational time are issues.

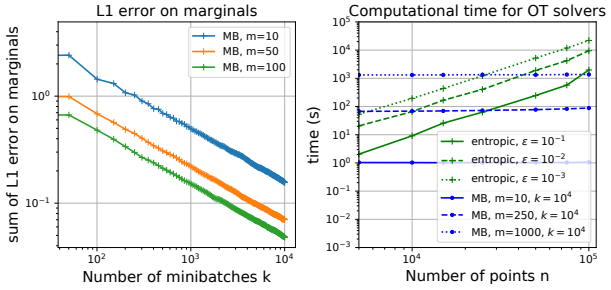

Figure 6: (left) L1 error on both marginals (loglog scale). We selected 1000 points from original images and computed the error on marginals for several m and k (loglog scale). (Right) Computation time for several OT solvers for several number of points in the input distributions, the computation time of the cost matrix is included.

5 Conclusion

In this paper, we studied the impact of using a minibatch strategy in order to reduce the Wasserstein distance complexity. We review the basic properties, and studied the asymptotic behavior of our estimator. We showed a deviation bound between our subsampled estimator and the expectation of our estimator. Furthermore, we studied the optimization procedure of our estimator and proved that it enjoys unbiased gradients. Finally, we demonstrated the effect of minibatch strategy with gradient flow experiments, color transfer and GAN experiments. Future works will focus on the geometry of minibatch Wasserstein (for instance on barycenters) and on investigating a debiasing approach similar to the one used for Sinkhorn Divergence.

Acknowledgements

Authors would like to thank Thibault Séjourné and Jean Feydy for fruitful discussions. This work is partially funded through the projects OATMIL ANR-17-CE23-0012 and 3IA Côte d'Azur Investments ANR-19-P3IA-0002 of the French National Research Agency (ANR).

References

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*. 1, 6

[Bassetti et al., 2006] Bassetti, F., Bodini, A., and Regazzini, E. (2006). On minimum kantorovich distance estimators. *Statistics & Probability Letters*, 76. 1

[Bellemare et al., 2017] Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. (2017). The cramer distance as a solution to biased wasserstein gradients. *CoRR*, abs/1705.10743. 3, 6

[Bernton et al., 2017] Bernton, E., Jacob, P., Gerber, M., and Robert, C. (2017). Inference in generative models using the Wasserstein distance. working paper or preprint. 1, 2

[Blondel et al., 2018] Blondel, M., Seguy, V., and Rolet, A. (2018). Smooth and sparse optimal transport. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. 4, 7

[Bonneel et al., 2011] Bonneel, N., van de Panne, M., Paris, S., and Heidrich, W. (2011). Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, New York, NY, USA.

[Bonnotte, 2013] Bonnotte, N. (2013). *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Université de Paris-Sud. 3

[Bunne et al., 2019] Bunne, C., Alvarez-Melis, D., Krause, A., and Jegelka, S. (2019). Learning generative models across incomparable spaces. In *Proceedings of the 36th International Conference on Machine Learning*.

[Clémençcon, 2011] Clémençcon, S. J. (2011). On u-processes and clustering performance. In *Advances in Neural Information Processing Systems*.

[Clémençon et al., 2016] Clémençon, S., Colin, I., and Bellet, A. (2016). Scaling-up empirical risk minimization: Optimization of incomplete u-statistics. *Journal of Machine Learning Research*.

[Clémençon et al., 2008] Clémençon, S., Lugosi, G., Vayatis, N., et al. (2008). Ranking and empirical minimization of u-statistics. *The Annals of Statistics*.

[Clémençon et al., 2013] Clémençon, S., Robbiano, S., and Tressou, J. (2013). Maximal deviations of incomplete u-statistics with applications to empirical risk sampling. In *Proceedings of the 2013 SIAM International Conference on Data Mining*.

[Courty et al., 2017] Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1

[Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*. 1, 2

[Damodaran et al., 2018] Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018). DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation. In *ECCV 2018 - 15th European Conference on Computer Vision*. Springer. 1, 2

[Ferradans et al., 2013] Ferradans, S., Papadakis, N., Rabin, J., Peyré, G., and Aujol, J.-F. (2013). Regularized discrete optimal transport. In *Scale Space and Variational Methods in Computer Vision*. Springer Berlin Heidelberg. 1, 7

[Feydy et al., 2019] Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using sinkhorn divergences. In *Proceedings of Machine Learning Research*. 2, 6, 7

[Flamary and Courty, 2017] Flamary, R. and Courty, N. (2017). Pot python optimal transport library. 6

[Frogner et al., 2015] Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. (2015). Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems 28*.

[Genevay et al., 2019] Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of sinkhorn divergences. In *Proceedings of Machine Learning Research*. 2

[Genevay et al., 2016] Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*. 1

[Genevay et al., 2018] Genevay, A., Peyre, G., and Cuturi, M. (2018). Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. 1, 2, 6

[Gerber and Maggioni, 2017] Gerber, S. and Maggioni, M. (2017). Multiscale strategies for computing optimal transport. *Journal of Machine Learning Research*. 2

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*. 6

[Gretton et al.,] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schlkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13. 2

[Gulrajani et al., 2017] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30*. 6

[Hoeffding, 1963] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*. 5

[Hull, 1994] Hull, J. (1994). Database for handwritten text recognition research. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16.

[Hunter, 2007] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

[J Lee, 2019] J Lee, A. (2019). U-statistics : theory and practice / a. j. lee. *SERBIULA (sistema Librum 2.0)*.

[Kolouri et al., 2016] Kolouri, S., Zou, Y., and Rohde, G. K. (2016). Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2, 4

[LeCun and Cortes, 2010] LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

[Liu et al., 2015] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*. 7

[Liutkus et al., 2019] Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. (2019). Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *Proceedings of the 36th International Conference on Machine Learning*. 2, 4, 7

[Mikołaj Bińkowski, 2018] Mikołaj Bińkowski, Dougal J. Sutherland, M. A. A. G. (2018). Demystifying MMD GANs. *International Conference on Learning Representations*.

[Papa et al., 2015] Papa, G., Clémençon, S., and Bellet, A. (2015). Sgd algorithms based on incomplete u-statistics: Large-scale minimization of empirical risk. In *Advances in Neural Information Processing Systems 28*.

[Paszke et al., 2017] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. 6

[Patrini et al., 2019] Patrini, G., van den Berg, R., Forré, P., Carioni, M., Bhargav, S., Welling, M., Genewein, T., and Nielsen, F. (2019). Sinkhorn autoencoders. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence*.

[Peyré, 2015] Peyré, G. (2015). Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*. 7

[Peyré and Cuturi, 2019] Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*. 1, 2, 4

[Seguy et al., 2018] Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. (2018). Large-scale optimal transport and mapping estimation. In *International Conference on Learning Representations (ICLR)*. 1

[Sommerfeld et al., 2019] Sommerfeld, M., Schrieber, J., Zemel, Y., and Munk, A. (2019). Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*. 1, 2

[Weed and Bach, 2019] Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*. 2

[Wu et al., 2019] Wu, J., Huang, Z., Acharya, D., Li, W., Thoma, J., Paudel, D. P., and Gool, L. V. (2019). Sliced wasserstein generative models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.