Stochastic Optimization for Regularized Wasserstein Estimators

Marin Ballu [1] **Quentin Berthet** [2] **Francis Bach** [3]

Abstract

Optimal transport is a foundational problem in optimization, that allows to compare probability distributions while taking into account geometric aspects. Its optimal objective value, the Wasserstein distance, provides an important loss between distributions that has been used in many applications throughout machine learning and statistics. Recent algorithmic progress on this problem and its regularized versions have made these tools increasingly popular. However, existing techniques require solving an optimization problem to obtain a single gradient of the loss, thus slowing down first-order methods to minimize the sum of losses, that require many such gradient computations. In this work, we introduce an algorithm to solve a regularized version of this problem of Wasserstein estimators, with a time per step which is sublinear in the natural dimensions of the problem. We introduce a dual formulation, and optimize it with stochastic gradient steps that can be computed directly from samples, without solving additional optimization problems at each step. Doing so, the estimation and computation tasks are performed jointly. We show that this algorithm can be extended to other tasks, including estimation of Wasserstein barycenters. We provide theoretical guarantees and illustrate the performance of our algorithm with experiments on synthetic data.

1. Introduction

Optimal transport is one of the foundational problems of optimization (Monge, 1781; Kantorovich, 2006), and an important topic in analysis (Villani, 2008). It asks how one can transport mass with distribution measure μ to another distribution measure ν, with minimal global transport cost. It can also be written with a probabilistic interpretation, known as the Monge-Kantorovich formulation, of finding a joint distribution π in the set $\Pi(\mu, \nu)$ of those with marginals μ and ν, minimizing an expected cost between variables X and Y. The minimum value gives rise to a natural statistical tool to compare distributions, known as the Wasserstein (or earth-mover's) distance,

$$W_c(\mu, \nu) = \mathrm{OT}(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi} \left[c(X, Y)\right].$$

In the case of finitely supported measures, taken with same support size n for ease of notation, such as two empirical measures from samples, it is written as a linear program (on the right). It can be solved by the Hungarian algorithm (Kuhn, 1955), which runs in time $O(n^3)$. While tractable, this is still relatively expensive for extremely large-scale applications in modern machine learning, where one hopes for running times that are linear in the size of the input (here n^2).

Attention to this problem has been recently renewed in machine learning, in particular due to recent advances to efficiently solve an entropic-regularized version (Cuturi, 2013), and its uses in many applications (see e.g. Peyré et al., 2019, for a survey), as it allows to capture the geometric aspects of the data. This problem has a strongly convex objective, and its solution converges to that of the optimal transport problem when the regularization parameter goes to 0. It can be easily solved with the Sinkhorn algorithm (Sinkhorn, 1964; Altschuler et al., 2017), or by other methods in time $O(n^2 \log n)$ (Dvurechensky et al., 2018).

These tools have been applied in a wide variety of fields, from machine learning (Alvarez-Melis et al., 2018; Arjovsky et al., 2017; Gordaliza et al., 2019; Flamary et al., 2018), natural language processing (Grave et al., 2019; Alaux et al., 2018; Alvarez-Melis et al., 2018), computer graphics (Feydy et al., 2017; Lavenant et al., 2018; Solomon et al., 2015), the natural sciences (del Barrio et al., 2019; Schiebinger et al., 2019), and learning under privacy (Boursier & Perchet, 2019).

Of particular interests to statistics and machine learning are analyses of this problem with only sample access to the distributions. There have been growing efforts to estimate

[1]Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, United Kingdom [2]Google Research, Brain Team, Paris, France [3]Inria, ENS, PSL Research University, Paris, France. Correspondence to: Marin Ballu <mb2193@cam.ac.uk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

either the objective value of this problem, or the unknown distribution, with this metric or associated regularized metrics (see below) (Weed et al., 2019; Genevay et al., 2019; Uppal et al., 2019). One of the motivations are variational Wasserstein problems, where the objective value of an optimal transport problem is used as a loss, and one seeks to minimize in a parameter θ an objective that depends on a known distribution ν_θ

$$\min_{\theta\in\Theta} \mathrm{OT}(\mu, \nu_\theta)\,,$$

where μ is only accessible through samples. This method for estimation, referred to as *minimum Kantorovich estimators* (Bassetti et al., 2006), mirrors the interpretation of likelihood maximization as the minimization of $\mathrm{KL}(\nu_\theta, \mu)$, with the Kullback-Leibler divergence.

The value of the entropic-regularized problem, or of the related *Sinkhorn divergence*, can also be used as a loss in learning tasks (Alvarez-Melis et al., 2018; Genevay et al., 2017; Luise et al., 2018), and compared to other metrics such as maximum mean discrepency (Gretton et al., 2012; Feydy et al., 2019; Arbel et al., 2019). One of the advantages of the regularized problem is the existence of gradients in the parameters of the problem (cost matrix, target measures).

The problem of minimizing this loss for the ℓ_2 cost over $\mathbb{R}^d$ has been shown to be equivalent to maximum likelihood Gaussian deconvolution (Rigollet & Weed, 2018). We show here that this result can be generalized for all cost functions to maximum likelihood estimation for a kernel inversion problem. It is not only the solution of a stochastic optimization problem, but also an estimator, referred to here as the *regularized Wasserstein estimator*.

In this work, we propose a new stochastic optimization scheme to minimize the OT_ε between an unknown discrete measure μ and another discrete measure $\nu \in \mathcal{M}$, with an additional regularization term on ν. There are many connections between this problem and stochastic optimization: by a dual formulation, the value $\mathrm{OT}_\varepsilon(\mu, \nu)$ can be written as the optimum of an expectation in μ, ν, allowing simple computations with only sample access (Genevay et al., 2016). Here, we take this one step further and design an algorithm to *optimize* in ν, not just evaluate this loss. A direct approach is to optimize by first-order methods, by the use of stochastic gradients in ν at each step (Genevay et al., 2017). However, these gradient estimates are based on dual solutions of the regularized problem, so obtaining them requires to solve an optimization problem, with running time scaling quadratically in the intrinsic dimension of the problem (the size of the supports of μ, ν). For the dual formulation that we introduce, stochastic gradients can be directly computed from samples. Algorithmic techniques exploiting the particular structure of the dual formulation for this regularization allow us to compute these gradients in constant time. We follow here the recent developments in *sublinear algorithms* based on stochastic methods (Clarkson et al., 2012).

We provide theoretical guarantees on the convergence of the final iterate ν_t to the true minimizer ν^*, and demonstrate these results on simulated experiments.

2. Problem Description

Definitions. Let μ be a probability measure on $\mathbb{R}^d$ with finite support $\mathcal{X} = \{x_i\}_{1\le i\le I} \subset \mathbb{R}^d$ and a family $\mathcal{M}$ of probability measures. The measures in $\mathcal{M}$ should all be absolutely continuous with respect to a known measure β supported in the finite set $\mathcal{Y} = \{y_j\}_{1\le j\le J} \subset \mathbb{R}^d$. We consider the following minimization problem:

$$\min_{\nu\in\mathcal{M}} \mathrm{OT}_\varepsilon(\mu, \nu) + \eta\,\mathrm{KL}(\nu, \beta). \tag{1}$$

In this expression, OT_ε is the regularised optimal transport cost defined by the following expression

$$\mathrm{OT}_\varepsilon(\mu, \nu) = \min_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\left[c(X, Y)\right] + \varepsilon\,\mathrm{KL}(\pi, \mu\otimes\nu), \tag{2}$$

where the minimum is taken over the set

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) : \pi_{\mathcal{X}} = \mu,\ \pi_{\mathcal{Y}} = \nu\}$$

of couplings of μ and ν, and c is a cost function in $\mathbb{R}^d$. The operator $\mathrm{KL}(\cdot,\cdot)$ is the Kullback-Leibler divergence, defined as

$$\mathrm{KL}(\mu_1, \mu_2) = \mathbb{E}_{Z\sim\mu_2}\left[\frac{d\mu_1}{d\mu_2}(Z)\log\left(\frac{d\mu_1}{d\mu_2}(Z)\right)\right],$$

for two measures μ_1 and μ_2 such that $\mu_1 \ll \mu_2$. We assume that $\mathcal{M}$ is convex for the problem to be a convex optimization problem, and compact to guarantee that the minimum is attained. We consider $\eta \ge \varepsilon$ to guarantee convexity (see Proposition 3.2).

Remark 1. If c is a distance and if $\varepsilon = \eta = 0$, then OT_ε is a Wasserstein distance and our problem can be seen as computing a projection of μ onto $\mathcal{M}$. In the discrete case, the solution to the unregularized problem is the distribution ν such that $\nu(y) = \mu(x)$, where y is the nearest neighbour in $\mathcal{Y}$ of x.

Remark 2. In (2), the addition of entropic regularization smoothes the transport plan π. In (1), the Kullback-Leibler divergence plays the role of a second regularization that smoothes the estimated measure ν. The effects of these two layers of regularization are further discussed in section 5.2.

Learning problem. Our objective is to solve the optimization problem in Equation (1), given observations X_i independent and identically distributed (i.i.d.) from μ that is

unknown, and sample access to β. These can be assumed to be simulated by the user if β is known, as part of the regularization. This problem can be either be interpreted as an unsupervised learning problem or as estimation in an inverse problem, and we refer to it as *regularized Wasserstein estimation*. The term in Kullback-Leibler (or entropy, up to an offset) are classical manners in which a probability can be regularized.

Maximum likelihood interpretation. While the unregularized problem has a trivial solution, there is in general no closed form for positive ε. When $\varepsilon > 0$, $\eta = 0$ and $\mathcal{M}$ is the set of all probability measures on Y, then our problem is equivalent to the maximum likelihood estimator for a kernel inversion problem. This corresponds to estimating the unknown initial distribution of a random variable Y, but only by observing it *after* the action of a specific transition kernel κ (see, e.g., Berthet & Kanade, 2019, for the statistical complexity of estimating initial dustributions under general Markov kernels).

Proposition 2.1 (MLE interpretation)**.** *Let $\mathcal{M}$ be the set of all probability measures on Y, let ν^* be a measure on Y, and let $\kappa : Y \to X$ be a transition kernel of the form*

$$\kappa(x, y) = \frac{\exp\left(-\frac{c(x,y)}{\varepsilon}\right)}{\sum_{x' \in X} \exp\left(-\frac{c(x',y)}{\varepsilon}\right)},$$

the observed measure is $\mu = \kappa\nu^*$, which can be written as

$$\mu(x) = \int_Y \kappa(x, y) d\nu^*(y).$$

The maximum likelihood estimation of ν^* for this observation is

$$\hat{\nu} := \arg\max_{\nu \in \mathcal{M}} \sum_i \log(\kappa\nu)(X_i).$$

This estimator also verifies

$$\hat{\nu} = \arg\min_{\nu \in \mathcal{M}} \mathrm{OT}_\varepsilon(\mu, \nu). \tag{3}$$

Remark 3. If $c(x, y) = \|x-y\|^2$, then $\kappa(x, y) =: \phi_\varepsilon(x-y)$ is a Gaussian convolution kernel with ϕ_ε being a centered gaussian distribution with covariance matrix $\frac{\varepsilon}{2}\mathrm{Id}$. The sample measure $\mu = \phi_\varepsilon \star \nu^*$ is a convolution, so the solution of (3) is the MLE of the Gaussian deconvolution problem, as already presented by Rigollet & Weed (2018).

As in the Gaussian case, these optimization problems share an optimum, but are not equal in value. Therefore, in our regularized setting, it is not possible to substitute one for the other.

Gaussian case. To illustrate the effect of each regularization term, we consider here the case where $c(x, y) = \|x - y\|^2$, $\mathcal{M}$ is the set of one dimensional Gaussian distributions and the target measure $\mu \sim \mathcal{N}(m_\mu, \sigma_\mu^2)$ as well as the prior measure $\beta \sim \mathcal{N}(0, 1)$ are Gaussian distributions. The multivariate case has a closed form, as showed by (Janati et al., 2020). We present this closed form of the objective in the 1-D case:

$$\begin{aligned}
&\mathrm{OT}_\varepsilon(\mu, \nu) + \eta\,\mathrm{KL}(\nu, \beta) = \\
&|m_\nu - m_\mu|^2 + \frac{\eta}{2}|m_\nu|^2 + \sigma_\mu^2 + \left(1 + \frac{\eta}{2}\right)\sigma_\nu^2 \\
&- \sqrt{4\sigma_\mu^2\sigma_\nu^2 + \frac{\varepsilon^2}{4}} - \frac{\eta}{2}\log\sigma_\nu^2 \\
&+ \frac{\varepsilon}{2}\log\left(\varepsilon + \sqrt{4\sigma_\mu^2\sigma_\nu^2 + \frac{\varepsilon^2}{4}}\right),
\end{aligned} \tag{4}$$

where m_ν and σ_ν are the mean and variance of the gaussian variable ν. The estimator $\hat{\nu}$ that minimizes (4) over the set of gaussian distributions verifies:

$$m_{\hat{\nu}} = \frac{m_\mu}{1 + \frac{\eta}{2}},$$

so we see that the regularization term in η centers the Wasserstein estimator. If $\eta = 0$ then

$$\sigma_{\hat{\nu}}^2 = \sigma_\mu^2 - \frac{\varepsilon^2}{4} \tag{5}$$

since the estimator is a gaussian deconvolution (see Remark 3). However in the general case $\eta > 0$, the variance $\sigma_{\hat{\nu}}^2$ does not have a closed form, but we can compare the effects of each regularization by looking at their asymptotic behaviour. Indeed, when ε tends to 0, we have

$$\sigma_{\hat{\nu}} = \frac{\sigma_\mu + \sqrt{\sigma_\mu^2 + \left(1 + \frac{\eta}{2}\right)(2\eta - \varepsilon)}}{2 + \eta} + O(\varepsilon^2),$$

where the bounds on $O(\varepsilon)$ only depend on σ_μ. This expression shows that a larger η will spread the distribution $\hat{\nu}$ if and only if the variance σ_μ^2 is smaller than 1. It also suggests that that a larger ε will reduce the variance of the estimator, which is already seen in equation (5). We will see in Proposition 3.2 that it is preferable that η is chosen greater than ε to guarantee convexity. We remark that if $\eta \geq \varepsilon$ and η tends to 0, then

$$\sigma_{\hat{\nu}}^2 = (1 - \eta)\sigma_\mu^2 + \eta + O(\eta^2),$$

so the variance of $\hat{\nu}$ can be approximated by the average of the variance of μ and the variance of the prior β. The effects of the regularization on the estimator are also further discussed in Section 5.2.

3. Dual formulations

As noted above, first-order optimization methods to solve directly in ν the regularized problem require at every step to

solve an optimization problem. We explore instead another approach, through a dual formulation of our problem. Such a formulation allows to change the minimization problem in (2) into a maximization problem.

Proposition 3.1 (Dual formulation)**.** *If $\varepsilon > 0$, then the problem (1) is equivalent to the following problem:*

$$\min_{f\in\mathcal{F}} \max_{a\in L^1(\mu), b\in L^1(\nu)} \mathbb{E}\left[a(X) + b(Y)f(Y) - \varepsilon \exp\left(\frac{a(X)+b(Y)-c(X,Y)}{\varepsilon}\right) + (\eta-\varepsilon)f(Y)\log f(Y)\right], \tag{6}$$

the expectation being over the variables $(X,Y) \sim \mu\otimes\beta$, with $f(y) = \frac{d\nu}{d\beta}(y)$ and $\mathcal{F} = \{\frac{d\nu}{d\beta} \,:\, \nu\in\mathcal{M}\}$.

If f is constant β-almost everywhere, with value 1, then the maximization problem for a and b in (6) is the dual of the regularized optimal transport problem 2, for which a block coordinate descent corresponds to Sinkhorn algorithm (Cuturi, 2013).

This dual formulation is a saddle point problem, and it is convex-concave if $\eta \geq \varepsilon$, so the Von Neumann minimax theorem applies: we can swap the minimum and the maximum.

Proposition 3.2. *If $\eta \geq \varepsilon > 0$ then the problem (1) is equivalent to the following maximization problem:*

$$\max_{a\in L^1(\mu), b\in L^1(\nu)} F(a,b), \tag{7}$$

with

$$F(a,b) = \mathbb{E}\left[a(X) - \varepsilon e^{\frac{a(X)+b(Y)-c(X,Y)}{\varepsilon}}\right] - (\eta-\varepsilon)H^*_\beta\left(-\frac{b}{\eta-\varepsilon}\right), \tag{8}$$

by writing

$$H^*_\beta(\alpha) = \max_{f\in\mathcal{F}} \mathbb{E}\left[\alpha(Y)f(Y) - f(Y)\log f(Y)\right],$$

with the variables $(X,Y)\sim\mu\otimes\beta$.

In its discrete formulation, the problem is written with the following notations: $C_{i,j} := c(x_i, y_j)$ for the cost matrix, $a_i = a(x_i)$ and $b_j = b(y_j)$ for the dual vectors, and $f_j = f(y_j)$ for the remaining primal variable.

The problem (7) is hence given by

$$\max_{(a,b)\in\mathbb{R}^I\times\mathbb{R}^J} F(a,b), \tag{9}$$

with

$$F(a,b) = \mathbb{E}\left[a_i - \varepsilon\exp\left(\frac{a_i+b_j-C_{i,j}}{\varepsilon}\right)\right] - (\eta-\varepsilon)H^*_{\beta,\mathcal{M}}\left(-\frac{b}{\eta-\varepsilon}\right). \tag{10}$$

The indices (i,j) are here independent random variables such that $x_i\sim\mu$ and $y_j\sim\beta$. The function $H^*_{\beta,\mathcal{M}}$ is the Legendre transform of the relative entropy to β on the set $\mathcal{F}$:

$$H^*_{\beta,\mathcal{M}}(\alpha) = \max_{f\in\mathcal{F}} \mathbb{E}\left[f_j(\alpha_j - \log f_j)\right], \tag{11}$$

with j a random index such that $y_j\sim\beta$.

If the maximum is attained on the relative interior of $\mathcal{M}$ at the point $\nu^*(\alpha)$, then we have $\nabla H^*_{\beta,\mathcal{M}}(\alpha) = \nu^*(\alpha)$. Moreover the optimum $\nu^*(-b^*/(\eta-\varepsilon))$ for the dual problem (6) is the optimal $\nu\in\mathcal{M}$ for our general problem (1).

Proposition 3.3. *The function F has the following properties.*

1. *The set of solutions to the problem (9) is a nonempty affine space spanned by the vector $((1,\dots,1),(-1,\dots,-1))$.*
2. *Every solution (a^*,b^*) of (9) verifies*

$$\forall\, i,j,\ |a^*_i + b^*_j - C_{i,j}| \leq B, \tag{12}$$

with $B := \varepsilon m + 2R_C$, where R_C is the range of the matrix C given by $R_C := \max_{i,j} C_{i,j} - \min_{i,j} C_{i,j}$, and $m := \max_j |\log f_j|$ with $f_j = \nu^*_j/\beta_j$.

3. *The function $-F$ is λ-strongly convex on the slice $\{\sum_i \mu_i a_i = \sum_j \beta_j b_j\}$ with*

$$\lambda := \frac{\min_{i,j}\{\mu_i,\beta_j\}}{\varepsilon} e^{-(m+2R_C/\varepsilon)}.$$

4. *For i and j independent random variables as for (10), we have the gradients of F are written as simple expectations*

$$\nabla_a F = \mathbb{E}\left[(1-D_{i,j})e_i\right], \tag{13}$$
$$\nabla_b F = \mathbb{E}\left[(f_j - D_{i,j})e'_j\right], \tag{14}$$

with $D_{i,j}(a,b) = \exp\left(\frac{a_i+b_j-C_{i,j}}{\varepsilon}\right)$, $(e_i)_{1\leq i\leq I}$ and $(e'_j)_{1\leq j\leq J}$ the canonical basis in $\mathbb{R}^I$ and $\mathbb{R}^J$ respectively.

4. Stochastic Optimization Methods

The formulas (13) and (14) suggest that our problem can be solved using a stochastic optimization approach. For random indices i drawn from μ and j drawn from β, we obtain the following stochastic gradients

$$G_a = (1-D_{i,j})e_i = \left(1-\exp\left(\frac{a_i+b_j-C_{i,j}}{\varepsilon}\right)\right)e_i$$
$$G_b = (f_j - D_{i,j})e'_j = \left(\frac{\nu^*_j}{\beta_j} - \exp\left(\frac{a_i+b_j-C_{i,j}}{\varepsilon}\right)\right)e'_j.$$

By Proposition 3.3, these are unbiased estimates of the gradients of F. The algorithm then proceeds with an averaged gradient ascent that uses these stochastic gradients updates at each step. The obtained iterates $(b^t)_{t\geq 1}$ are averaged, producing the sequence $\left(\overline{b^t}\right)_{t\geq 0}$ of iterates defined by

$$\overline{b^t} := \frac{1}{t}\sum_{1\leq t'\leq t} b^{t'}.$$

The computation of G_a can be done in $O(1)$, however G_b necessitates the value ν_j^* in (11) to be computed. The complexity of this computation depends on the set $\mathcal{M}$, and we will present here two cases where it can be done with low complexity.

Initialization. To guarantee that the gradients will not get exponentially big, we choose the initial value of the dual variables so that it verifies

$$\forall i, j,\ a_i + b_j - C_{i,j} \leq -\varepsilon m,$$

with m being defined in (12). We define

$$\text{ini}(C, \varepsilon, m) := (\min C_{i,j} - \varepsilon m)/2,$$

and we initialize

$$a_i = b_j = \overline{b_j} = \text{ini}(C, \varepsilon, m). \tag{15}$$

Usually, m is unknown and should be determined by heuristics.

Simple case. We analyze the case where $\mathcal{M}$ is the family of all probability measures supported in the finite set $\{y_j\}_{1\leq j\leq J} \subset \mathbb{R}^d$, with the assumption that $\eta > \varepsilon$. Then, if the max is attained on the interior of the simplex, we have the optimum

$$\nu_j^* = \frac{\beta_j e^{-b_j/(\eta-\varepsilon)}}{\sum_k \beta_k e^{-b_k/(\eta-\varepsilon)}}. \tag{16}$$

The algorithm needs $O(1)$ complexity for each time step. If the values of $C_{i,j}$ are accessible without having the whole matrix stored (such as a simple function of x_i and y_j), the storage is only $O(I + J)$ in this algorithm, because we do not need to store any $D_{i,j}$. The complexity at each step of the algorithm is better than with the non regularized form, where j is taken as $\arg\max_j \beta_j e^{-b_j/(\eta-\varepsilon)}$, instead of randomly. This enhancement in complexity mostly comes from the storage of the sum $S^t = \sum_j g_j(b_j^t)$ with

$$g_j(b_j^t) := \beta_j e^{-b_j^t/(\eta-\varepsilon)}.$$

Indeed, instead of computing the entire sum at each iterates, which costs $O(J)$ operations, the algorithm simply updates the part of the sum that was modified:

$$S^{t+1} = S^t + g_j(b_j^{t+1}) - g_j(b_j^t).$$

Algorithm 1 SGD for Wasserstein estimator

The entries are the learning rates (γ_t), the probabilities $\mu = (\mu_i)_i$, $\beta = (\beta_j)_j$, the cost matrix $C_{i,j}$ and the logarithmic gap m between the solution and the prior.
Initialize $a_i = b_j = \overline{b_j} = \text{ini}(C, \varepsilon, m)$, $S = e^{-\frac{\text{ini}(C,\varepsilon,m)}{\eta-\varepsilon}}$..

for $t = 1$ to T **do**
  Sample $i \in \{1, \dots, I\}$ with probability μ_i.
  Sample $j \in \{1, \dots, J\}$ with probability β_j.
  $D_{i,j} = e^{\frac{a_i + b_j - C_{i,j}}{\varepsilon}}$.
  $f_j = e^{-b_j/(\eta-\varepsilon)}/S$.
  $a_i \leftarrow a_i + \gamma_t(1 - D_{i,j})$.
  $b_j \leftarrow b_j + \gamma_t(f_j - D_{i,j})$ with the previous as b'_j.
  $\overline{b_j} \leftarrow \left(1 - \frac{1}{t}\right)\overline{b_j} + \frac{1}{t}b_j$
  $S \leftarrow S + \beta_j e^{-b_j/(\eta-\varepsilon)} - \beta_j e^{-b'_j/(\eta-\varepsilon)}$
end for
for $j = 1$ to J **do**
  $\nu_j = \beta_j e^{-\overline{b_j}/(\eta-\varepsilon)} / \sum_{j'} \beta_{j'} e^{-\overline{b_{j'}}/(\eta-\varepsilon)}$
end for
Return ν.

This method assures updates in $O(1)$. In a context focused entirely on optimization, where μ and β are known in advance, we could also pick i and j uniformly, and add μ_i and β_j as factors in the formulas. This would not reduce the complexity.

Mixture models. We also consider a set of measures $(\nu^k)_{1\leq k\leq K}$ supported in supported in the set $\{y_j\}_{1\leq j\leq J} \subset \mathbb{R}^d$, and take $\mathcal{M} = \{\sum_k \theta_k \nu^k : \theta \in \Delta_K\}$ to be their convex hull. We define the matrix $M = (\nu^k(y_j))_{j,k}$. Then $\mathcal{M} = \{M\theta : \theta \in \Delta_K\}$, and Equation (11) becomes

$$H^*_{\beta,\mathcal{M}}(\alpha) = \max_{\theta\in\Delta_K} (\alpha - \log(M\theta) + \log(\beta))^T M\theta, \tag{17}$$

with the log being taken component-wise.

Proposition 4.1. *The maximization problem* (17) *has a solution*

$$\theta^* = \frac{M^\dagger \exp\left(P_{Im(M)}(-b/(\eta-\varepsilon) - 1 - \log(\beta))\right)}{1^T M^\dagger \exp\left(P_{Im(M)}(-b/(\eta-\varepsilon) - 1 - \log(\beta))\right)},$$

with $M^\dagger$ being the Moore-Penrose inverse of the matrix M. It gives the measure

$$\nu^* = \frac{\exp\left(P_{Im(M)}(-b/(\eta-\varepsilon) - 1 - \log(\beta)\right)}{1^T \exp\left(P_{Im(M)}(-b/(\eta-\varepsilon) - 1 - \log(\beta)\right)}.$$

We can replace it in equation (14) to get the stochastic gradients. However at each new computed step, every coefficient changes, and there is a need to do J computations for each step. The solution computed here is also valid for the case when it is not unique.

We can, however, consider another regularization to the entropy of θ to improve the algorithm. The problem is the following:

$$\min_{\theta\in\Delta_K} \mathrm{OT}_\varepsilon(\nu,\mu) + \eta\,\mathrm{KL}(\theta, M^\dagger\beta).$$

The other computations are unchanged, apart from Equation (11), replaced by

$$\begin{aligned} H^*_{\beta,\mathcal{M}}(\alpha) &= \max_{\theta\in\Delta_K} \alpha^T M\theta - (\eta-\varepsilon)\,\mathrm{KL}(\theta, M^\dagger\beta) \\ &= \max_{\theta\in\Delta_K} (M^T\alpha - \log(\theta) + \log(M^\dagger\beta))^T\theta. \end{aligned} \tag{18}$$

Proposition 4.2. *The maximization problem* (18) *has a solution*

$$\theta^* = \frac{\exp\left(M^T(-b/(\eta-\varepsilon)-1) + \log(M^\dagger\beta)\right)}{1^T\exp\left(M^T(-b/(\eta-\varepsilon)-1) + \log(M^\dagger\beta)\right)}.$$

Both regularizations $\mathrm{KL}(\theta, M^\dagger\beta)$ and $\mathrm{KL}(\nu,\beta)$ are minimal when $\nu=\beta$, and can therefore be used as a suitable proxy. The solution to the regularized problem is similar to the solution to the unregularized one. For this modified problem, the computations are accessible, and they can be done in time $O(K)$, a great improvement if $K \ll J$. We apply the stochastic gradient scheme in Algorithm 2.

Wasserstein barycenters. Algorithm 1 can be used to compute an approximation of the Wasserstein barycenter of K measures $\mu^1,\dots,\mu^K$. If the cost funtion in the optimal transport problem is of the form $c(x,y) = d(x,y)^p$ with d being a distance and $p\geq 1$, then the transport cost $\mathrm{OT}(\cdot,\cdot)$ defines the p-Wasserstein distance. In these conditions, the Wasserstein barycenter of the measures $\mu_1,\dots,\mu_K$ with nonnegative weights $w_1,\dots,w_K$ is the solution of the minimization problem

$$\min_\nu \sum_{k=1}^K w_k\,\mathrm{OT}(\mu^k,\nu). \tag{19}$$

This optimization and the barycenter that it defines was introduced by Agueh & Carlier (2011), these objects and their regularized versions have attracted a lot of attention, for their statistical and algorithmic aspects (Zemel et al., 2019; Cuturi & Doucet, 2014; Claici et al., 2018; Luise et al., 2019).

As an analogy with our original problem (1), we consider an entropic regularization of the Wasserstein barycenter problem (19):

$$\min_{\nu\in\mathcal{M}} \sum_{k=1}^K w_k\,\mathrm{OT}_\varepsilon(\mu^k,\nu) + \eta\,\mathrm{KL}(\nu,\beta).$$

Algorithm 2 SGD for Wasserstein projection

The entries are the learning rates (γ_t), the probabilities $\mu = (\mu_i)_i$, $\beta = (\beta_j)_j$, the stochastic matrix $M = (\nu_j^k)_{j,k}$, the cost matrix $C_{i,j}$ and the logarithmic gap m between the solution and the prior.
Initialize a_i, b_j, $\overline{b_j}$, $\alpha = \log\left(M^\dagger\beta\right)$, $\theta_k = 1/K$.
for $t = 1$ to T **do**
  Sample $i\in\{1,\dots,I\}$ with probability μ_i.
  Sample $j\in\{1,\dots,J\}$ with probability β_j.
  $D_{i,j} = e^{\frac{a_i+b_j-C_{i,j}}{\varepsilon}}$.
  $f_j = \sum_{k=1}^K \theta_k\nu_j^k/\beta_j$.
  $a_i \leftarrow a_i + \gamma_t(1-D_{i,j})$.
  $b_j \leftarrow b_j + \gamma_t(f_j - D_{i,j})$.
  for $k = 1$ to K **do**
    $\alpha_k \leftarrow \alpha_k - \frac{\gamma_t}{\eta-\varepsilon}\nu_j^k(f_j - D_{i,j})$.
    $\overline{\alpha_k} \leftarrow \left(1-\frac{1}{t}\right)\overline{\alpha_k} + \frac{1}{t}\alpha_k$
  end for
  for $k = 1$ to K **do**
    $\theta_k = e^{\alpha_k}/\sum_{k'} e^{\alpha_{k'}}$.
  end for
end for
for $k = 1$ to K **do**
  $\theta_k = e^{\overline{\alpha_k}}/\sum_{k'} e^{\overline{\alpha_{k'}}}$.
end for
for $j = 1$ to J **do**
  $\nu_j = \sum_{k=1}^K \theta_k\nu_j^k$.
end for
Return ν.

Our approach can be translated to this setting, as well as the theoretical results found for (1). We have the equivalent dual formulation

$$\max_{a\in L^1(\mu), b\in L^1(\nu)} \tilde{F}(a^1,\dots,a^K,b),$$

with

$$\tilde{F}(a^1,\dots,a^K,b) := \sum_{k=1}^K w_k F_k(a^k,b).$$

Here F_k is defined like the function F in (8) by replacing μ by μ^k. The only difference in the algorithm is that there should be K dual variables $a^1,\dots,a^K$ that play the role of the variable a for each measure μ^k while one variable b is used to obtain the target measure.

The complexity of the algorithm is $O(K)$ for each stochastic gradient step, which gains a factor $\log K$ compared to the state-of-the-art stochastic Wasserstein barycenter (Staib et al., 2017), that solves the unregularized minimisation problem (19). The complexity of a gradient step could be further reduced to $O(1)$ at the cost of more randomization, by sampling k randomly at each step with probability proportional to w_k, and updating a_k and b as in algorithm 1 with μ_k playing the role of μ.

If $\eta \approx \varepsilon$, the approximation error of this estimated Wasserstein Barycenter is of the same order as that of the problem

$$\min_{\nu \in \mathcal{M}} \sum_{k=1}^{K} w_k \, \mathrm{OT}_\varepsilon(\mu^k, \nu),$$

considered in Cuturi & Peyré (2015), that solves the problem with a single layer of regularization.

5. Results

5.1. Convergence bounds

The following convergence bounds are valid for both algorithms presented in the previous section. They come from general convergence bounds for averaged stochastic gradient descent with decreasing stepsize (Shamir & Zhang, 2012). For $\nu^* \in \mathcal{M}$ be the optimal Wasserstein estimator, let $\nu^t \in \mathcal{M}$ be the estimator obtained by stopping the algorithm at step t. Since the measures in $\mathcal{M}$ are all supported on $\mathcal{Y}$, we consider the Kullback-Leibler divergence to express how close the estimated measure ν^t is to ν^*. As ν^t is obtained with the dual variable b^t, the estimation error of b^t can translate to an entropic error in the following two bounds. The first result uses the stepsize for SGD associated to strongly convex functions and the second one uses the stepsize for SGD associated to convex functions. Both results are presented here: even though the theoretical bound of the second one is asymptotically worse, its stepsize can yield better performance in practice.

Theorem 5.1. *With stepsize $\gamma_t = \frac{1}{\lambda t}$, the estimator verifies the following bound:*

$$\mathbb{E}\left[\mathrm{KL}(\nu^*, \nu^t)\right] \leq 34 \frac{e^{2m}}{(\eta - \varepsilon)\lambda^2} \frac{1 + \log t}{t}.$$

Theorem 5.2. *With stepsize $\gamma_t = \frac{c_0 \varepsilon}{\sqrt{t}}$, with any chosen constant $c_0 \leq B e^{-m}/\varepsilon$, the estimator verifies the following bound:*

$$\mathbb{E}\left[\mathrm{KL}(\nu^*, \nu^t)\right] \leq 2 \frac{B^2 e^m}{c_0 \varepsilon (\eta - \varepsilon)\lambda} \frac{2 + \log t}{\sqrt{t}}.$$

In order to prove both theorems, we present two lemmas whose proofs are provided in the appendix.

Lemma 5.3. *Let a^t, b^t be the iterations of the stochastic gradient descent, seen as random variables. If the initialization is done as in (15), then the second order moments of the stochastic gradients are bounded:*

$$\mathbb{E}\left[\|\nabla_a F_{i,j}(a^t, b^t)\|^2 + \|\nabla_b F_{i,j}(a^t, b^t)\|^2\right] \leq 2e^{2m}.$$

Lemma 5.4. *The convergence of the primal variable $\nu(b)$ is linked to the convergence of the objective by the following bound:*

$$KL(\nu(b^*), \nu(b)) \leq \frac{F(a^*, b^*) - F(a, b)}{(\eta - \varepsilon)\lambda}.$$

Proof of Theorem 5.1. The result from Shamir & Zhang (2012) on strongly convex functions gives the bound

$$\mathbb{E}\left[F(a^*, b^*) - F(a^t, b^t)\right] \leq 17 \frac{G^2}{\lambda} \frac{1 + \log t}{t},$$

with G^2 being a bound on the second order moments of the stochastic gradients. The lemma 5.3 provides $G^2 = 2e^{2m}$. We conclude with lemma 5.4. □

Proof of theorem 5.2. With stepsize $\gamma_t = \frac{B}{G\sqrt{t}}$, the result from Shamir & Zhang (2012) on convex functions gives the bound

$$\mathbb{E}\left[F(a^*, b^*) - F(a^t, b^t)\right] \leq 2(BG) \frac{2 + \log t}{\sqrt{t}},$$

with G^2 being a bound on the second order moments of the stochastic gradients. The lemma 5.3 provides $G \geq \sqrt{2} e^m$, here we choose $G = \frac{B}{c_0 \varepsilon}$ where we assume $c_0 \leq B e^{-m}/\varepsilon$. We conclude with Lemma 5.4. □

Remark 4. The term in $\log t$ can be removed by using adaptive averaging schemes: by averaging only the past αt iterates, the the term $1 + \log t$ can be replaced by $\frac{1 - \log(1-\alpha)}{\alpha}$.

Remark 5. The strong convexity coefficient

$$\lambda = \frac{\min_{i,j}\{\mu_i, \beta_j\}}{\varepsilon} e^{-B/\varepsilon}$$

is negligible when $\varepsilon \ll B$, thus the stepsize of the first theorem is large: it can lead the dual variables to grow out of their normal range and produces an exponential overflow in experiments. One solution is to cap the dual variables to the range provided by (12), but the algorithm would then not provide any useful solution until a high number of steps is performed, i.e. $t \gtrapprox 1/B\lambda$. Instead, we recommend using the stepsize $\gamma_t = \min\{1/\lambda t, c_0 \varepsilon/\sqrt{t}\}$ that provides a quick convergence at the earlier steps, then gives a better asymptotic convergence rate.

5.2. Simulations

We demonstrate the performance of the algorithm on simulated experiments.

Regularization term. In order to exhibit clearly the impact of regularization parameters, We analyze a simple case, where $\mathcal{X} = \mathcal{Y}$, and $C_{i,j} = |i - j|$. In this case the solution is given by $\nu^* = \mu$ for $\varepsilon = \eta = 0$, with a diagonal transportation matrix. The prior β is chosen as the uniform measure on $\mathcal{Y}$, and we use the learning rate provided by Theorem 5.2.

The regularization coefficient η should be greater than ε to guarantee convexity, and has a smoothing effect on the solution. Indeed, the solution converges to β when η tends

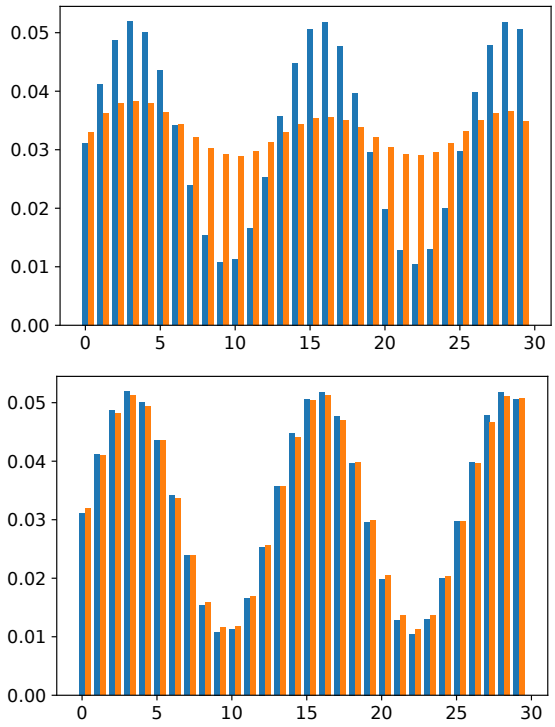

Figure 1. Effect of the regularization on the target measure. Target measure μ in blue, estimator in orange. *Upper plot:* $\varepsilon = \eta - \varepsilon = 0.1$. *Lower plot:* $\varepsilon = \eta - \varepsilon = 0.01$.

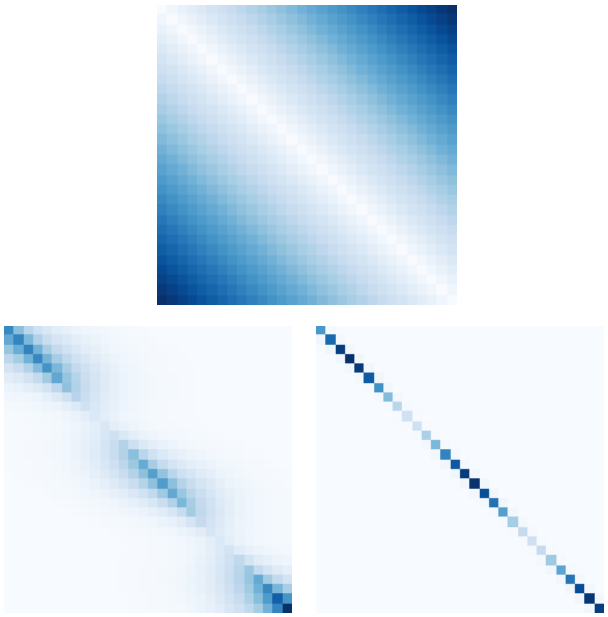

Figure 2. Effect of the regularization on the transport matrix. *Upper plot:* cost matrix used. *Lower plots, from left to right:* transportation matrix for $\varepsilon = \eta - \varepsilon = 0.1$ after 10^6 iterations, and for $\varepsilon = \eta - \varepsilon = 0.01$.

to infinity, and generally, an estimator that is regularized with a larger η will be closer (in KL divergence) to the reference measure β. For example, if β is chosen as a uniform law on the discrete set $\mathcal{Y}$, the regularized estimator will be more spread out than the unregularized solution, i.e. will have a larger entropy - as in Figure 1. We choose to take $\eta = 2\varepsilon$ to conserve a similar degree of regularization as in the case $\eta = 0$, while guaranteeing that the exponentials in (16) do not overflow. We also note that the introduction of the positive regularization in ε noticeably spreads the transportation matrix - see Figure 2.

Sensitivity to dimension. We consider the relationship between the convergence rate and the dimensions (I, J) of the problem. The theoretical results 5.1 and 5.2 depend on $(\min_i \mu_i) + (\min_j \beta_j)$, which scales with $1/\min(I, J)$ if μ and β are uniform on their support. We generate $\mathcal{X}$ and $\mathcal{Y}$ randomly by drawing two sets of independent Gaussian vectors of respective sizes I and J. We pick μ to be the uniform measure on $\mathcal{X}$, and the cost matrix is taken so that $C_{i,j}$ is the distance between X_i and Y_j. We compute the gradient norm of the objective function F at the averaged iterates $\overline{a_t}$, $\overline{b_t}$. The results can be seen in Figure 3.

The observed rate for the convergence of the gradient norm is $O(t^{-\delta})$, with δ of order $1/2$ as would be predicted from

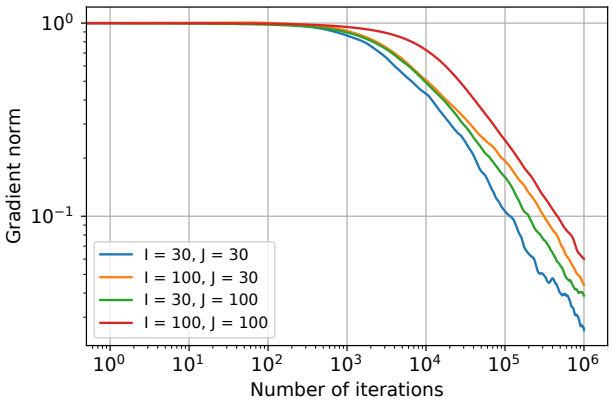

Figure 3. Convergence of the gradient norm for different dimensions.

the theorem 5.1. Overall, a lower dimension increases performance, especially after a small number of iterations. An increase of the support size J of the target decreases performance less than an increase of the sample size I for the input measure.

Choice of the learning rate. As noted above, a choice of learning rate that is large compared to ε can lead to a divergence of the dual variables. This is due to the exponential dependency of the gradients in a and b. Experiments suggest the learning rate

$$\gamma_t = \min\left\{\frac{1}{\lambda t}, \frac{c_0 \varepsilon}{\sqrt{t}}\right\}.$$

Figure 4 shows the convergence to the target with different choices of c_0. Here $\varepsilon = 0.01$, $\eta = 0.02$, with the same problem is the same as in the experiments on the regularization term shown in Figure 1.

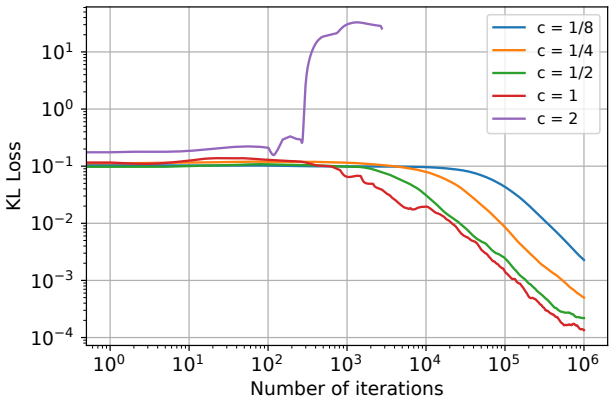

Figure 4. Comparison of the learning rates.

A regression on the curves shows that the empirical convergence rate is of order $O\left(t^{-\delta}\right)$ with δ close to 1, which matches with theorem 5.1. We remark that the greater c_0 is, the better the algorithm converges, until it becomes unstable and does not converge anymore for $c_0 > 1$. This instability was observed consistently for a large range of values of ε and η. The choice $c_0 = 1/2$ appears to be reasonable for both stability and convergence.

6. Conclusion

We consider the problem of minimizing a doubly regularized optimal transport cost over a set of finitely supported measures with fixed support. Using an entropic regularization on the target measure, we derive a stochastic gradient descent on the dual formulation with sublinear (even constant in the simplest case) complexity at each step of the optimization. The algorithm is thus highly paralellizable, and can be used to compute a regularized solution to the Wasserstein barycenter problem. We also provide convergence bounds for the estimator that this algorithm yields after t steps, and demonstrate its performs on randomly generated data.

Aknowledgements. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grant SEQUOIA 724063).

References

Agueh, M. and Carlier, G. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011.

Alaux, J., Grave, E., Cuturi, M., and Joulin, A. Unsupervised hyperalignment for multilingual word embeddings. *CoRR*, abs/1811.01124, 2018.

Altschuler, J., Niles-Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pp. 1964–1974, 2017.

Alvarez-Melis, D., Jaakkola, T., and Jegelka, S. Structured optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pp. 1771–1780, 2018.

Arbel, M., Korba, A., Salim, A., and Gretton, A. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, pp. 6481–6491, 2019.

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

Bassetti, F., Bodini, A., and Regazzini, E. On minimum kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302, 2006.

Berthet, Q. and Kanade, V. Statistical windows in testing for the initial distribution of a reversible markov chain. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 246–255. PMLR, 16–18 Apr 2019.

Boursier, E. and Perchet, V. Private learning and regularized optimal transport. *arXiv preprint arXiv:1905.11148*, 2019.

Claici, S., Chien, E., and Solomon, J. Stochastic wasserstein barycenters. In *International Conference on Machine Learning*, pp. 999–1008, 2018.

Clarkson, K. L., Hazan, E., and Woodruff, D. P. Sublinear optimization for machine learning. *Journal of the ACM (JACM)*, 59(5):1–49, 2012.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

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

Cuturi, M. and Peyré, G. A smoothed dual approach for variational wasserstein problems, 2015.

del Barrio, E., Inouzhe, H., Loubes, J.-M., Matrán, C., and Mayo-Íscar, A. optimalflow: Optimal-transport approach to flow cytometry gating and population matching. *arXiv preprint arXiv:1907.08006*, 2019.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International Conference on Machine Learning*, pp. 1367–1376, 2018.

Feydy, J., Charlier, B., Vialard, F.-X., and Peyré, G. Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 291–299. Springer, 2017.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

Flamary, R., Cuturi, M., Courty, N., and Rakotomamonjy, A. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pp. 3440–3448, 2016.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. *arXiv preprint arXiv:1706.00292*, 2017.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1574–1583, 2019.

Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pp. 2357–2365, 2019.

Grave, E., Joulin, A., and Berthet, Q. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1880–1890, 2019.

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

Janati, H., Muzellec, B., Peyré, G., and Cuturi, M. Entropic optimal transport between (unbalanced) gaussian measures has a closed form, 2020.

Kantorovich, L. V. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4):1381–1382, 2006.

Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Lavenant, H., Claici, S., Chien, E., and Solomon, J. Dynamical optimal transport on discrete surfaces. *ACM Transactions on Graphics (TOG)*, 37(6):1–16, 2018.

Luise, G., Rudi, A., Pontil, M., and Ciliberto, C. Differential properties of sinkhorn approximation for learning with wasserstein distance. In *Advances in Neural Information Processing Systems*, pp. 5859–5870, 2018.

Luise, G., Salzo, S., Pontil, M., and Ciliberto, C. Sinkhorn barycenters with free support via frank-wolfe algorithm. In *Advances in Neural Information Processing Systems*, pp. 9318–9329, 2019.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

Peyré, G., Cuturi, M., et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607, 2019.

Rigollet, P. and Weed, J. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11-12):1228–1235, 2018.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Shamir, O. and Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. *CoRR*, abs/1212.1824, 2012.

Sinkhorn, R. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

Staib, M., Claici, S., Solomon, J. M., and Jegelka, S. Parallel streaming wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pp. 2647–2658, 2017.

Uppal, A., Singh, S., and Poczos, B. Nonparametric density estimation & convergence rates for gans under besov ipm losses. In *Advances in Neural Information Processing Systems*, pp. 9086–9097, 2019.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Weed, J., Bach, F., et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Zemel, Y., Panaretos, V. M., et al. Fréchet means and procrustes analysis in wasserstein space. *Bernoulli*, 25 (2):932–976, 2019.