Stochastic Optimization for Regularized Wasserstein Estimators

Marin Ballu* Quentin Berthet† Francis Bach‡

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 optimisation (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

*Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, mb2193@cam.ac.uk
†Google Research, Brain team, Paris, qberthet@google.com
‡INRIA - DI, ENS, PSL Research University Paris, francis.bach@inria.fr

aspects of the data. This problem has a strongly convex objective, and its solution converges to that of the optimal transport problem whenthe 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., 2019; 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 and 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 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

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

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 and 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 $\mathcal{X} = \{x_i\}_{1 \leqslant i \leqslant I} \subset \mathbb{R}^d$ 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 $\mathcal{Y} = \{y_j\}_{1 \leqslant j \leqslant J} \subset \mathbb{R}^d$. 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

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

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.

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.

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 and 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 2. If $c(x, y) = \|x - y\|^2$, then $\kappa(x, y) =: \phi_\varepsilon(x - y)$ is a Gaussian convolution kernel and the sample measure $\mu = \phi \star \nu^*$ is a convolution, so the solution of (3) is the MLE of the Gaussian deconvolution problem, as already presented by Rigollet and 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.

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 minimisation problem in (2) into a maximisation problem.

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

$$\min_{f \in \mathcal{F}} \max_{a \in L^1(\mu), b \in L^1(\nu)} \mathbb{E}\left[a(X) + b(Y)f(Y) - \varepsilon \exp\left(\frac{a(X) + b(Y) - c(X, Y)}{\varepsilon}\right) + (\eta - \varepsilon)f(Y)\log f(Y)\right], \tag{4}$$

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 (4) 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 $\eta \geqslant \varepsilon$, so the Von Neumann minimax theorem applies: we can swap the minimum and the maximum.

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

$$\max_{a \in L^1(\mu), b \in L^1(\nu)} F(a, b), \tag{5}$$

with

$$F(a, b) = \mathbb{E}\left[a(X) - \varepsilon e^{\frac{a(X) + b(Y) - c(X,Y)}{\varepsilon}}\right] - (\eta - \varepsilon)H^*_\beta\left(-\frac{b}{\eta - \varepsilon}\right), \tag{6}$$

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 (5) is hence given by

$$\max_{(a,b) \in \mathbb{R}^I \times \mathbb{R}^J} F(a, b), \tag{7}$$

with

$$F(a,b) = \mathbb{E}\left[a_i - \varepsilon \exp\left(\frac{a_i + b_j - C_{i,j}}{\varepsilon}\right)\right] - (\eta - \varepsilon) H^*_{\beta,\mathcal{M}}\left(-\frac{b}{\eta - \varepsilon}\right). \tag{8}$$

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 $\mathcal{F}$:

$$H^*_{\beta,\mathcal{M}}(\alpha) = \max_{f \in \mathcal{F}} \mathbb{E}\left[f_j(\alpha_j - \log f_j)\right], \tag{9}$$

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 (4) 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 (7) is a nonempty affine space spanned by the vector $((1,\dots,1),(-1,\dots,-1))$.*
2. *Every solution (a^*,b^*) of (7) verifies*

$$\forall\ i,j,\ |a^*_i + b^*_j - C_{i,j}| \leqslant B, \tag{10}$$

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 (8), we have the gradients of F are written as simple expectations*

$$\nabla_a F = \mathbb{E}\left[(1 - D_{i,j})e_i\right], \tag{11}$$
$$\nabla_b F = \mathbb{E}\left[(f_j - D_{i,j})e_j\right], \tag{12}$$

with $D_{i,j}(a,b) = \exp\left(\frac{a_i + b_j - C_{i,j}}{\varepsilon}\right)$.

4 Stochastic Optimization Methods

The formulas (11) and (12) 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 $(b^t)_{t \geqslant 1}$ are averaged, producing the sequence $\left(\overline{b^t}\right)_{t \geqslant 0}$ of iterates defined by

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

The computation of G_a can be done in $O(1)$, however G_b necessitates the value ν_j^* in (9) 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

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

with m being defined in (10). We define

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

and we initialize

$$a_i = b_j = \overline{b_j} = \mathrm{ini}(C, \varepsilon, m). \tag{13}$$

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 $\{y_j\}_{1 \leqslant j \leqslant J} \subset \mathbb{R}^d$, with the assumption that $\eta > \varepsilon$. Then, if the max is attained on the interior of the simplex, we have the optimum

$$\nu_j^* = \frac{\beta_j e^{-b_j/(\eta-\varepsilon)}}{\sum_k \beta_k e^{-b_k/(\eta-\varepsilon)}}. \tag{14}$$

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 $a_i = b_j = \overline{b_j} = \mathrm{ini}(C, \varepsilon, m)$, $S = e^{-\frac{\mathrm{ini}(C,\varepsilon,m)}{\eta-\varepsilon}}$..
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.

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:

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

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 $(\nu^k)_{1\leqslant k\leqslant K}$ supported in supported in the set $\{y_j\}_{1\leqslant j\leqslant J} \subset \mathbb{R}^d$, 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 (9) becomes

$$H^*_{\beta,\mathcal{M}}(\alpha) = \max_{\theta\in\Delta_K} (\alpha - \log(M\theta) + \log(\beta))^T M\theta, \tag{15}$$

with the log being taken component-wise.

Proposition 4.1. *The maximization problem (15) has a solution*

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

which 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 (12) 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:

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

with $M^\dagger$ being the Moore-Penrose inverse of the matrix M. The other computations are unchanged, apart from Equation (9), replaced by

$$\begin{aligned} H^*_{\beta,\mathcal{M}}(\alpha) &= \max_{\theta\in\Delta_K} \alpha^T M\theta - (\eta-\varepsilon)\,\mathrm{KL}(\theta, M^\dagger\beta) \\ &= \max_{\theta\in\Delta_K} (M^T\alpha - \log(\theta) + \log(M^\dagger\beta))^T\theta. \end{aligned} \tag{16}$$

Proposition 4.2. *The maximization problem (16) 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 $K \ll J$. The algorithm is the following:

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.

Wasserstein barycenters. Algorithm 4 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 $p \geqslant 1$, 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 weights $\theta^1, \dots, \theta^K$ is the solution of the problem

$$\min_{\nu} \sum_{k=1}^{K} \theta_k \, \mathrm{OT}(\mu^k, \nu). \tag{17}$$

This optimization and the barycenter that it defines was introduced by Agueh and 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 and 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 (17):

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

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

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

Here F_k is defined like the function F in (6) 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 minimisation problem

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

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 θ_k, and updating a_k and b as in algorithm 4 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 by Staib et al. (2017).

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 averaged stochastic gradient descent with decreasing stepsize (Shamir and 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. 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:*

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

Theorem 5.2. *With stepsize $\gamma_t = \frac{c_0 \varepsilon}{\sqrt{t}}$, $c_0 \leqslant Be^{-m}/\varepsilon$, the estimator verifies the following bound:*

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

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 (13), then the second order moments of the stochastic gradients are bounded:*

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

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

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

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

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

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 and Zhang (2012) on convex functions gives the bound

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

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

Remark 3. 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 4. 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 to growth of the dual variables 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 (10), 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 introduction of the positive regularization in η noticeably spreads the transportation matrix, and provides a solution that is closer to the uniform law on $\mathcal{Y}$. We use the learning rate provided by Theorem 5.2.

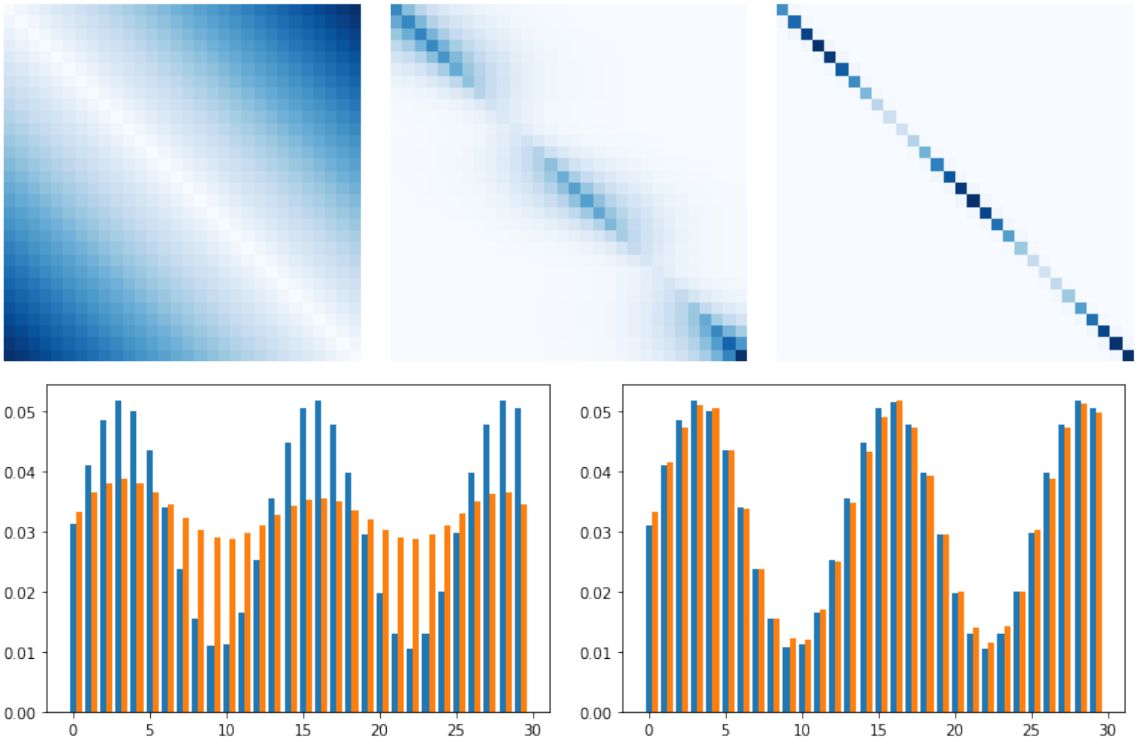

Figure 1: Effect of the regularization. *Upper plots, from left to right:* cost matrix used, transportation matrix for $\varepsilon = \eta - \varepsilon = 0.1$ after 10^5 iterations, and for $\varepsilon = \eta - \varepsilon = 0.01$. *Lower plots, from left to right:* Target measure μ in blue, estimator in orange, $\varepsilon = \eta - \varepsilon = 0.1$, then for $\varepsilon = \eta - \varepsilon = 0.01$.

The regularization term η should be greater than ε, and brings the estimated measure closer to the uniform measure. 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 (14) do not overflow.

Sensibility 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 $\mathcal{Y}$ as two samples of I and J independent Gaussian vectors, μ is the uniform measure on $\mathcal{X}$, and $C_{i,j}$ is the distance matrix 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 gradient norm here converges at rate $O(T^{-\delta})$, with $\delta \geqslant 1/2$ as would be predicted from the theorem 5.1, except in the case $I > J$ where $1/4 \leqslant \delta < 1/2$, which matches better with the bound in Theorem 5.2. An increase of the sample size I for the input measure seems to decrease performance

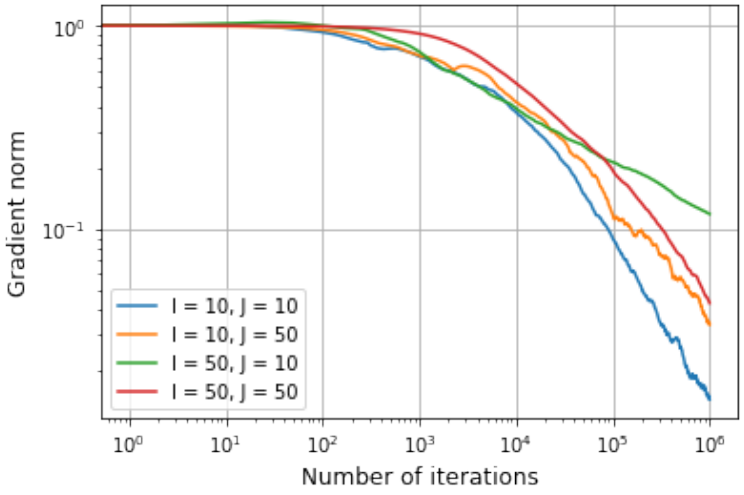

Figure 2: Convergence of the gradient norm for different dimensions.

while an increase of the support size J of the target increases performance. It means that a finer grid of points in $\mathcal{Y}$ will provide a faster convergence to the optimal estimator.

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

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

The following graphs show the convergence to the target with different choices of c_0. Here $\varepsilon = 0.001$, $\eta = 0.002$, with the same problem is the same as in the experiments on the regularization term.

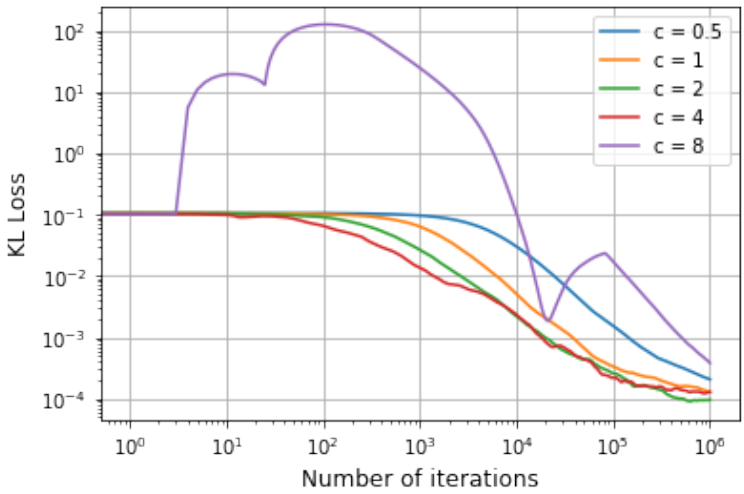

Figure 3: Comparison of the learning rates.

A regression on the curves shows that the empirical convergence rate is of order $O\left(T^{-\delta}\right)$ with $\delta > 1$, 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 $c_0 > 5$. This instability was observed consistently for a large range of values of ε and η. The choice $c_0 = 2$ 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).

7 References

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

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. Unsupervised hyper-alignment for multilingual word embeddings. 2019.

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

David Alvarez-Melis, Tommi Jaakkola, and Stefanie Jegelka. Structured optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1771–1780, 2018.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, pages 6481–6491, 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

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

Quentin Berthet and Varun Kanade. Statistical windows in testing for the initial distribution of a reversible markov chain. pages 246–255, 2019.

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

Sebastian Claici, Edward Chien, and Justin Solomon. Stochastic wasserstein barycenters. In *International Conference on Machine Learning*, pages 999–1008, 2018.

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

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

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

Eustasio del Barrio, Hristo Inouzhe, Jean-Michel Loubes, Carlos Matrán, and Agustín Mayo-Íscar. optimalflow: Optimal-transport approach to flow cytometry gating and population matching. *arXiv preprint arXiv:1907.08006*, 2019.

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

Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 291–299. Springer, 2017.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pages 3440–3448, 2016.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. *arXiv preprint arXiv:1706.00292*, 2017.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583, 2019.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365, 2019.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890, 2019.

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

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

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

Hugo Lavenant, Sebastian Claici, Edward Chien, and Justin Solomon. Dynamical optimal transport on discrete surfaces. *ACM Transactions on Graphics (TOG)*, 37(6):1–16, 2018.

Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with wasserstein distance. In *Advances in Neural Information Processing Systems*, pages 5859–5870, 2018.

Giulia Luise, Saverio Salzo, Massimiliano Pontil, and Carlo Ciliberto. Sinkhorn barycenters with free support via frank-wolfe algorithm. In *Advances in Neural Information Processing Systems*, pages 9318–9329, 2019.

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

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

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

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

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

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

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

Matthew Staib, Sebastian Claici, Justin M Solomon, and Stefanie Jegelka. Parallel streaming wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pages 2647–2658, 2017.

Ananya Uppal, Shashank Singh, and Barnabas Poczos. Nonparametric density estimation & convergence rates for gans under besov ipm losses. In *Advances in Neural Information Processing Systems*, pages 9086–9097, 2019.

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

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

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

A Proofs of technical results

Proof of proposition 2.1. This proof follows the reasoning in Rigollet and Weed (2018). Let $\mu = \frac{1}{I}\sum_i \delta_{X_i}$ be the empirical measure of the sample (X_i). We first remark that the log-likelihood of X_i defined by

$$\ell_\nu(X_i) := \log \int \kappa(X_i, y) d\nu(y)$$

verifies

$$\ell_\nu(X_i) = \log \mathbb{E}_{Y\sim\nu}\left[\kappa(X_i, Y)\right].$$

With the Legendre transform of the relative entropy, we obtain

$$\ell_\nu(X_i) = \sup_{\gamma_i} \mathbb{E}_{Y\sim\gamma_i}\left[\log \kappa(X_i, Y)\right] - \mathrm{KL}(\gamma_i, \nu)$$

with the minimum being over every probability measures γ_i on $\mathcal{Y}$. The MLE maximizes

$$\frac{1}{I}\sum_i \ell_\nu(X_i) = \mathbb{E}_{X\sim\mu}\left[\ell_\nu(X)\right]$$

over $\nu \in \mathcal{M}$, it can be written as

$$\max_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\left[\log \kappa(X, Y)\right] - \mathbb{E}_{X\sim\mu}\left[KL(\pi(X,\cdot),\nu)\right],$$

with $\pi(X,\cdot)$ being the conditional probability of π, defined by $\pi(X_i,\cdot) := \gamma_i$. We have

$$\begin{aligned}
\mathbb{E}_{X\sim\mu}\left[KL(\pi(X,\cdot),\nu)\right] =& \frac{1}{I}\sum_i \mathbb{E}_{Y\sim\nu}\left[\log \frac{d\pi(X_i,\cdot)}{d\nu}(Y)\right], \\
=& \frac{1}{I}\sum_i \mathbb{E}_{Y\sim\nu}\left[\log \frac{d\pi}{d\mu\otimes\nu}(X_i, Y)\right] - \log I, \\
=& KL(\pi, \mu\otimes\nu) - \log I.
\end{aligned}$$

Thus the MLE minimizes

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

which is the regularized optimal transport cost between μ and ν. □

Proof of Proposition 3.3.

1. The function $H^*_{\beta,\mathcal{M}}$ is a Legendre transform, so it is convex, and thus $-F$ is convex as a sum of convex functions. Moreover F is bounded from above:

$$\begin{aligned}
F(a,b) \leqslant& C_1\mathbb{E}[a_i + b_j] - C_2\mathbb{E}\left[e^{\frac{a_i+b_j}{\varepsilon}}\right], \\
\leqslant& C_3,
\end{aligned}$$

   where C_3 does not depend on a or b. Thus the set of solutions is nonempty. F is invariant by the translation $(a, b) \mapsto (a_1 + c, \ldots, a_I + c, b_1 - c, \ldots, b_J - c)$, so each solution generates an affine set of solutions spanned by the vector $((1, \ldots, 1), (-1, \ldots, -1))$. We can conclude using the strong convexity on the slice $\{\sum_i \mu_i a_i = \sum_j \beta_j b_j\}$, which implies that there exists only one solution on this slice.

2. The solution (a^*, b^*) solves the following system

$$\begin{cases}
\nabla_a F(a^*, b^*) = 0, \\
\nabla_b F(a^*, b^*) = 0.
\end{cases}$$

With notations $A_i = e^{a_i^*/\varepsilon}$, $B_j = e^{b_j^*/\varepsilon}$, $\Gamma_{i,j} = e^{-C_{i,j}/\varepsilon}$, the two equations can be written as

$$\begin{cases} \forall\ 1 \leqslant i \leqslant I, & 1 - A_i \sum_j \beta_j B_j \Gamma_{i,j} = 0, \\ \forall\ 1 \leqslant j \leqslant J, & f_j - B_j \sum_i \mu_i A_i \Gamma_{i,j} = 0. \end{cases} \tag{18}$$

Thus

$$\begin{cases} \forall\ 1 \leqslant i \leqslant I, & A_i = \frac{1}{\sum_j \beta_j B_j \Gamma_{i,j}}, \\ \forall\ 1 \leqslant j \leqslant J, & B_j = \frac{f_j}{\sum_i \mu_i A_i \Gamma_{i,j}}. \end{cases} \tag{19}$$

We also remark that by multiplying the second term of (18) by β_j and summing over j we get

$$\sum_{i,j} \mu_i A_i \beta_j B_j \Gamma_{i,j} = 1. \tag{20}$$

By multiplying the equations in (19) we have for all i, j:

$$A_i B_j \Gamma_{i,j} = \frac{f_j \Gamma_{i,j}}{\sum_{k,l} \mu_k A_k \Gamma_{k,j} \beta_l B_l \Gamma_{i,l}}$$

thus using (20):

$$f_j \min_{k,l} \frac{\Gamma_{i,j}\Gamma_{k,l}}{\Gamma_{k,j}\Gamma_{i,l}} \leqslant A_i B_j \Gamma_{i,j} \leqslant f_j \max_{k,l} \frac{\Gamma_{i,j}\Gamma_{k,l}}{\Gamma_{k,j}\Gamma_{i,l}}$$

finally

$$e^{-m-2R_C/\varepsilon} \leqslant A_i B_j \Gamma_{i,j} \leqslant e^{m+2R_C/\varepsilon}.$$

3. We now prove that $-F$ is strongly convex. We compute

$$-\nabla_a^2 F = \mathbb{E}\left[\frac{1}{\varepsilon} D_{i,j} E_{i,i}\right],$$

$$-\nabla_b^2 F = -\nabla_b \nu^* + \mathbb{E}\left[\frac{1}{\varepsilon} D_{i,j} E_{j,j}\right],$$

$$-\nabla_a \nabla_b F = \mathbb{E}\left[\frac{1}{\varepsilon} D_{i,j} E_{i,j}\right].$$

We remark that

$$\nu^* = \mathrm{softmax}(-b_j/\eta + \log \beta_j),$$

so

$$-\nabla_b \nu^* = \frac{1}{\eta} S$$

with

$$S := (\nabla \mathrm{softmax})(-b_j/\eta + \log \beta_j),$$
$$S = (\nu_i(\delta_{i,j} - \nu_j))_{i,j}.$$

We remark that $S \geqslant 0$ since

$$\begin{aligned} u^T S u &= \sum_i \nu_i u_i^2 - \left(\sum_i \nu_i u_i\right)^2 \\ &= \mathbb{E}_\nu[U^2] - (\mathbb{E}_\nu[U])^2 \geqslant 0 \end{aligned}$$

by Jensen, with $U = u_j$ with probability ν_j. It implies $-\nabla_b \nu_j^* \geqslant 0$. So

$$-\nabla^2_{a,b} F \geqslant \frac{1}{\varepsilon} M,$$

with

$$M := \mathbb{E}\left[D_{i,j} \begin{pmatrix} E_{i,i} & E_{i,j} \\ E_{j,i} & E_{j,j} \end{pmatrix} \right].$$

As we want to prove strong convexity on the slice $\sum_i \mu_i a_i = \sum_j \beta_j b_j$, we compute

$$(a,b)^T M (a,b) = \mathbb{E}\left[D_{i,j} (a_i + b_j)^2 \right] \geqslant e^{-B/\varepsilon} \mathbb{E}\left[(a_i + b_j)^2 \right].$$

We add that

$$\mathbb{E}\left[(a_i + b_j)^2 \right] = \sum_i \mu_i a_i^2 + \sum_j \beta_j b_j^2 + 2(\sum_i \mu_i a_i)(\sum_j \beta_j b_j)$$

thus

$$\mathbb{E}\left[(a_i + b_j)^2 \right] = \sum_i (\mu_i + \mu_i^2) a_i^2 + \sum_j (\beta_j + \beta_j^2) b_j^2$$

since we are on the slice. So $M \geqslant \lambda \mathrm{Id}$ and finally $-F$ is $\lambda-$strongly convex with

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

4. We compute the gradients of F:

$$\frac{\partial F}{\partial a_i}(a,b) = \mu_i - \mu_i \sum_{j=1}^{J} \beta_j D_{i,j}(a,b), \tag{21}$$

$$\frac{\partial F}{\partial b_j}(a,b) = \nu_j^*(-b/\eta) - \beta_j \sum_{i=1}^{I} \mu_i D_{i,j}(a,b), \tag{22}$$

with $D_{i,j}(a,b) = e^{\frac{a_i + b_j - C_{i,j}}{\varepsilon}}$. If we take i and j to be independent random variables following the laws (μ_i) and (β_j) respectively, we have the desired expression for the gradients.

□

Proof of Lemma 1. With the initial conditions, we guarantee that $0 \leqslant G_a^0 \leqslant 1$ and $0 \leqslant G_b^0 \leqslant f_j \leqslant e^m$. At each timestep t, we have

$$\|\nabla F_{i,j}^t\|^2 \leqslant \max\{2e^{2m}, 2(D_{i,j}^t)^2\},$$

with i, j being two independent random variables following the laws μ and β respectively. If $D_{i,j}^t \geqslant e^m$, then $G_a + G_b \leqslant 0$ and

$$D_{i,j}^{t+1} = D_{i,j}^t e^{\frac{G_a + G_b}{\varepsilon}} \leqslant D_{i,j}^t.$$

Moreover if $D_{i,j}^t \leqslant e^m$ then $\|\nabla F_{i,j}^t\|^2 \leqslant 1 + e^{2m}$ thus $\mathbb{E}\left[\max\{2e^{2m}, (D_{i,j}^t)^2\}\right]$ is a decreasing function of t. Thus we have the bound

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

□

Proof of Lemma 2. We first assume that (a, b) and (a^*, b^*) are on the slice $\{\sum_i \mu_i a_i = \sum_j \beta_j b_j\}$. By strong convexity of $-F$ on this slice we have

$$|b - b^*|^2 \leqslant \frac{2(F(a^*, b^*) - F(a, b))}{\lambda}. \tag{23}$$

We remark that the function $g : b \mapsto KL(\nu(b^*), \nu(b))$ verifies

$$\begin{aligned}
\partial_i g(b) &= -\sum_j \nu_j(b^*)\partial_i \log \nu_j(b), \\
&= -\sum_j \nu_j(b^*)\nu_j(b)^{-1}\partial_i \nu_j(b), \\
&= \frac{1}{\eta}\sum_j \nu_j(b^*)\nu_j(b)^{-1}\nu_i(\delta_{ij} - \nu_j(b)), \\
&= \frac{\nu_i(b^*) - \nu_i(b)}{\eta - \varepsilon},
\end{aligned}$$

thus

$$\begin{aligned}
\partial_i\partial_j g(b) &= -\frac{\partial_j \nu_i(b)}{\eta - \varepsilon}, \\
&= -\frac{\nu_j(b)(\delta_{ij} - \nu_i(b))}{\eta - \varepsilon},
\end{aligned}$$

so the Hessian matrix $\nabla^2 g(b)$ of g is a sum of a diagonal matrix with the negative values $-\nu_j(b)/(\eta-\varepsilon)$ and the one-rank matrix $(\nu_j(b)\nu_i(b)/(\eta - \varepsilon))_{i,j}$. Hence the eigenvalues of $\nabla^2 g(b)$ are contained in $[-1/(\eta - \varepsilon), 1/(\eta - \varepsilon)]$, thus Taylor's inequality gives

$$\begin{aligned}
g(b) &\leqslant g(b^*) + |b - b^*|\|\nabla g(b^*)\| + \frac{|b - b^*|^2}{2(\eta - \varepsilon)}, \\
&\leqslant \frac{|b - b^*|^2}{2(\eta - \varepsilon)},
\end{aligned}$$

because $g(b^*) = 0$ and $\nabla g(b^*) = 0$. We complete the proof with (23). For the case where the vector (a, b) or (a^*, b^*) is not on the slice $\{\sum_i \mu_i a_i = \sum_j \beta_j b_j\}$, we note that adding a constant vector $c = (c_1, \ldots, c_1)$ to b does not change the value of $\nu(b)$, and that F is invariant by translation in the direction $(-c, +c)$. With $c_1 = \left(\sum_i \mu_i a_i - \sum_j \beta_j b_j\right)/2$, the vectors $(a', b') = (a + c, b - c)$ are on the slice and verify $\nu(b') = \nu(b)$ and $F(a', b') = F(a, b)$. Hence the result for (a', b') implies the result for (a, b). □