Learning with minibatch Wasserstein : asymptotic and gradient properties

Supplementary material

Outline. The supplementary material of this paper is organized as follows:

- In section A, we first review the formalism with definitions, basic property proofs, statistical proofs and optimization proofs. Then we give details about the 1D case.
- In section B, we give extra experiments for domain adaptation, minibatch Wasserstein gradient flow in 2D and on the celebA dataset and finally, color transfer.

A Formalism

In what follows, without any loss of generality and in order to simplify the notations we will work with the cost matrix $C = C(X, Y) = (|X_i - Y_i|)_{1 \leqslant i,j \leqslant n}$.

A.1 Definitions

We start giving the formal definitions for the transportation plan Π_m. We recall that the discrete entropy of a coupling matrix is defined as $H(P) = -\sum_{i,j} P_{i,j} \left(\log\left(P_{i,j}\right) - 1\right)$ [chapitre 4, [Peyré and Cuturi, 2019]]. The entropic regularization parameter $\varepsilon \in \mathbb{R}_+$.

Definition 1 (Mini-batch Transport)**.** *Let $A \in \mathcal{P}_m(\alpha_n)$ and $B \in \mathcal{P}_m(\beta_n)$ be two sets. We denote by $\Pi^0_{A,B}(\alpha_n, \beta_n) = (\Pi^0_{A,B}(i,j))_{1 \leqslant i,j \leqslant m} \in \mathbb{R}^{m \times m}$ an optimizer of the optimal transport. Formally,*

$$\Pi^0_{A,B} = \underset{\Pi \in U(A,B)}{\operatorname{argmin}} \langle \Pi, C_{|A,B} \rangle - \varepsilon H(\Pi) \tag{1}$$

where $C_{|A,B} \in \mathbb{R}^{m \times m}$ is the matrix extracted from C by considering elements of the lines (resp. columns) of C which belong to A (resp. B) and H the entropy term. ε is a positive real number that can be equal to 0 to get the original OT problem.

For two sets $A \in \mathcal{P}_m(\alpha_n)$ and $B \in \mathcal{P}_m(\beta_n)$ we denote by $\Pi_{A,B}(\alpha_n, \beta_n) \in \mathbb{R}^{n \times n}$ the matrix

$$\Pi_{A,B} = (\Pi^0_{A,B}(i,j)\mathbf{1}_A(i)\mathbf{1}_B(j))_{(i,j) \in \alpha_n \times \beta_n} \tag{2}$$

Definition 2 (Averaged mini-batch transport)**.** *We define the empirical averaged mini-batch transport matrix $\Pi_m(\alpha_n, \beta_n)$ by the formula*

$$\Pi_m := \frac{1}{\binom{n}{m}^2} \sum_{A \in \mathcal{P}_m(\alpha_n)} \sum_{B \in \mathcal{P}_m(\beta_n)} \Pi_{A,B} \tag{3}$$

Moreover, we can define the averaged Wasserstein distance over all mini batches as :

$$U_W(\alpha_n, \beta_n) = \langle \Pi_m, C \rangle \tag{4}$$

Remark 1. *Note that this construction is consistent with $U_h(\alpha_n, \beta_n)$.*

A.2 Basic properties

Proposition 1. *Π_m is a transportation plan between the empirical distributions α_n, β_n.*

Proof. We need to verify that the marginals sum to one -e.g. that the sum over any row (resp. column) is equal to $\frac{1}{n}$. Without loss of generality, we will fix a source sample (or row): i_0. A simple combinatorial argument gives that $\sum_{A \in \mathcal{P}_m(\alpha_n)} \mathbf{1}_A(i_0) = \binom{n-1}{m-1}$. Now we are ready to sum over the row i_0.

$$\sum_{j=1}^{n} \Pi_m(i_0, j) = \frac{1}{\binom{n}{m}\binom{n}{m}} \sum_{j=1}^{n} \sum_{A \in \mathcal{P}_m(\alpha_n)} \sum_{B \in \mathcal{P}_m(\beta_n)} \Pi_{A,B}(i_0, j) \tag{5}$$

$$= \frac{1}{\binom{n}{m}\binom{n}{m}} \sum_{B \in \mathcal{P}_m(\beta_n)} \sum_{j=1}^{n} \Pi^0_{A,B}(i_0, j) \mathbf{1}_B(j) \sum_{A \in \mathcal{P}_m(\alpha_n)} \mathbf{1}_A(i_0) \tag{6}$$

$$= \frac{1}{\binom{n}{m}\binom{n}{m}} \sum_{B \in \mathcal{P}_m(\beta_n)} \underbrace{\sum_{j=1}^{n} \Pi^0_{A,B}(i_0, j) \mathbf{1}_B(j)}_{=1/m} \binom{n-1}{m-1} \tag{7}$$

$$= \frac{1}{\binom{n}{m}\binom{n}{m}} \binom{n}{m} \frac{1}{m} \binom{n-1}{m-1} \tag{8}$$

$$= \frac{1}{n} \tag{9}$$

The argument is similar for the summation over any column.

□

Remark 2 (Positivity, symmetry and bias)**.** *Let $m < n$, the quantity U_h is positive and symmetric but also stricly positive, i.e $U_h(\alpha_n, \alpha_n) > 0$. Indeed,*

$$U_h(\alpha_n, \alpha_n) := \frac{1}{\binom{n}{m}^2} \sum_{A \in \mathcal{P}(\alpha_n)} \sum_{A' \in \mathcal{P}(\alpha_n)} h(A, A') \tag{10}$$

$$= \frac{1}{\binom{n}{m}^2} \sum_{(A,A') \in \mathcal{P}(\alpha_n) \times \mathcal{P}(\beta_n), A \neq A'} h(A, A') > 0 \tag{11}$$

Convexity We introduce a few notations. Let $\mathcal{D}(\mathbb{R}^d)$ be the space defined by

$$\mathcal{D}(\mathbb{R}^d) := \{\sum_{i=1}^{p} \gamma_i \delta_{x_i} : (\gamma_i)_{1 \leq i \leq p} \in (\mathbb{R}_+)^p, \sum_{i=1}^{p} \gamma_i = 1; p \in \mathbb{N}; (x_i)_{1 \leq i \leq p} \in (\mathbb{R}^d)^p\} \tag{12}$$

It is easy to see that $\mathcal{D}(\mathbb{R}^d)$ is convex. One can actually extend in a natural way the definition of U_h to the set $\mathcal{D}(\mathbb{R}^d) \times \mathcal{D}(\mathbb{R}^d)$. Assuming this can be done, the intuition for convexity is that U_h is an average of convex terms [(section 9.1 and prop 4.6, [Peyré and Cuturi, 2019]]. We then claim the convexity of the following maps:

$$\begin{aligned} (\alpha_n, \beta_n) &\mapsto U_W(\alpha_n, \beta_n) \\ \mathcal{D}(\mathbb{R}^d) \times \mathcal{D}(\mathbb{R}^d) &\to \mathbb{R} \end{aligned}$$

and for $h = W_\epsilon$ or $h = S_\epsilon$:

$$\begin{aligned} \alpha_n &\mapsto U_h(\alpha_n, \beta_n) \\ \mathcal{D}(\mathbb{R}^d) &\to \mathbb{R} \\ \beta_n &\mapsto U_h(\alpha_n, \beta_n) \\ \mathcal{D}(\mathbb{R}^d) &\to \mathbb{R} \end{aligned}$$

A.3 Statistical proofs

Note that because the distributions α and β are compactly supported, there exists a constant $M > 0$ such that for any $1 \leqslant i, j \leqslant n$, $|X_i - Y_j| \leqslant M$ with $M := \text{diam}(\text{Supp}(\alpha) \cup \text{Supp}(\beta))$. We define the following quantity depending on the

OT loss h:

$$M_h = \begin{cases} \text{diam}(\text{Supp}(\alpha) \cup \text{Supp}(\beta)) & \text{if } h = W \\ \frac{3}{2}\left\{\text{diam}(\text{Supp}(\alpha) \cup \text{Supp}(\beta)) + \varepsilon(2\log_2(m)+1)\right\} & \text{if } h = W_\varepsilon \text{ or } S_\varepsilon \end{cases} \tag{13}$$

Lemma 1 (Upper bounds)**.** *Let $(A, B) \in \mathcal{P}(\alpha_n) \times \mathcal{P}(\beta_n)$. We have the following bound for each of the above considered OT losses h:*

$$|h(A, B)| \leqslant 2M_h \tag{14}$$

Proof. We start with the case $h = W$. Note that with our choice of cost matrix $C = (C_{i,})$ one has $0 \leqslant C_{i,j} \leqslant M_W$. We have for a transport plan $\Pi = (\Pi_{i,j})$ between A and B (with respect to the cost matrix $C_{|A,B}$)

$$|\langle \Pi, C_{|A,B}\rangle| \leqslant \sum_{1\leqslant i,j\leqslant m} (C_{|A,B})_{ij}\Pi_{i,j} \leqslant M_W \sum_{1\leqslant i,j\leqslant m} \Pi_{i,j} = M_W$$

Hence, $h(A, B) \leqslant M_W$.
If $h = W_\varepsilon$ for an $\varepsilon > 0$. Let us denote by $E(q) = -\sum_{i=1}^r q_i \log(q_i)$ the Shannon entropy of the discrete probability distribution $q = (q_i)_{1\leqslant i\leqslant r}$. Using the classical fact : $0 \leqslant E(q) \leqslant \log_2(r)$ one estimates for a transport plan Π:

$$|\langle \Pi, C_{|A,B}\rangle - \varepsilon H(\Pi)| \leqslant M_W + \varepsilon(E(\Pi)+1) \leqslant M_W + \varepsilon(\log_2(m^2)+1) \leqslant 2M_h$$

which gives the intended bound by definition of W_ε. Lastly, for $h = S_\varepsilon$, since it is basically the sum of three terms of the form W_ε one can conclude. □

Proof of Theorem 1 We now give the details of the proof of theorem 1. We start by recalling the definitions of our losses.

Definition 3 (Minibatch Wasserstein definitions)**.** *Given an OT loss h and an integer $m \leq n$, we define the following quantities:*

The continuous loss:

$$U_h(\alpha, \beta) := \mathbb{E}_{(X,Y)\sim\alpha^{\otimes m}\otimes\beta^{\otimes m}}[h(X, Y)] \tag{15}$$

The semi-discrete loss:

$$U_h(\alpha_n, \beta) := \binom{n}{m}^{-1} \sum_{A\in\mathcal{P}(\alpha_n)} \mathbb{E}_{Y\sim\beta^{\otimes m}}[h(A, Y)] \tag{16}$$

The discrete-discrete loss:

$$U_h(\alpha_n, \beta_n) := \binom{n}{m}^{-2} \sum_{A\in\mathcal{P}(\alpha_n)} \sum_{B\in\mathcal{P}(\beta_n)} h(A, B) \tag{17}$$

The subsample discrete-discrete loss. Pick an integer $k > 0$. We define:

$$\widetilde{U}_h^k(\alpha_n, \beta_n) := k^{-1} \sum_{(A,B)\in D_k} h(A, B) \tag{18}$$

where D_k is a set of cardinality k whose elements are drawn at random from the uniform distribution on $\Gamma := \mathcal{P}_m(\{X_1, \cdots, X_n\}) \times \mathcal{P}_m(\{Y_1, \cdots, Y_n\})$. Where h can be the Wasserstein distance W, the entropic loss W_ε or the sinkhorn divergence S_ε for a cost $c(\boldsymbol{x}, \boldsymbol{y})$.

Lemma 2 (U-statistics concentration bound)**.** *Let $\delta \in (0, 1)$ and m be fixed, we have a concentration bound between $U_h(\alpha_n, \beta_n)$ and the expectation over minibatches $U_h(\alpha, \beta)$ depending on the number of empirical data n which follow α and β.*

$$|U_h(\alpha_n, \beta_n) - U_h(\alpha, \beta)| \leq M_h\sqrt{\frac{\log(2/\delta)}{2\lfloor n/m\rfloor}} \tag{19}$$

with probability at least $1 - \delta$. Furthermore, a Bernstein concentration bound is available. Let us denote the variance of the OT loss h over the batches σ_h^2, i.e., $\sigma_h^2 = Var(h(X_1, \cdots, X_m, Y_1, \cdots, Y_m))$. The variance is bounded by M_h^2. Then we have with probability at least ε:

$$P(|U_h(\alpha_n, \beta_n) - U_h(\alpha, \beta)| \geq \varepsilon) \leqslant 2\, exp\left(\frac{-\lfloor n/m\rfloor\varepsilon^2}{2(\sigma_h^2 + \frac{M_h}{3}\varepsilon)}\right) \leqslant 2\, exp\left(\frac{-\lfloor n/m\rfloor\varepsilon^2}{2(M_h^2 + \frac{M_h}{3}\varepsilon)}\right) \tag{20}$$

Proof. $U_h(\alpha_n, \beta_n)$ is a two-sample U-statistic and $U_h(\alpha, \beta)$ is its expectation as α_n and β_n have *iid* random variables. $U_h(\alpha_n, \beta_n)$ is a sum of dependant variables and Hoeffding found a way to rewrite $U_h(\alpha_n, \beta_n)$ as a sum of independent random variables. As our data are *iid* and our OT loss is bounded, we can apply its third theorem to our U-statistic. The proof can be found in [Hoeffding, 1963, Section 5] (the two sample U-statistic case is discussed in 5.b) . □

Lemma 3 (Deviation bound)**.** *Let α_n and β_n be empirical distributions of respectively α and β, let $\delta \in (0, 1)$ and $k \geqslant 1$. We have a deviation bound between $\widetilde{U}_h^k(\alpha_n, \beta_n)$ and $U_h(\alpha_n, \beta_n)$ depending on the number of batches k.*

$$|\widetilde{U}_h^k(\alpha_n, \beta_n) - U_h(\alpha_n, \beta_n)| \leqslant M_h \sqrt{\frac{2\log(2/\delta)}{k}} \tag{21}$$

with probability at least $1 - \delta$.

Proof. First note that $\widetilde{U}_h^k(\alpha_n, \beta_n)$ is an incomplete U-statistic of $U_h(\alpha_n, \beta_n)$. Let us consider the sequence of random variables $((\mathbf{1}_l(A, B)_{(A,B)\in\Gamma})_{1\leqslant l\leqslant k}$ such that $\mathbf{1}_l(A, B)$ is equal to 1 if (A, B) has been selected at the $l-$th draw and 0 otherwise. By construction of $\widetilde{U}_h^k$, the aforementioned sequence is an i.i.d sequence of random vectors and the $\mathbf{1}_l(A, B)$ are bernoulli random variables of parameter $1/|\Gamma|$. We then have

$$\widetilde{U}_h^k(\alpha_n, \beta_n) - U_h(\alpha_n, \beta_n) = \frac{1}{k}\sum_{l=1}^{k} \omega_l \tag{22}$$

where $\omega_l = \sum_{(A,B)\in\Gamma}(\mathbf{1}_l(A, B) - \frac{1}{|\Gamma|})h(A, B)$. Conditioned upon $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_n)$, the variables ω_l are independent, centered and bounded by $2M_h$ thanks to lemma 1. Using Hoeffding's inequality yields

$$\begin{aligned}
\mathbb{P}(|\widetilde{U}_h^k(\alpha_n, \beta_n) - U_h(\alpha_n, \beta_n)| > \varepsilon) &= \mathbb{E}[\mathbb{P}(|\widetilde{U}_h^k(\alpha_n, \beta_n) - U_h(\alpha_n, \beta_n)| > \varepsilon|X, Y)] &(23)\\
&= \mathbb{E}[\mathbb{P}(|\frac{1}{k}\sum_{l=1}^{k} \omega_l)| > \varepsilon|X, Y)] &(24)\\
&\leqslant \mathbb{E}[2e^{\frac{-k\varepsilon^2}{2M^2}}] = 2e^{\frac{-k\varepsilon^2}{2M^2}} &(25)
\end{aligned}$$

which concludes the proof. □

Theorem 1 (Maximal deviation bound)**.** *Let $\delta \in (0, 1)$, $k \geqslant 1$ and m be fixed, we have a maximal deviation bound between $\widetilde{U}_h^k(\alpha_n, \beta_n)$ and the expectation over minibatches $U_h(\alpha, \beta)$ depending on the number of empirical data n which follow α and β and the number of batches k.*

$$|\widetilde{U}_h^k(\alpha_n, \beta_n) - U_h(\alpha, \beta)| \leq M_h \sqrt{\frac{\log(2/\delta)}{2\lfloor n/m \rfloor}} + M_h \sqrt{\frac{2\log(2/\delta)}{k}} \tag{26}$$

with probability at least 1 - δ

Proof. Thanks to lemma 3 and A.3 we get

$$\begin{aligned}
|\widetilde{U}_h^k(\alpha_n, \beta_n) - U_h(\alpha, \beta)| &\leq |\widetilde{U}_h^k(\alpha_n, \beta_n) - U_h(\alpha_n, \beta_n)| + |U_W(\alpha_n, \beta_n) - U_h(\alpha, \beta)| &(27)\\
&\leq M_h \sqrt{\frac{\log(2/\delta)}{2\lfloor n/m \rfloor}} + M_h \sqrt{\frac{2\log(2/\delta)}{k}} &(28)
\end{aligned}$$

with probability at least $1 - (\frac{\delta}{2} + \frac{\delta}{2}) = 1 - \delta$. We can get a sharper bound using the Bernstein inequality instead of the Hoeffding inequality as detailed in lemma . □

Proof of Theorem 2 We now give the details of the proof of theorem 2. In what follows, we denote by $\Pi_{(i)}$ the i-th row of matrix Π. Let us denote by $\mathbf{1} \in \mathbb{R}^n$ the vector whose entries are all equal to 1.

Theorem 2 (Distance to marginals)**.** *Let $\delta \in (0,1)$, we have for all $k \geqslant 1$ and all $1 \leqslant j \leqslant n$:*

$$|\Pi_k(\alpha_n, \beta_n)_{(i)}\mathbf{1} - \frac{1}{n}| \leqslant \sqrt{\frac{2\log(2/\delta)}{k}} \tag{29}$$

with probability at least $1-\delta$.

Proof. We would like to remind that Π_m is a transportation plan between the full input distributions α_n and β_n and hence, it verifies the marginals, i.e $\Pi_m(\alpha_n, \beta_n)_i \times \mathbf{1} = \frac{1}{n}$. Let us consider the sequence of random variables $((\mathbf{1}_p(A,B)_{(A,B)\in\Gamma})_{1\leqslant p\leqslant k}$ such that $\mathbf{1}_p(A,B)$ is equal to 1 if (A,B) has been selected at the $p-$th draw and 0 otherwise. By construction of $\Pi_k(\alpha_n, \beta_n)$, the aforementioned sequence is an i.i.d sequence of random vectors and the $\mathbf{1}_p(A,B)$ are bernoulli random variables of parameter $1/|\Gamma|$. We then have

$$\Pi_k(\alpha_n, \beta_n)_{(i)}\mathbf{1} = \frac{1}{k}\sum_{p=1}^{k} \omega_p \tag{30}$$

where $\omega_p = \sum_{(A,B)\in\Gamma}\sum_{j=1}^{n}(\Pi_{A,B})_{i,j}\mathbf{1}_p(A,B)$. Conditioned upon $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_n)$, the random vectors ω_p are independent, and bounded by 1. Moreover, one can observe that $\mathbb{E}[\Pi_k(\alpha_n, \beta_n)_i\mathbf{1}] = \Pi_m(\alpha_n, \beta_n)_i\mathbf{1}$. Using Hoeffding's inequality yields

$$\mathbb{P}(|\Pi_k(\alpha_n, \beta_n)_i\mathbf{1} - \Pi_m(\alpha_n, \beta_n)_i\mathbf{1})| > \varepsilon) = \mathbb{E}[\mathbb{P}(|\frac{1}{k}\sum_{p=1}^{k}\omega_p - \mathbb{E}[\frac{1}{k}\sum_{p=1}^{k}\omega_p])| > \varepsilon|X,Y)] \tag{31}$$

$$\leqslant 2e^{-2k\varepsilon^2} \tag{32}$$

which concludes the proof. □

A.4 Optimization

The main goal of this section is to give a justification of optimization for our minibatch OT losses by giving the **proof of theorem 3**. More precisely, we show that for the losses W_ε and S_ε, one can exchange the gradient symbol ∇ and the expectation $\mathbb{E}$. It shows for example that a stochastic gradient descent procedure is unbiased and as such legitimate.

Main hypothesis. We assume that the map $\lambda \mapsto C(A, Y_\lambda)$ is differentiable. For instance for GANs, it is verified when the neural network in the generator is differentiable -which is the case if the nonlinear activation functions are all differentiable- and when the cost chosen in the Wasserstein distance is also differentiable.
We introduce the map

$$g : (\Pi, C) \mapsto \langle \Pi, C\rangle - \varepsilon H(\Pi)$$

To prove this theorem, we first define a map we will use the "Differentiation Lemma".

Lemma 4 (Differentiation lemma)**.** *Let V be a nontrivial open set in $\mathbb{R}^p$ and let $\mathcal{P}$ be a probability distribution on $\mathbb{R}^d$. Define a map $C : \mathbb{R}^d \times \mathbb{R}^d \times V \to \mathbb{R}$ with the following properties:*

- *For any $\lambda \in V, \mathbb{E}_\mathcal{P}[|C(X,Y,\lambda)|] < \infty$*
- *For P-almost all $(X,Y) \in \mathbb{R}^d \times \mathbb{R}^d$, the map $V \to \mathbb{R}$, $\lambda \to C(X,Y,\lambda)$ is differentiable.*
- *There exists a P-integrable function $\varphi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $|\partial_\lambda C(X,Y,\lambda)| \leq g(x)$ for all $\lambda \in V$.*

Then, for any $\lambda \in V$, $E_\mathcal{P}[|\partial_\lambda C(X,Y,\lambda)|] < \infty$ and the function $\lambda \to E_\mathcal{P}[C(X,Y,\lambda)]$ is differentiable with differential:

$$E_\mathcal{P}\partial_\lambda[C(X,Y,\lambda)] = \partial_\lambda E_\mathcal{P}[C(X,Y,\lambda)] \tag{33}$$

The following result will also be useful.

Lemma 5 (Danskin, Rockafellar)**.** *Let $g : (z, w) \in \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a function. We define $\varphi : z \mapsto \max_{w \in W} g(z, w)$ where $W \subset \mathbb{R}^d$ is compact. We assume that for each $w \in W$, the function $g(\cdot, w)$ is differentiable and that $\nabla_z g$ depends continuously on (z, w). If in addition, $g(z, w)$ is convex in z, and if $\overline{z}$ is a point such that $\operatorname{argmax}_{w \in W} g(\overline{z}, w) = \{\overline{w}\}$, then φ is differentiable at $\overline{z}$ and verifies*

$$\nabla \varphi(\overline{z}) = \nabla_z g(\overline{z}, \overline{w}) \tag{34}$$

The last theorem shows that the entropic loss is differentiable with respect to the cost matrix. Indeed, the theorem directly applies since the problem is strongly convex. This remark enables us to obtain the following result.

Theorem 3 (Exchange gradient and expectation)**.** *Let us suppose that we have two distributions α and β on two bounded subsets $\mathcal{X}$ and $\mathcal{Y}$, a $\mathcal{C}^1$ cost. Assume $\lambda \mapsto Y_\lambda$ is differentiable. Then for the entropic loss and the Sinkhorn divergence:*

$$\nabla_\lambda \mathop{\mathbb{E}}_{Y_\lambda \sim \beta_\lambda^{\otimes m}} h(A, Y_\lambda) = \mathop{\mathbb{E}}_{Y_\lambda \sim \beta_\lambda^{\otimes m}} \nabla_\lambda h(A, Y_\lambda) \tag{35}$$

Proof. Regarding the Sinkhorn divergence, as it is the sum of three terms of the form W_ε, it suffices to show the theorem for $h = W_\varepsilon$

The first and the third conditions of the Differentiation Lemma are trivial as we have supposed that our distributions have compact supports. Hence, the minibatch Wasserstein exists and is bounded on a finite set. We can also build a measurable function φ which takes the biggest cost value $||c||_\infty$ inside $\mathcal{X}$ and 0 outside. As $\mathcal{X}$ is compact, the integral of the function over $\mathbb{R}^d$ is finite.
The problem is in the second hypothesis where we need to prove that W_ε is differentiable almost everywhere. We have to show that the following function $\lambda \mapsto \varphi_A(\lambda)$ is differentiable:

$$\varphi_A : \lambda \mapsto \min_{\Pi \in U(a,b)} \langle \Pi, C(A, \lambda) \rangle - \varepsilon H(\Pi)$$

where $C(A, \lambda)$ is the cost computed using pairwise distances between A and Y_λ. Since $\lambda \mapsto C(A, \lambda)$ is differentiable almost everywhere by our hypothesis on $\lambda \mapsto y_\lambda$, it suffices, by composition, to show that $C \mapsto \min_{\Pi \in U(a,b)} \langle \Pi, C \rangle - \varepsilon H(\Pi)$ is differentiable in $C \in \mathbb{R}^{m \times m}$. We obtain this using lemma 5 and the fact that there is one unique solution to the entropically regularized optimal transport problem.

□

A.5 1D case

We now give the full combinatorial calculus for the 1D case. We start by sorting all the data and give to each of them an index which reprensents their position after the sorting phase. Then we select and sort all the minibatches. x_j can not be at a position superior to its index j inside a batch. For a fixed x_j, a simple combinatorial arguments tells you that there are $C_{x_j}^i$ sets where x_j is at the i-th position:

$$C_{i,x_j}^{m,n} = \binom{j-1}{i-1}\binom{n-j}{m-i} \tag{36}$$

Suppose that x_j is transported to a y_k points in the target mini batch. Then, they both share the same positions i in their respective minibatch. As there are several i where x_j is transported to y_k, we sum over all those possible positions. Hence our current transportation matrix coefficient $\Pi_{j,k}$ can be calculated as :

$$\Pi_{j,k} = \sum_{i=i_{\text{min}}}^{i_{\text{max}}} C_{i,x_j}^{m,n} C_{i,y_k}^{m,n} \tag{37}$$

Where $i_{\text{min}} = \max(0, m - n + j, m - n + k)$ and $i_{\text{max}} = \min(j, k)$. i_{min} and i_{max} represent the sorting constraints. Furthermore, as we have uniform weight histograms, we will transport a mass of $\frac{1}{m}$ and averaged it by the total number of transportation. So finally, our transportation matrix coefficient $\Pi_{j,k}$ are:

$$\Pi_{j,k} = \frac{1}{m\binom{n}{m}^2} \sum_{i=i_{\text{min}}}^{i_{\text{max}}} C_{i,x_j}^{m,n} C_{i,y_k}^{m,n} \tag{38}$$

B Extra experiments

In this section, we present extra experiments on the utility of using minibatch Wasserstein loss for domain adaptation, gradient flow and color transfer. We also give the algorithm which computes the barycentric mapping incrementally.

B.1 Generative models

We give implementation details of our batch Wasserstein generative models. We use a normal Gaussian noise in a latent space of dimension 10 and the generator is designed as a simple multilayer perceptron with 2 hidden layers of respectively 128 and 32 units with ReLu activation functions, and one final layer with 2 output neurons. For the different OT losses, the generator is trained with the same learning rate equal to 0.05. The optimizer is the Adam optimizer with $\beta_1 = 0$ and $\beta_2 = 0.9$. For the Sinkhorn divergence we set ε to 0.01. For WGAN and WGAN-GP we train a discriminator with the same hidden layers than the generator. We update the discriminator 5 times before one update of the generator. WGAN is trained with RMSprop optimizer and WGAN-GP with Adam optimizer ($\beta_1 = 0$, $\beta_2 = 0.9$) as done in their original papers. The learning rate is set to 10^{-4} for both. WGAN-GP has a gradient penalty parameter set to 10. All models are trained for 30000 iterations with a batch size of 100. Our minibatch OT losses use $k = 1$, which means that we compute the stochastic gradient on only one minibatch, and larger k was not needed to get meaningful results.

B.2 Domain adaptation

Domain adaptation problems consist to transfer knowledge from a source domain to a target domain. The goal is to use the labeled data in the source domain in order to classify the unlabeled data in the target domain. [Courty et al., 2017] used optimal transport to transport the source data to the target data by computing an OT map. Then they used a barycentric mapping to transport the source data to the target domain with their label. Optimal transport has been successful on this problem and we now want to study the impact of the minibatch OT losses and different OT variants.

We consider two common datasets for domain adaptation problems : MNIST [LeCun and Cortes, 2010] and USPS [Hull, 1994]. The datasets are composed of hand written digits betwenn 0 and 9. MNIST have 60000 training samples and USPS have 7291 training samples. We select 7000 samples from each dataset. The used cost for those experiments is a normalized squared euclidean cost. We want to study the number of samples which are transported on same labeled data from the source dataset to the target dataset. That is why we will study the proportion of mass between same labeled data in the transportation matrix.

The experiments use minibatch Wasserstein loss. We will use several k and m values, while for the entropic OT loss we will consider values of epsilon between 10^{-3} and 1. For each m and k, we conducted the experiments 10 times and we plot the mean and standard deviation for each m and k.

This experiment shows that considering a very small batch size hurts the number of images transported on correct labels and taking a large number of batches does not correct the performance. We also see that the number of batches k reduces the variance and should decrease when the batch size increases. Furthermore, we see that when m decreases, we have a similar performance than for the entropic OT loss with a large regularization parameter ε. We conjecture, that doing the minibatch entropic loss with a large ε parameter can lead to over regularization and can hurt the performance.

B.3 Minibatch Wasserstein gradient flow

We experimented the minibatch OT gradient flow to distributions in 2D. The purpose is to see the relevance of minibatch Wasserstein gradient flow for shape matching applications. We used the same experiments as in [Feydy et al., 2019] and relied on the geomloss package. In 2D we selected 500 data points following the image's pixel distribution. The experiments were conducted with the minibatch Wasserstein loss. We observe that we are not able to recover the target distribution, it is expected as our loss is strictly positive. However, for large enough batch size, the final distribution fits almost perfectly the target distribution and our loss leads to a good approximation.

Nevertheless we can see that taking a batch size too small results in a loss of information and drives the data toward the high density area as pointed in the 2D experiments. Regarding the number of minibatches k, it does not influence the shape of the final distribution.

Regarding the gradient flow on the celebA dataset, we now show the results when we use the minibatch Sinkhorn divergence instead of the minibatch Wasserstein distance. The minibatch Sinkhorn divergence is slower in practice than the minibatch

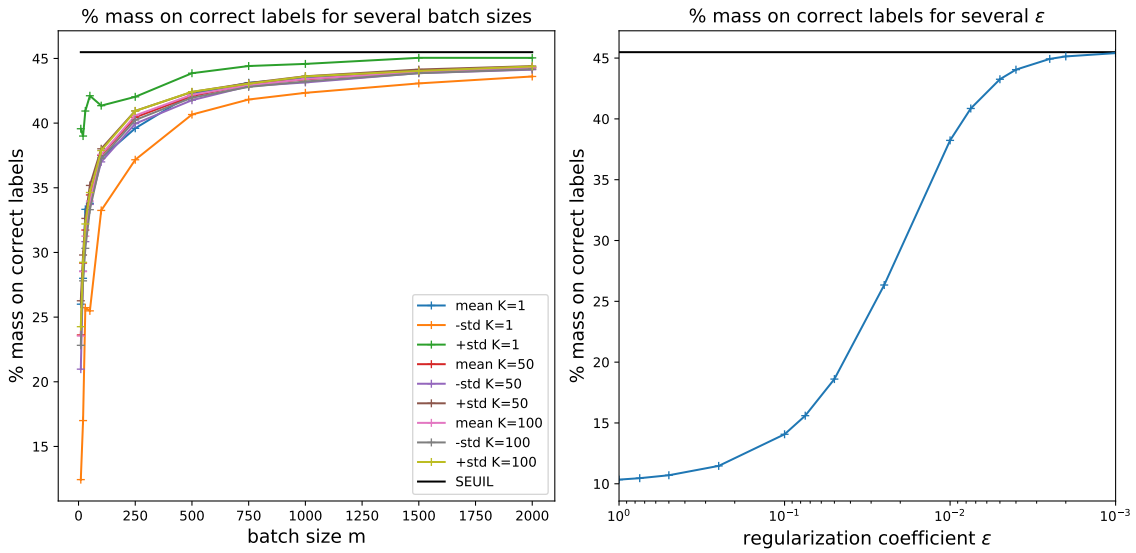

Figure 1: Proportion of correct transfered data between S/T domains for OT MB.

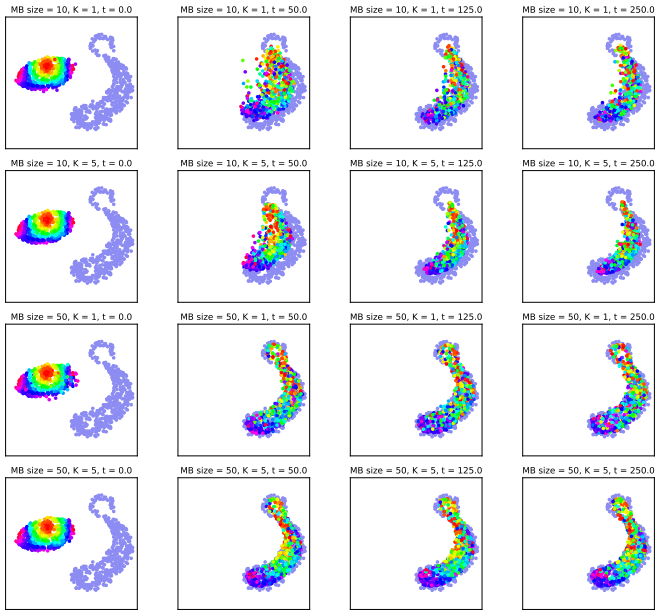

Figure 2: Gradient flow between 2D distributions for several batch sizes m and several number of batches k. The source and the target distributions have 500 samples each.

Wasserstein distance and the samples converge toward different pictures. However, we can still see a natural evolution in the images along the gradient flow.

B.4 Color transfer between subset of images

In order to present the influence of k for barycentric mapping, we present extra experiments for color transfer. We compute a k-means clustering with l clusters for each point cloud. For each image, we computed 1000 k-means clusters of the point clouds and applied the optimal transport algorithms between those subsets. We consider batch size of 10, 50 and 100. We show the color transfer for each image for $k = 5000$ and $k = 20000$ batches.

In what follows, we present the algorithm which computes the color transfer vectors incrementally without requiring the

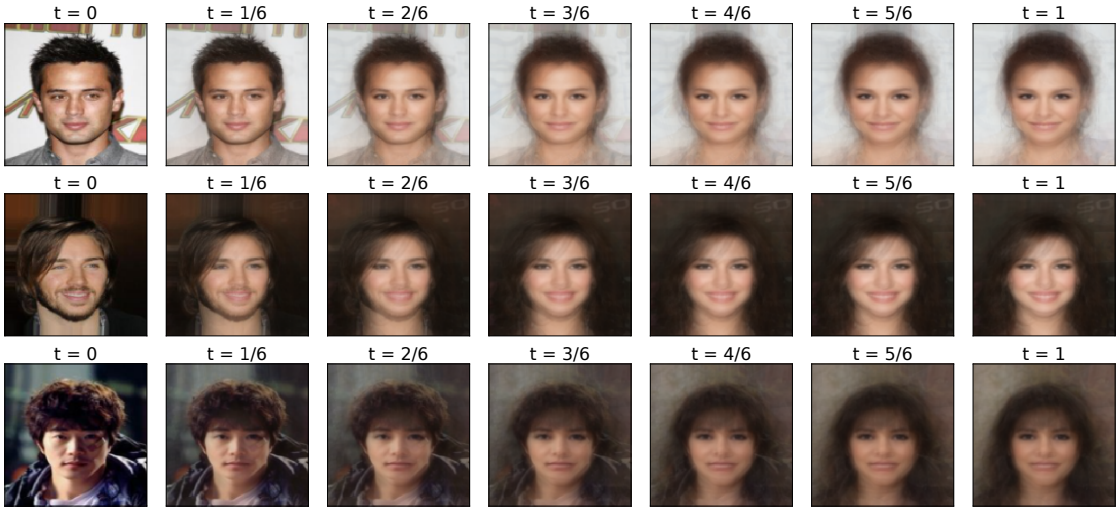

Figure 3: Gradient flow on the CelebA dataset. Source data are 5000 male images while target data are 5000 female images. The batch size m is set to 500 and the number of minibatch k is set to 10. The results were computed with the minibatch Sinkhorn divergence.

storage of the full cost matrix neither the full transportation matrix Π_k.

Algorithm 1: Computation of incremental color transfer

Inputs: m, k, source domain $\boldsymbol{X_s} \in \mathbb{R}^{n\times d}$, target domain $\boldsymbol{X_t} \in \mathbb{R}^{n\times d}$;
Results : $\boldsymbol{Y_s}$, $\boldsymbol{Y_t}$;
Initialisation : $\boldsymbol{Y_s} \in \mathbb{R}^{n\times d}$, $\boldsymbol{Y_t} \in \mathbb{R}^{n\times d}$;
for *t=1, $\cdots$, k* **do**
  Select a set A of m samples in $\boldsymbol{X_s}$;
  Select a set B of m samples in $\boldsymbol{X_t}$;
  Compute the restricted cost $C_{A,B}$;
  $G \leftarrow \underset{\Pi\in U(A,B)}{\text{argmin}} \ \langle C_{A,B}, \Pi\rangle$;
  $\boldsymbol{Y_s}\big|_A \leftarrow \boldsymbol{Y_s}\big|_A + G.\boldsymbol{X_t}\big|_B$;
  $\boldsymbol{Y_t}\big|_B \leftarrow \boldsymbol{Y_t}\big|_B + G^T.\boldsymbol{X_s}\big|_A$;
end
return $\frac{n}{k}\boldsymbol{Y_s}$, $\frac{n}{k}\boldsymbol{Y_t}$

We see that for each batch size m, when the number of batches k increases, we get better resolution for our images. It is expected as our matrix Π_k gets closer to Π_m. However, when m is small, we will need to have a large k to get good resolutions for images. We can see this phenomenon for $m = 10$, where $k = 5000$ was not enough to have a good resolution. However, $k = 5000$ was enough to get good resolutions for $m = 1000$.

References

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

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

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

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

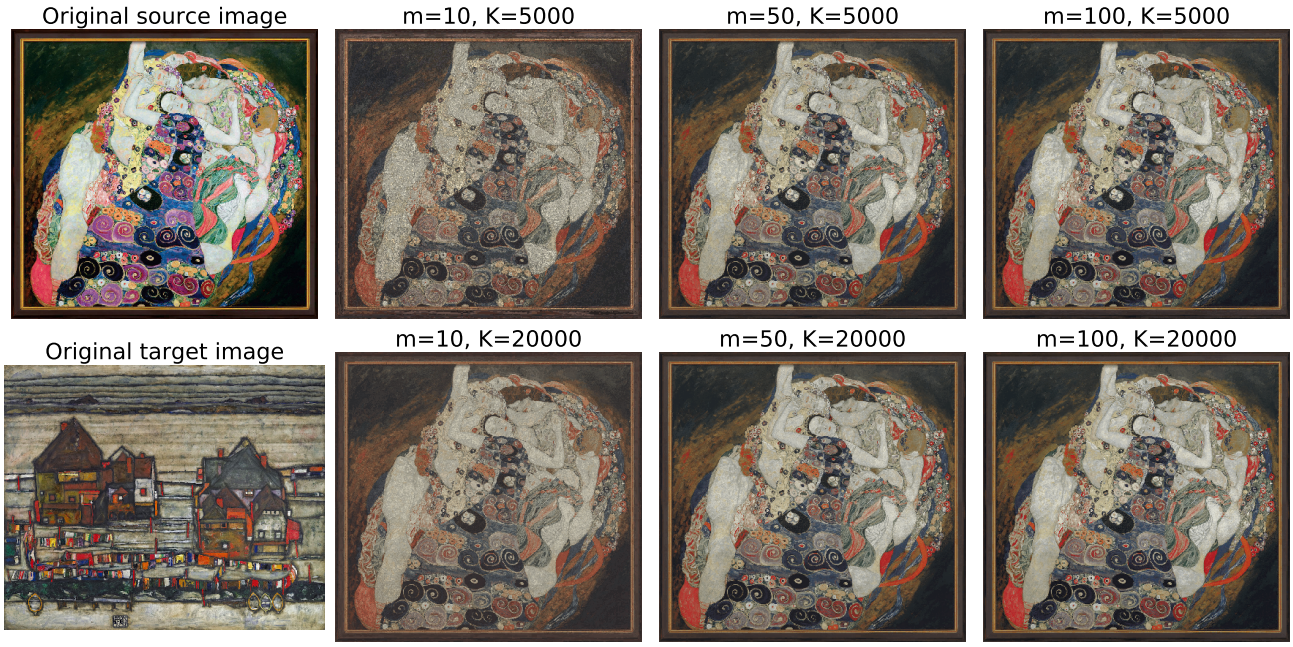

Figure 4: Color transfer from MB Wasserstein loss for several m and K. The minibatch Wasserstein distance is computed between subsets of original images.

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

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

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

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

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

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

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

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

[Courty et al., 2017] Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 7

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

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

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

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

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

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

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

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

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

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

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

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

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

[Hoeffding, 1963] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*. 4

[Hull, 1994] Hull, J. (1994). Database for handwritten text recognition research. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16. 7

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

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

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

[LeCun and Cortes, 2010] LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database. 7

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

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

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

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

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

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

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

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

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

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

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

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