Learning with minibatch Wasserstein: asymptotic and gradient properties

Quentin Bourbon[1]

[1]Master MVA
Télécom Paris

Table of Contents

Table of Contents

Presentation of the problem

Using Optimal Transport in Machine Learning problem has emerged these years because OT has a great behavior comparing to other metric, for example MMD.
However in practice, using OT distance is impossible since we are dealing with too large data and the algorithm are in $\mathcal{O}(n^3 \log(n))$ (Wasserstein distance) and $\mathcal{O}(n^2)$ (Sinkhorn divergence).

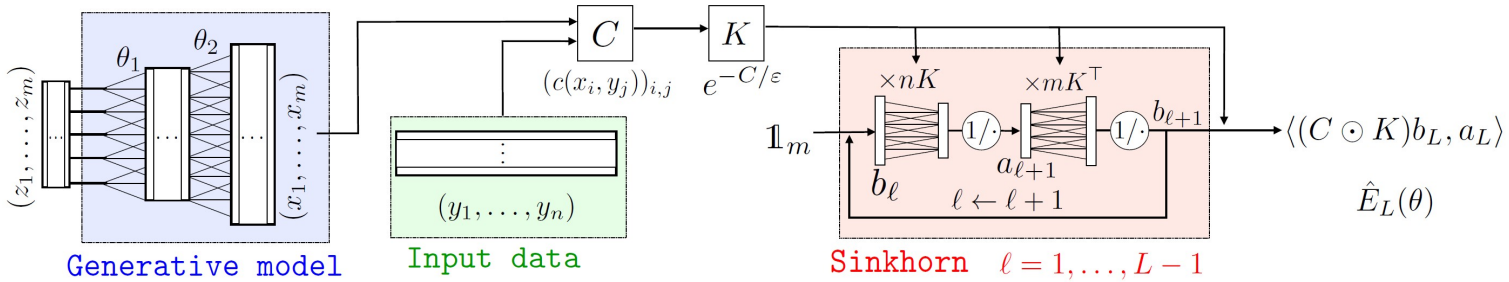


Figure: Example of a generative model using OT.

1

[1] (Genevay et al., 2017)

Table of Contents

Proposed method

To still take benefits from OT when dealing with much data we can:

- Consider a batch of data of size m with $m < n$.
- Compute the OT distance on this batch of size m.
- Repeat k times the to get an approximation of the OT quantity on all the data.

Definitions

Minibatch Wasserstein:

$$U_h(\alpha_n, \beta_n) = \binom{n}{m}^{-2} \sum_{A \in \mathcal{P}_m(\alpha_n)} \sum_{B \in \mathcal{P}_m(\beta_n)} h(A, B) \tag{1}$$

associated with the optimal transport plan $\Pi_m(\alpha_n, \beta_n)$.
But $\text{Card}(\mathcal{P}_m(\alpha_n) \times \mathcal{P}_m(\beta_n)) = \binom{n}{m}^2$, so we define the Minibatch Subsampling:

$$\tilde{U}^k_{h,m}(\alpha_n, \beta_n) = k^{-1} \sum_{(A,B) \in D_k} h(A, B) \tag{2}$$

associated with the optimal transport plan $\Pi^k_m(\alpha_n, \beta_n)$.

Table of Contents

Basic properties

- $\Pi_m(\alpha_n, \beta_n)$ is an admissible transport plan.
- To be worthwhile, (k, m) must verifies:

$$\begin{aligned} k &\leq \left(\frac{n}{m}\right)^3 \frac{\log(n)}{\log(m)} \quad \text{(Wasserstein distance)} \\ k &\leq \left(\frac{n}{m}\right)^2 \qquad\qquad\quad \text{(Sinkhorn divergence)} \end{aligned} \tag{3}$$

- With the probability at least $1 - \delta$, we sample all the points at least one time when

$$k \geq \frac{\log(\delta)}{\log(1 - \frac{m}{n})} \tag{4}$$

Convergence of our estimators

- With probability $1-\delta$,

$$|U_h^k(\alpha_n, \beta_n) - U_h(\alpha, \beta)| \leq M_h\left(\sqrt{\frac{\log(\frac{2}{\delta})}{2\mathrm{Ent}(\frac{n}{m})}} + \sqrt{\frac{2\log(\frac{2}{\delta})}{k}}\right) \tag{5}$$

  where M_h depends on h and scales at most as $\mathcal{O}(\log(m))$.
- With probability $1-\delta$,

$$|\Pi_m^k(\alpha_n, \beta_n)_{(i)}1 - \frac{1}{n}| \leq \sqrt{\frac{2\log(2/\delta)}{k}} \tag{6}$$

Gradient property

In Machine Learning, given some discrete samples from an unkwown distribution α, we want to fit a parametric model $\lambda \mapsto \beta_\lambda$ to α. Therefore we search to solve the following problem:

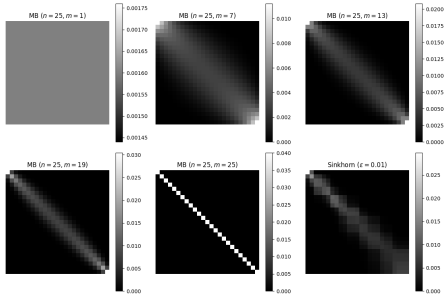
$$\min_{\lambda} U_h(\alpha_n, \beta_\lambda) \tag{7}$$

To use SGD, the gradient of U_h must be unbiased. (Fatras et al., 2021) shows this:

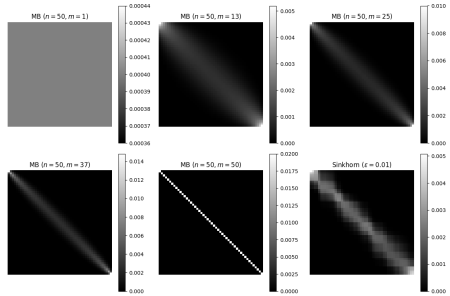
$$\nabla_\lambda \int_{X^{\otimes m}} \int_{Z^{\otimes m}} h(X, \psi_\lambda(Z))\, d\alpha^{\otimes m}(X)\, d\zeta^{\otimes m}(Z) = \int_{X^{\otimes m}} \int_{Z^{\otimes m}} \nabla_\lambda h(X, \psi_\lambda(Z))\, d\alpha^{\otimes m}(X)\, d\zeta^{\otimes m}(Z)$$

Table of Contents

Behavior of exact Minibatch Wasserstein



(a) OT matrix with $n = 25$



(b) OT matrix with $n = 50$

Figure: OT matrices between distributions in 1D depending on the number of samples n and batch size m

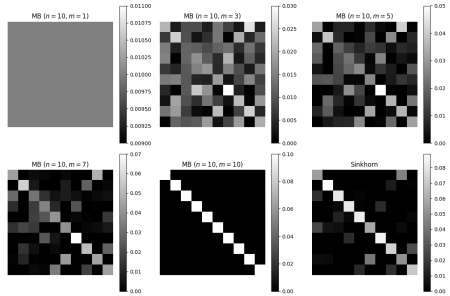


Figure: OT matrices in 2D with $n = 10$ depending on the batch size m

Admissible couple (m, k)

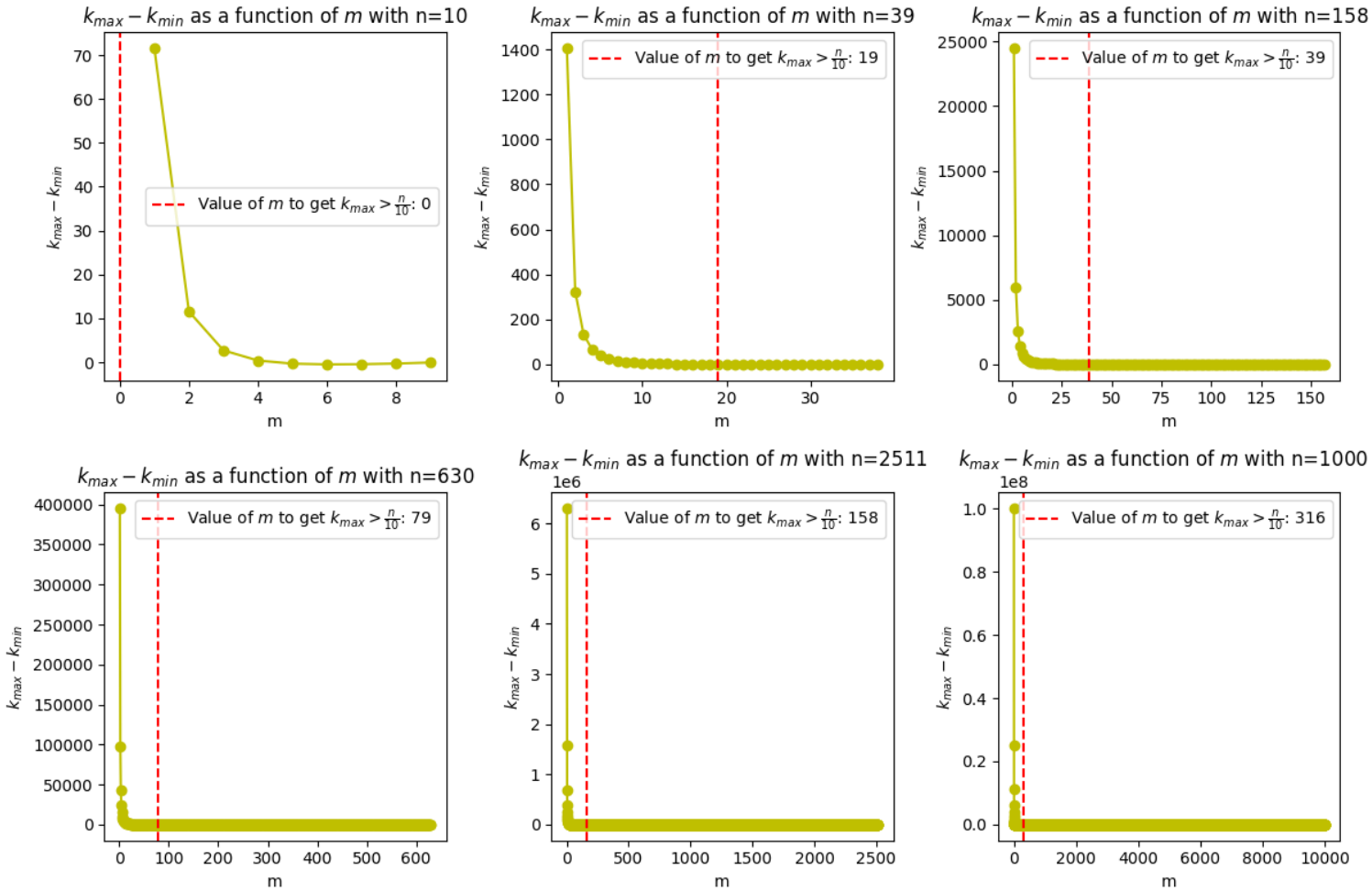
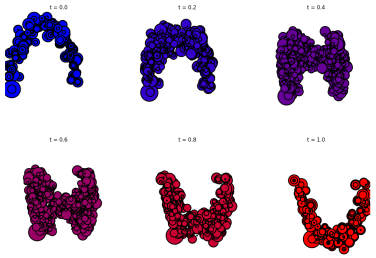
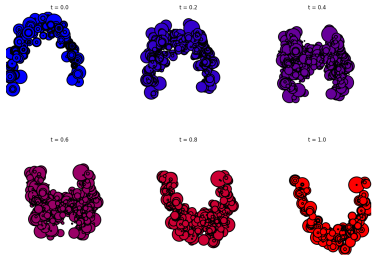


Figure: Value of $k_{\max} - k_{\min}$ with respect to m for different values of n (Sinkhorn)

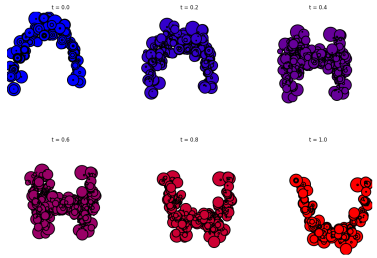
Minibatch Wasserstein Interpolation



(a) Interpolation with $m = 25,\ k = 16$

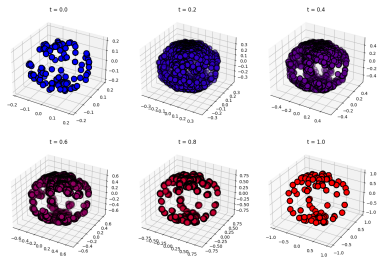


(b) Interpolation with $m = 50,\ k = 4$

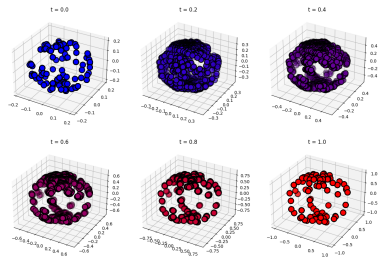


(c) Interpolation with $m = 75,\ k = 1$

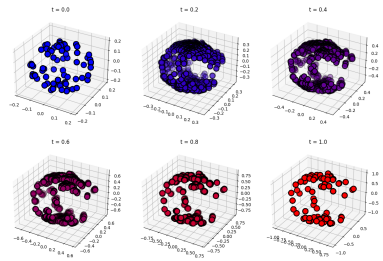
Figure: Minibatch Wasserstein Interpolation in 2D depending on the batch size m and the corresponding k



(a) Interpolation with $m = 25,\ k = 16$

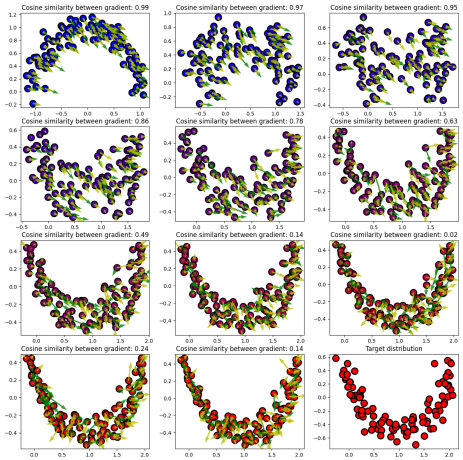


(b) Interpolation with $m = 50,\ k = 4$

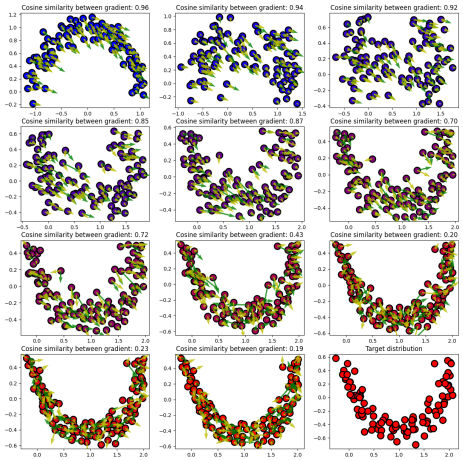


(c) Interpolation with $m = 75,\ k = 1$

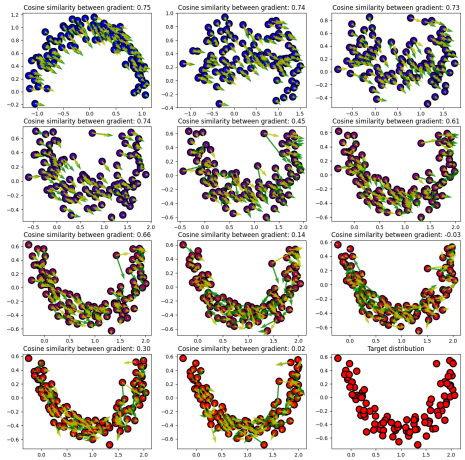
Minibatch Wasserstein gradient flow 2D



(a) Minibatch Wasserstein Flow with $m = 25,\ k = 16$



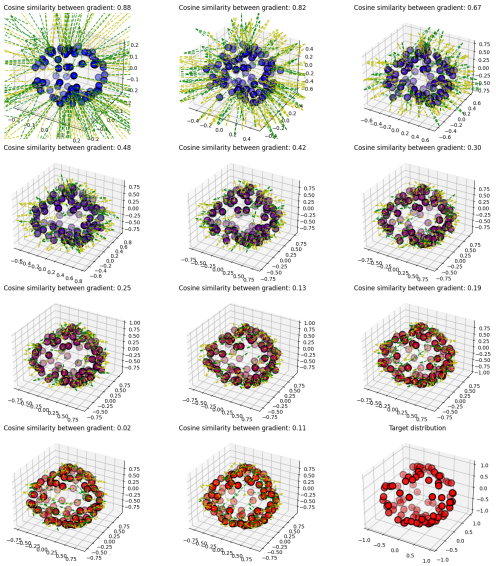
(b) Minibatch Wasserstein Flow with $m = 50,\ k = 4$



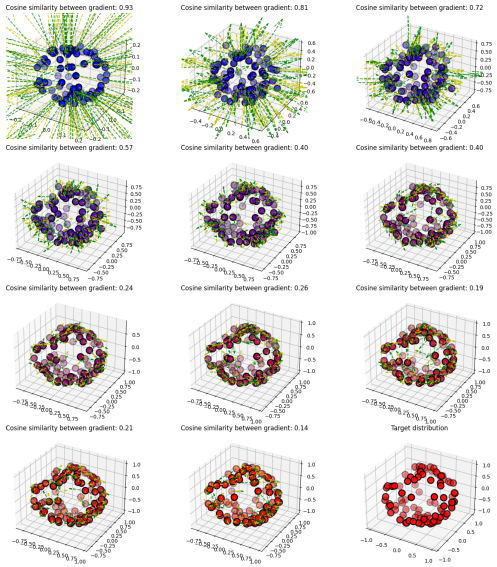
(c) Minibatch Wasserstein Flow with $m = 75,\ k = 1$

Figure: OT matrices between distributions in 2D depending on the number of samples n and batch size m

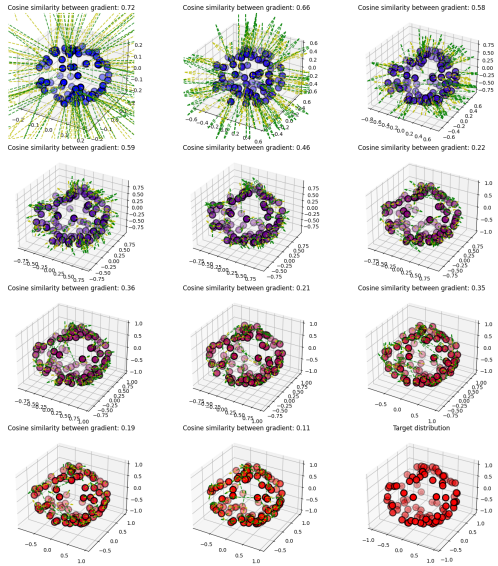
Minibatch Wasserstein gradient flow 3D



(a) Minibatch Wasserstein Flow with $m = 25,\ k = 16$



(b) Minibatch Wasserstein Flow with $m = 50,\ k = 4$



(c) Minibatch Wasserstein Flow with $m = 75,\ k = 1$

Figure: OT matrices between distributions in 3D depending on the number of samples n and batch size m

Table of Contents

Critics

- Everything is shown by considering the same number of points from α and β. We know that in this case, the solution of the Kantorovitch problem is the same that the solution of the Monge problem (the optimal transport plan is a permutation). Therefore, this is a specific case, the results may changeif we do not consider the same number of points.
- The hard problem of choosing m and k is not presented, even briefly.

Table of Contents

Conclusion

- Theoretical results on the minibatch Wasserstein.
- Minibatch Wasserstein can be seen as a regularization of the problem.
- The behavior of the estimator depending on m and k

Extension

- Develop a more complex strategy to gradient flow problem, increasing m at the end.
- Extend with the $n \neq m$. What should be the batch size between the two sets ? The same or a proportional value of n and m ?
- How to chose m and k ?
- The unbiased version of minibatch Wasserstein. What would the unbiased version do ?

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein : asymptotic and gradient properties, 2021.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences, 2017.