Learning with minibatch Wasserstein: asymptotic and gradient properties

Kilian Fatras*, Younès Zine, Rémi Flamary, Rémi Gribonval, Nicolas Courty
*Univ. Bretagne-Sud, CNRS, IRISA, INRIA

Why does optimal transport on minibtaches work?

Wasserstein distance (Peyre et al. 2019)

Wasserstein distance is defined as:

$$W_c(\alpha_n, \beta_n) = \min_{\Pi \in \boldsymbol{U}(\alpha_n, \beta_n)} \langle C, \Pi \rangle,$$

Where c is a ground cost and $\boldsymbol{U}(\alpha_n, \beta_n)$ is the set of joint probability distribution with marginals α_n and β_n.

U-statistics (Hoeffding, 1963)

Let $\underline{Z}_n := \{Z_i\}_{i=1}^n$ and $\mathcal{P}_m(\underline{Z}_n)$ the set of all m-combination. We define a one sample U-statistic of order m as:

$$U_h(\underline{Z}_n) := \binom{n}{m}^{-1} \sum_{A \in \mathcal{P}_m(\underline{Z}_n)} h(A)$$

And a subsample U-statistic as:

$$\widetilde{U}_h^k(\underline{Z}_n) := k^{-1} \sum_{A \in \mathcal{P}_m^k(\underline{Z}_n)} h(A)$$

Contributions

- Formalism and metric properties of minibatch OT.
- Concentration bounds of MB OT.
- Unbiased gradients of MB OT.
- Gradient Flow and large scale color transfer experiments.

References

Clémençon, Stephan, Igor Colin, and Aurélien Bellet (2016). "Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics". In: *Journal of Machine Learning Research*.

Ferradans, Sira et al. (2013). "Regularized Discrete Optimal Transport". In: *Scale Space and Variational Methods in Computer Vision*. Springer Berlin Heidelberg. ISBN: 978-3-642-38267-3.

Genevay, Aude, Gabriel Peyre, and Marco Cuturi (2018). "Learning Generative Models with Sinkhorn Divergences". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*.

Hoeffding, Wassily (1963). "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association*.

Peyré, Gabriel and Marco Cuturi (2019). "Computational Optimal Transport". In: *Foundations and Trends® in Machine Learning*. ISSN: 1935-8237. DOI: 10.1561/2200000073.

Problem setting and Goals

Goal: consider two distributions α, β with bounded support. We use optimal transport losses with minibatches to compare them as:

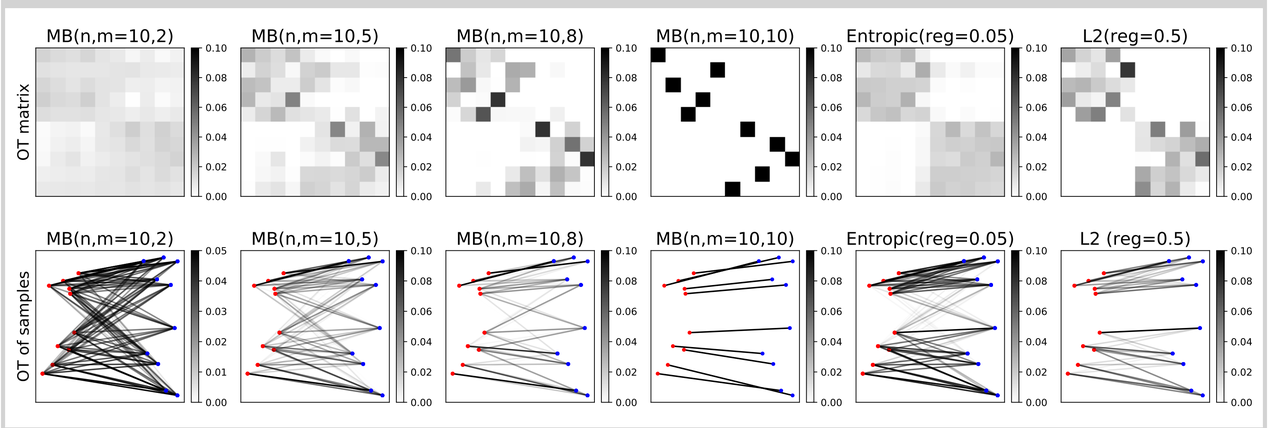
$$U_h(\alpha, \beta) := \mathbb{E}_{(X,Y) \sim \alpha^{\otimes m} \otimes \beta^{\otimes m}}[h(X, Y)] \qquad \text{(OPT)}$$

where m is the batch size.

- h is an OT loss such as (entropic-) Wasserstein distance or Sinkhorn divergence.
- Reduces complexity $\mathcal{O}(n^3) \mapsto \mathcal{O}(m^3)$
- Minibatch OT is not equivalent to minimizing original OT.
- Study of properties, statistics and gradients of minibatch OT.

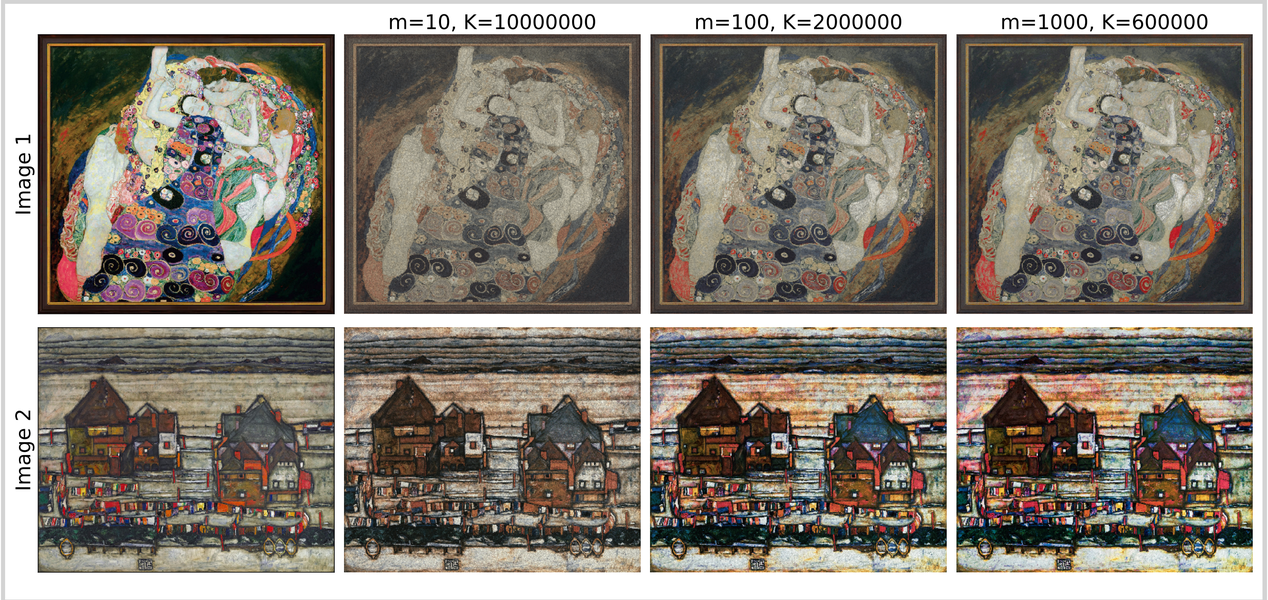
SUMMARY

- Minibatch acts like a regularization.
- ⚠ MB Wasserstein is not a ~~distance~~.
- Concentration bound.
- Unbiased estimator and gradients.



Experiments

Minibatch OT for large scale color transfer Ferradans et al. 2013. Two different images of 1M pixels and transfer their respective colors to the other image.



Estimators

Given an OT loss h, a batch size $m \leq n$ and a number of batches k, we define:

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

and a subsample quantity:

$$\widetilde{U}_h^k(\alpha_n, \beta_n) := k^{-1} \sum_{(A,B) \in \mathcal{P}_m^k(\underline{Z}_n)} h(A, B)$$

Maximal deviation bound

Theorem Let $\delta \in (0, 1)$ and consider an OT loss $h \in \{W, W_\epsilon, S_\epsilon\}$. We have with probability at least $1 - \delta$:

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

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

Unbiased gradients

Theorem Consider a $\mathcal{C}^1$ cost and assume $\lambda \mapsto Y_\lambda$ is differentiable. Then we are allowed to exchange gradients and expectation when $h \in \{W_\epsilon, S_\epsilon\}$:

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

Papers

All details can be found in our AISTATS 2020 paper and our code:

https://github.com/kilianFatras/minibatch_Wasserstein

email: kilian.fatras@irisa.fr