Minibatch Wasserstein distance

Asymptotic and gradients

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval and Nicolas Courty

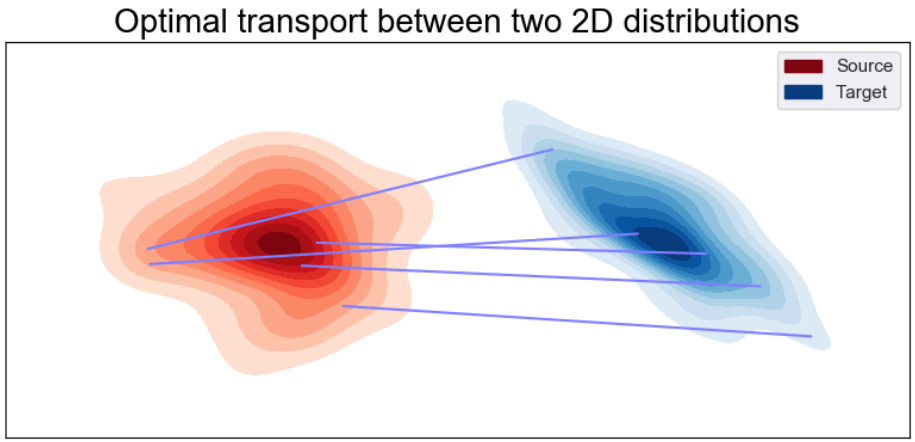
March 21, 2020

IRISA, INRIA, CNRS

Table of contents

Intro

Wasserstein distance



Optimal Transport seeks a probability coupling $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ between $\mathcal{X}$ and $\mathcal{Y}$ which minimizes a ground cost c.

$$\pi^* = \underset{\pi}{\operatorname{argmin}} \int_{\mathcal{X}\times\mathcal{Y}} c(\mathbf{x},\mathbf{y})\pi(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}$$

$$\text{s.t.} \quad \pi \in \mathcal{P} = \left\{\pi \geq \mathbf{0}, \int_{\mathcal{X}} \pi(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} = \mu_s, \int_{\mathcal{Y}} \pi(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x} = \mu_t\right\}$$

π is a joint probability measure with marginals μ_s and μ_t

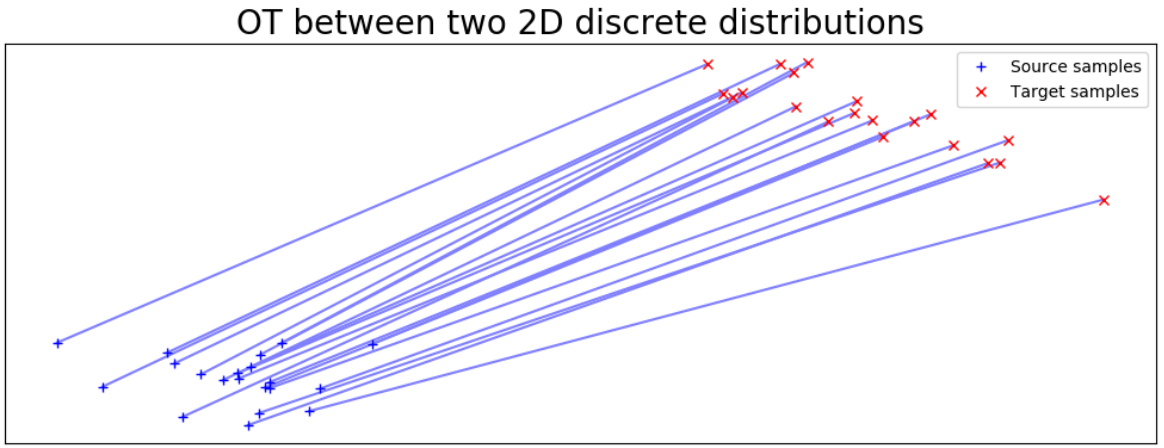
Discrete Optimal Transport

For discrete measures $\mu_s = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n} b_i \delta_{\mathbf{x}_i^t}$, OT becomes a linear program:

$$\pi^* = \underset{\pi \in \mathcal{U}(\mu_s, \mu_t)}{\operatorname{argmin}} \left\{ \langle \pi, \mathbf{C} \rangle_F = \sum_{i,j} \pi_{i,j} c_{i,j} \right\}$$

where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c\left(\mathbf{x}_i^s, \mathbf{x}_j^t\right)$ and the marginals constraints are

$$\mathcal{U}(\mu_s, \mu_t) = \left\{ \pi \in \left(\mathbb{R}^+\right)^{\mathbf{n_s} \times \mathbf{n_t}} | \pi \mathbf{1_{n_t}} = \mathbf{a}, \pi^{\mathbf{T}} \mathbf{1_{n_s}} = \mathbf{b} \right\}$$



Wasserstein distance

Definition

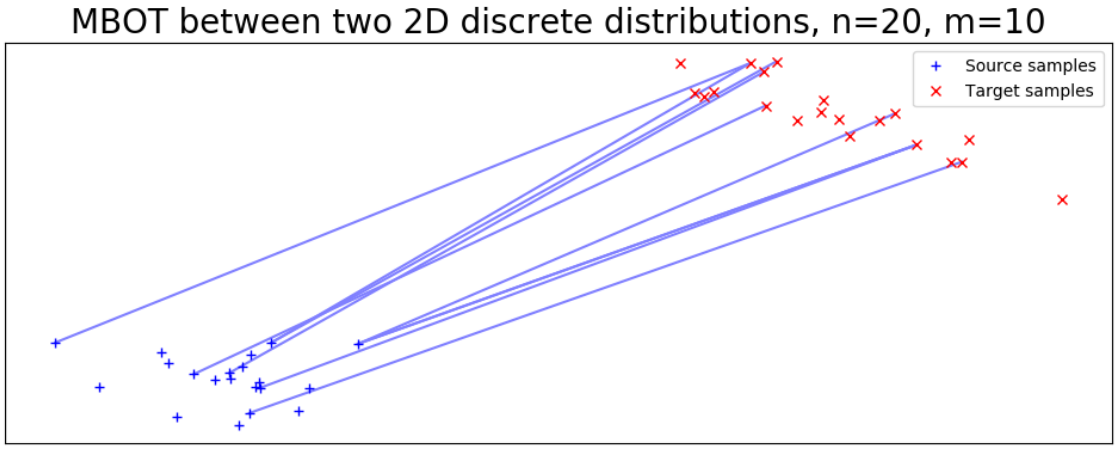
Optimal transport measures the distance between probablity distributions:

$$W_{\mathbf{C}} = \min_{\pi \in \mathcal{U}(\mu_s, \mu_t)} \langle \pi, \mathbf{C} \rangle_F$$

- If $\mathbf{C}$ is a metric then $W_{\mathbf{C}}$ becomes a metric
- A solution always exists
- Complexity of discrete OT is $\mathcal{O}(n^3)$

MBOT

Minibatch-Wasserstein distance



Let $m \leq n$, [Damodaran et al., 2018, Genevay et al., 2018] compute optimal transport between minibatch of sub-measures:

$$U_W(\alpha, \beta) := \mathbb{E}_{(X,Y)\sim\alpha^{\otimes m}\otimes\beta^{\otimes m}}[W(X, Y)]$$

- Complexity of discrete MBOT is $\mathcal{O}(m^3)$
- Left unjustified until now !

Estimate mini-batch Wasserstein distance

We define the following estimators:

Definition (Complete estimator)

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

where $\mathcal{P}_m(\alpha_n)$ is the set of all m-combination without replacement.

Definition (Incomplete estimator)

Pick an integer $k > 0$,

$$\widetilde{U}_W^k(\alpha_n, \beta_n) := k^{-1} \sum_{(A,B) \in D_k} W(A, B)$$

where D_k is a set of cardinality k whose elements are minibatches drawn at random.

Estimator properties

Proposition

The loss function $U_W(\alpha_n, \beta_n)$ *is:*

- *Unbiased estimator*
- *Strictly positive loss:* $U_W(\alpha_n, \alpha_n) > 0$
- *Not a metric*

Estimate mini-batch transportation plan

Similar estimator can be built for the mini-batch transportation plan π^*

Definition (Complete plan estimator)

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

and a subsample one:

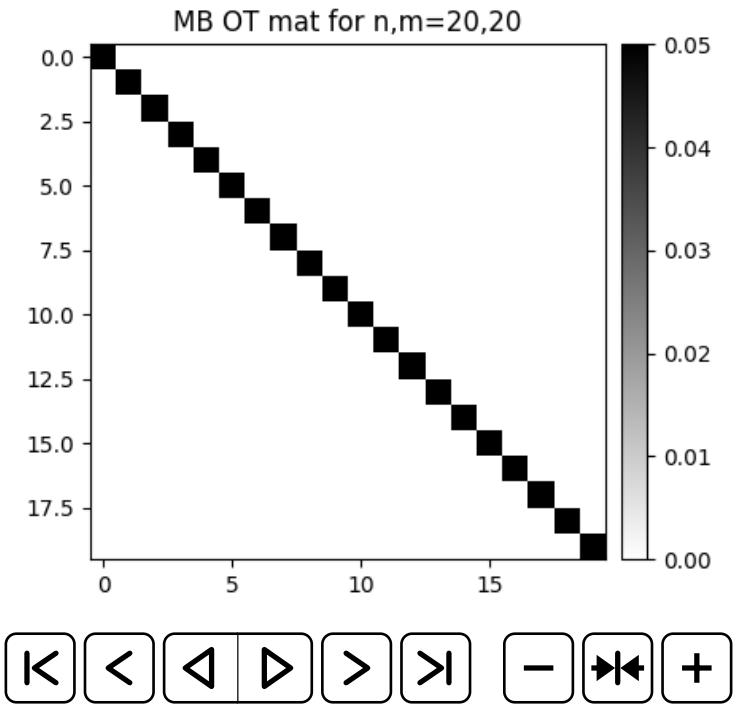
Definition (Incomplete plan estimator)

$$\Pi_k(\alpha_n, \beta_n) := k^{-1} \sum_{(A,B) \in D_k} \Pi_{A,B}$$

Proposition: Π_m is a transportation plan! (i.e. $\Pi_m \in \mathcal{U}(\mu_s, \mu_t)$)

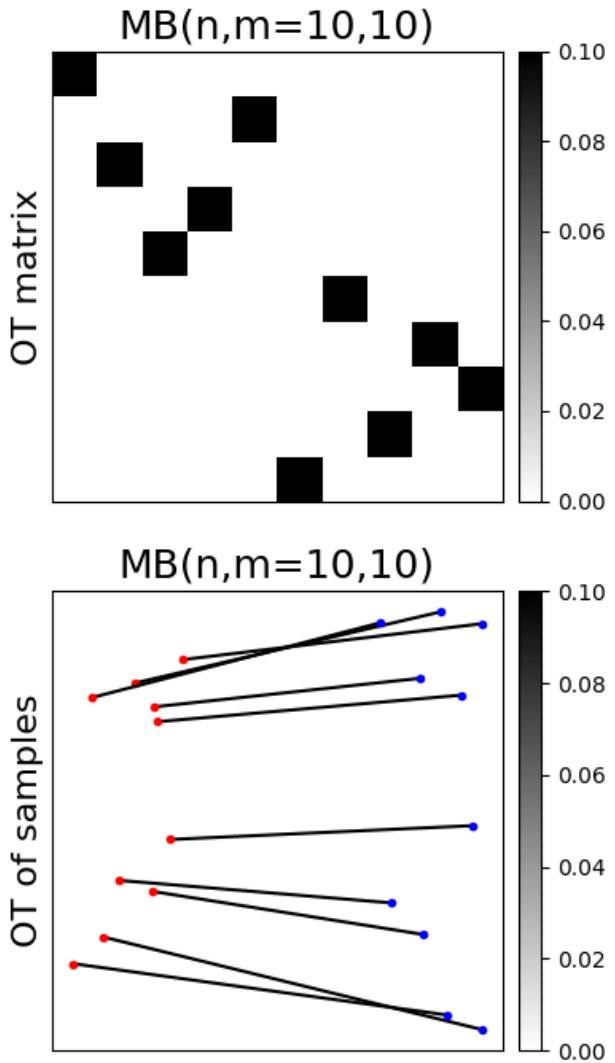
1D case

OT has a close form when measures lie in 1D space. Let us consider two 1D-measures with $n = 20$.



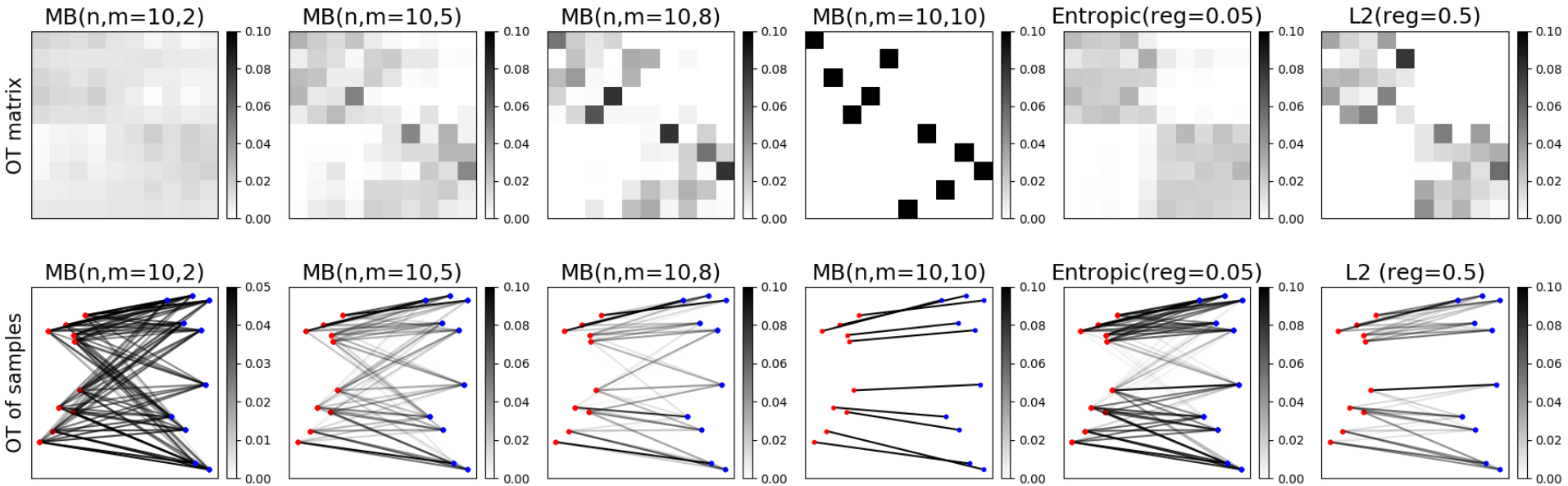
2D examples

Let us consider two 2D-measures with $n = 10$.



2D examples

Let us consider two 2D-measures with $n = 10$.



Learning properties

Asymptotic loss estimator

How far is our incomplete estimator $\widetilde{U}_h^k(\alpha_n, \beta_n)$ to the loss between full measures $U_h(\alpha, \beta)$?

Theorem (Maximal deviation bound)

Let $\delta \in (0, 1)$ and consider two compact distributions α, β. We have a deviation bound between $\widetilde{U}_h^k(\alpha_n, \beta_n)$ and $U_h(\alpha, \beta)$, with probability at least $1 - \delta$ on the draw of α_n, β_n and D_k:

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

where M_W is a constant.

Deviation transportation plan

How far is our incomplete estimator Π_k to the complete minibatch transportation plan Π_m?

Theorem (Distance to marginals)

Let $\delta \in (0,1)$, and consider two distributions α_n, β_n. For all $k \geqslant 1$, all $1 \leqslant i \leqslant n$, with probability at least $1-\delta$ on the draw of α_n, β_n and D_k we have:

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

Optimization

Can we optimize our loss with modern optimization methods such as SGD ?

Theorem (Exchange of Gradient and expectation)

Consider 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 we are allowed to exchange gradients and expectation when h is the entropic loss or the Sinkhorn divergence:

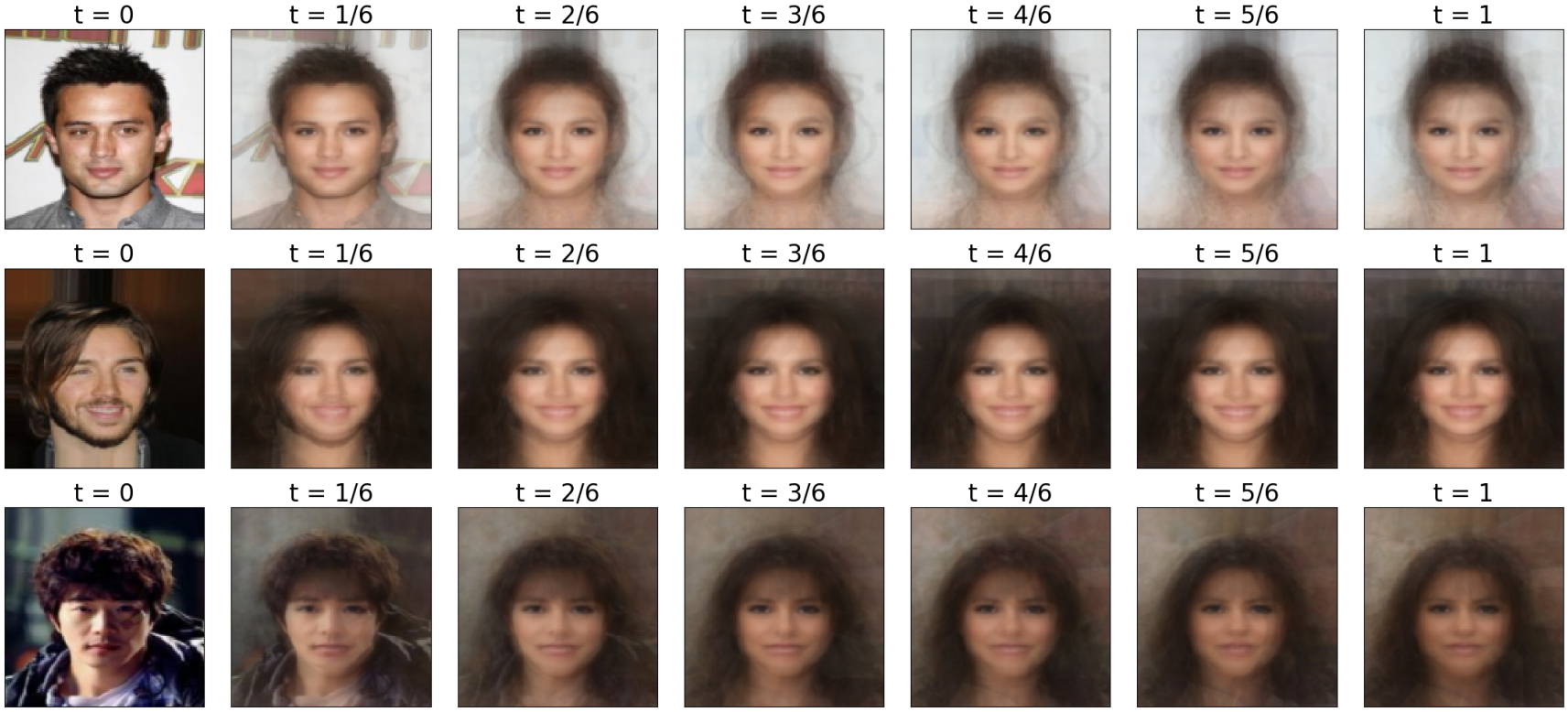
$$\nabla_\lambda \mathop{\mathbb{E}}_{Y_\lambda \sim \beta_\lambda^{\otimes m}} h(A, Y_\lambda) = \mathop{\mathbb{E}}_{Y_\lambda \sim \beta_\lambda^{\otimes m}} \nabla_\lambda h(A, Y_\lambda) \tag{2}$$

Experiments

Gradient flow

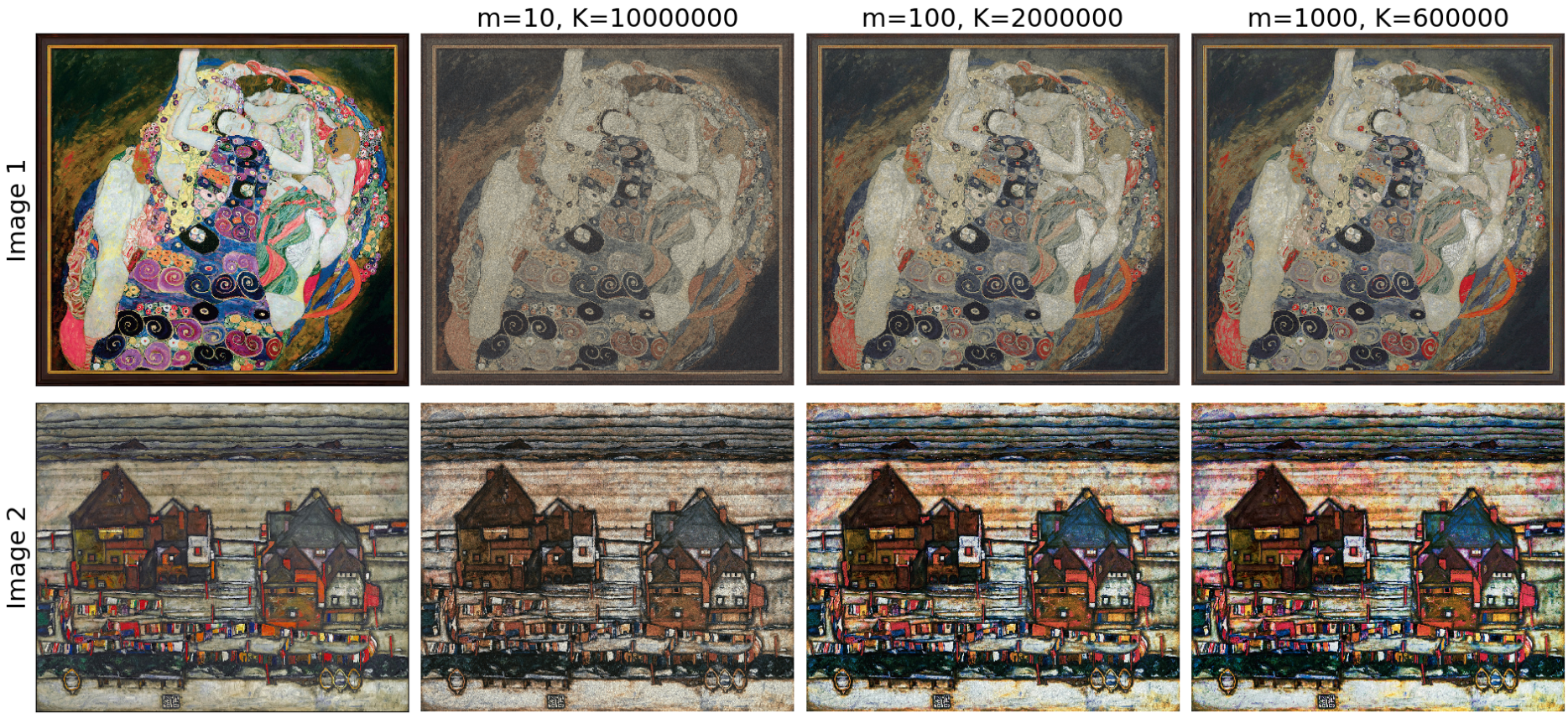
Minibatch Wasserstein gradient flow between male and female images.

$$\dot{\boldsymbol{x}}(t) = -m\nabla_{\boldsymbol{x}}\left[\widetilde{U}_h^k(\alpha_n, \beta_n)\right](\boldsymbol{x}(t))$$



Color transfer

Transfert of color between two input images (1M pixels) using Π_k.



Paper

Full details in our AISTATS 2020 paper !

Check it out : https://arxiv.org/abs/1910.04091

References i

[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.

[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.

[Fatras et al., pear] Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. (2020 (To appear)).
Learning with minibatch wasserstein: asymptotic and gradient properties.
In *AISTATS*.

[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.

References ii

[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*.

[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*.