Stochastic Optimization for Regularized Wasserstein Estimators

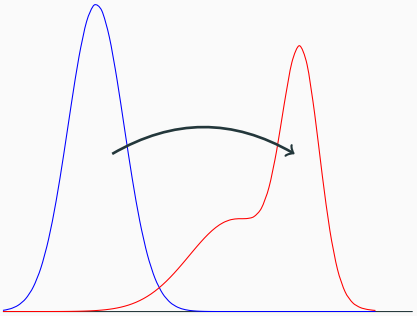
ICML 2020

Marin Ballu

Quentin Berthet

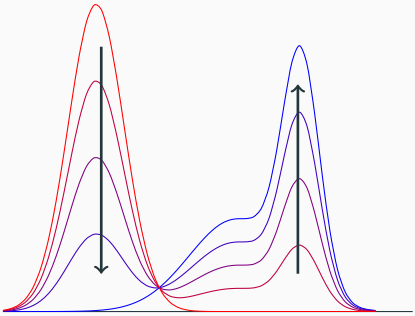
Francis Bach

Wasserstein Distance: a natural geometry for distributions



How does one compute the distance between two data distributions?

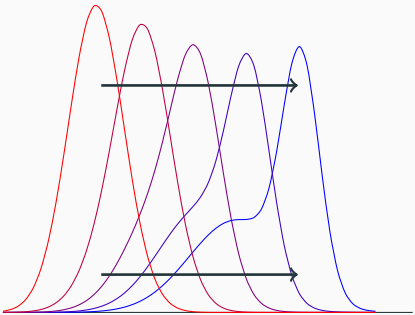
Wasserstein Distance: a natural geometry for distributions



How does one compute the distance between two data distributions?

- Relative entropy and other f-divergences allow classical statistical approaches.

Wasserstein Distance: a natural geometry for distributions

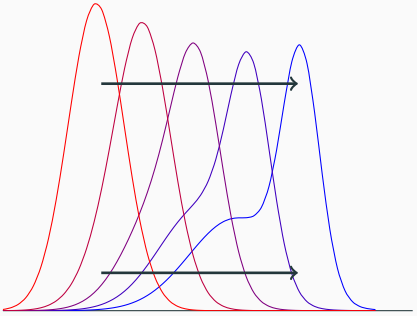


How does one compute the distance between two data distributions?

- Relative entropy and other f-divergences allow classical statistical approaches.
- Optimal transport theory allows us to capture the geometry of the data distributions, with the *Wasserstein distance*.

$$W_c(\mu, \nu) = \mathrm{OT}(\mu, \nu) = \min_{T_{\#}\mu=\nu} \mathbb{E}_{X\sim\mu}\left[c(X, T(X))\right]$$

Wasserstein Distance: a natural geometry for distributions

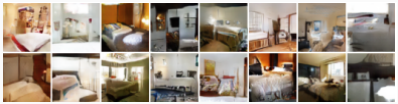


How does one compute the distance between two data distributions?

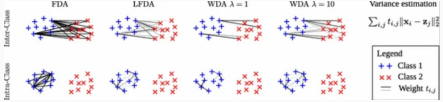
- Relative entropy and other f-divergences allow classical statistical approaches.
- Optimal transport theory allows us to capture the geometry of the data distributions, with the *Wasserstein distance*.

$$W_c(\mu,\nu) = \mathrm{OT}(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\left[c(X,Y)\right]$$

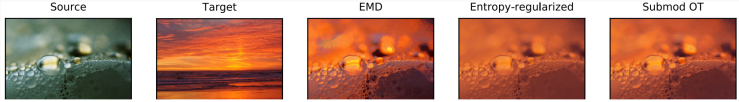
Wasserstein distance in machine learning



Wasserstein GAN (Arjovsky et al., 2017)

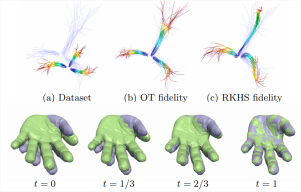


Wasserstein Discriminant Analysis (Flamary et al., 2018)

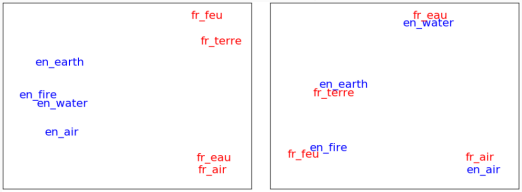


Clustered point-matching (Alvarez-Melis et al., 2018)

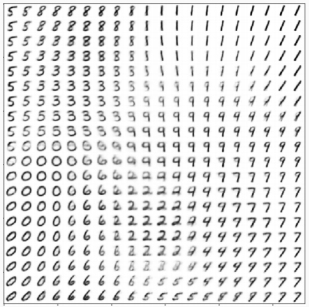
Wasserstein distance in machine learning



Diffeomorphic registration (Feydy et al., 2017)



Alignment of embeddings (Grave et al., 2019)



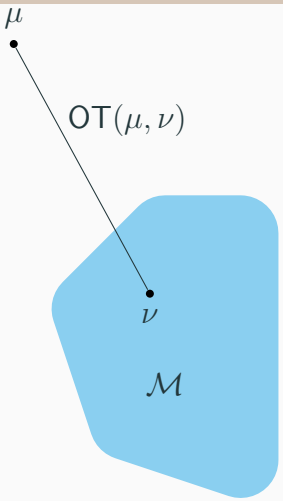
Sinkhorn divergence for generative models (Genevay et al., 2019)

Our contribution

We consider the minimum Kantorovich estimator (Bassetti et al., 2006), or *Wasserstein estimator* of the measure μ:

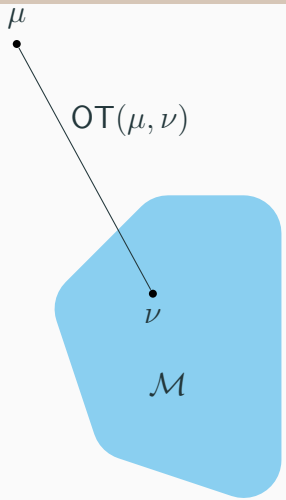
$$\min_{\nu \in \mathcal{M}} \mathrm{OT}(\mu, \nu)\,,$$

which is often used for $\mu = \sum_i \delta_{x_i}$ to fit a parametric model $\mathcal{M}$ (as with MLE, where KL divergence replaces OT).



Our contribution

- We add two layers of entropic regularization.
- We propose a new stochastic optimization scheme to minimize the regularized problem.
- Time per step is sublinear in the natural dimension of the problem.
- We provide theoretical guarantees, and simulations.



Regularized Wasserstein Distance

Wasserstein distance

$$W_c(\mu, \nu) = \mathrm{OT}(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\left[c(X, Y)\right]$$

Regularized Wasserstein Distance

Wasserstein distance

$$W_c(\mu, \nu) = \mathrm{OT}(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\left[c(X, Y)\right]$$

Regularized Wasserstein distance

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

Computed at lightspeed by Sinkhorn algorithm (Cuturi 2013)

SGD on dual problem (Genevay et al. 2016)

Regularized Wasserstein Estimator

Wasserstein estimator

$$\min_{\nu \in \mathcal{M}} \mathsf{OT}(\mu, \nu)$$

Regularized Wasserstein Estimator

Wasserstein estimator

$$\min_{\nu \in \mathcal{M}} \mathsf{OT}(\mu, \nu)$$

First layer of regularization

$$\min_{\nu \in \mathcal{M}} \mathsf{OT}_{\varepsilon}(\mu, \nu)$$

Regularized Wasserstein Estimator

Wasserstein estimator

$$\min_{\nu \in \mathcal{M}} \mathsf{OT}(\mu, \nu)$$

First layer of regularization

$$\min_{\nu \in \mathcal{M}} \mathsf{OT}_{\varepsilon}(\mu, \nu)$$

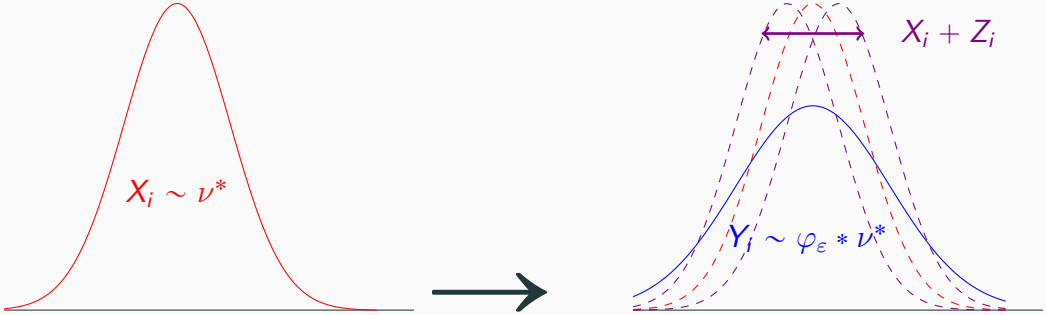
Second layer of regularization

$$\boxed{\min_{\nu \in \mathcal{M}} \mathsf{OT}_{\varepsilon}(\mu, \nu) + \eta\, \mathsf{KL}(\nu, \beta)}$$

First layer: Gaussian deconvolution

This is a recent interpretation (Rigollet, Weed 2018).

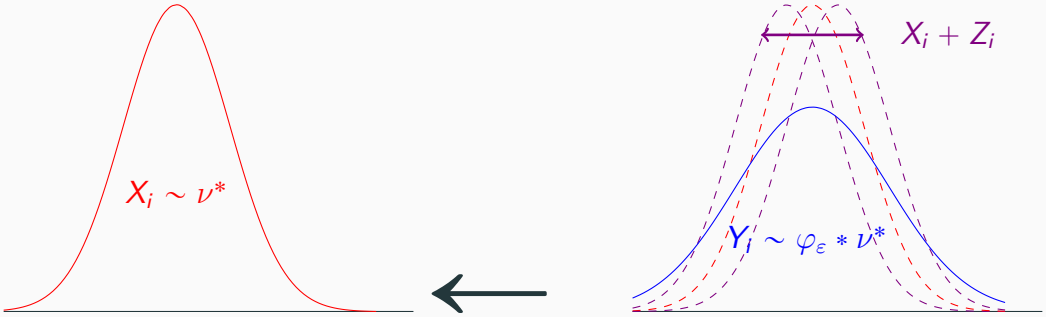
Let X_i be iid random variables following ν^*, $Z_i \sim \varphi_\varepsilon = \mathcal{N}(0, \varepsilon \mathrm{Id})$ an iid gaussian noise and $Y_i = X_i + Z_i$ the perturbed observation with distribution μ.



First layer: Gaussian deconvolution

For $c(x, y) = \|x - y\|^2$, the MLE for ν^* is

$$\hat{\nu} := \arg\max_{\nu \in \mathcal{M}} \sum_i \log(\varphi_\varepsilon * \nu)(X_i) \Leftrightarrow \hat{\nu} = \arg\min_{\nu \in \mathcal{M}} \mathrm{OT}_\varepsilon(\mu, \nu).$$



First layer: adds entropy to the transport matrix

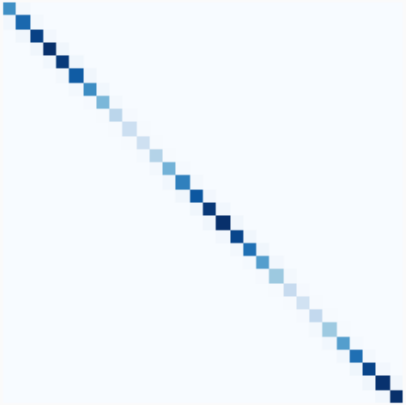


Figure 1: Small regularization $\varepsilon = 0.01$

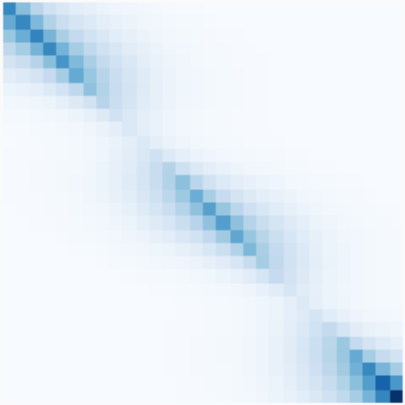


Figure 2: Big regularization $\varepsilon = 0.1$

Second Layer: Interpolation with likelihood estimators

Wasserstein Estimator

$$\min_{\nu \in \mathcal{M}} \mathsf{OT}(\mu, \nu)$$

Maximum Likelihood Estimator

$$\min_{\nu \in \mathcal{M}} \mathsf{KL}(\nu, \beta)$$

Regularized Wasserstein Estimator

$$\min_{\nu \in \mathcal{M}} \mathsf{OT}_{\varepsilon}(\mu, \nu) + \eta\, \mathsf{KL}(\nu, \beta)$$

Second Layer: adds entropy to the target measure

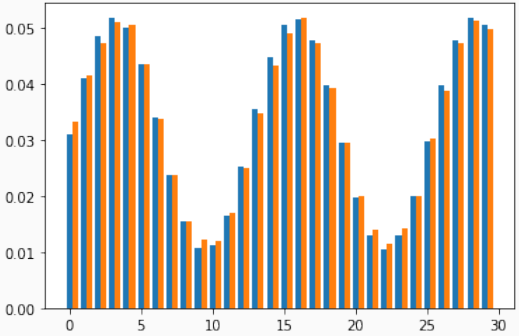


Figure 3: Small regularization $\eta = 0.02$

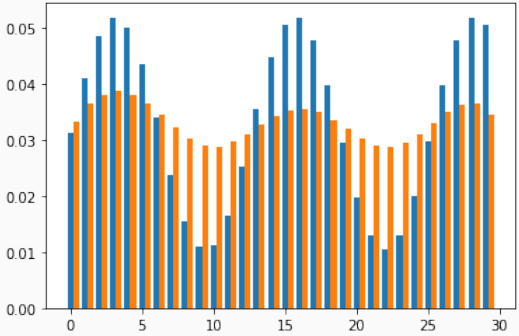


Figure 4: Big regularization $\eta = 0.2$

Dual Formulation of the problem

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

with

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

is

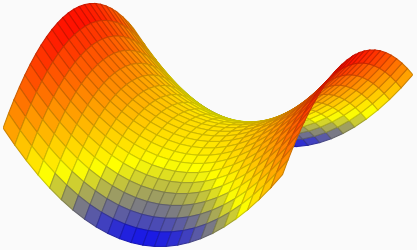
$$\min_{\nu \in \mathcal{M}} \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\left[c(X, Y)\right] + \varepsilon\, \mathrm{KL}(\pi, \mu \otimes \nu) + \eta\, \mathrm{KL}(\nu, \beta).$$

We consider the dual of the second min.

Dual Formulation

The dual problem can be written as a saddle point problem, where the min and the max can be swapped. The final formulation is of the form

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



Properties of the function F in the discrete case

1. F is λ-strongly convex on the hyperplane $E = \{\sum_i \mu_i a_i = \sum_j \beta_j b_j\}$.
2. There exists a solution of $\boxed{\max_{(a,b)\in\mathbb{R}^I\times\mathbb{R}^J} F(a,b),}$ which is in E, and it is unique.
3. The gradients of F can be written as expectations

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

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

Stochastic Gradient Descent

We have stochastic gradients for F

$$G_a = (1 - D_{i,j})e_i$$
$$G_b = (f_j - D_{i,j})e_j.$$

SGD algorithm:

- Sample $i \in \{1, \ldots, I\}$ with probability μ_i,
- Sample $j \in \{1, \ldots, J\}$ with probability β_j,
- Compute G_a and G_b
- $a \leftarrow a + \gamma_t G_a$,
- $b \leftarrow b + \gamma_t G_b$.

Stochastic Gradient Descent

We only have to compute a and b one coefficient at a time

- Sample $i \in \{1, \dots, I\}$ with probability μ_i,
- Sample $j \in \{1, \dots, J\}$ with probability β_j,
- Compute f_j and $D_{i,j}$
- $a_i \leftarrow a_i + \gamma_t(1 - D_{i,j})$,
- $b_j \leftarrow b_j + \gamma_t(f_j - D_{i,j})$.

The sum memorization trick

The computation of $D_{i,j}(a,b) = \exp\left(\frac{a_i + b_j - C_{i,j}}{\varepsilon}\right)$ and $f_j = \frac{\nu_j(b)}{\beta_j}$ is $O(1)$.

However

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

but we can do it in $O(1)$ if we update

$$S^{(t)} = \sum_k \beta_k e^{-b_k^{(t)}/(\eta-\varepsilon)},$$

with

$$S^{(t+1)} = S^{(t)} + \beta_j e^{-b_j^{(t+1)}/(\eta-\varepsilon)} - \beta_j e^{-b_j^{(t)}/(\eta-\varepsilon)}.$$

Convergence Bounds

With stepsize $\gamma_t = \frac{1}{\lambda t}$, the estimator verifies

$$\mathbb{E}\left[\mathrm{KL}(\nu^*, \nu^t)\right] \leqslant \frac{C_1}{(\eta - \varepsilon)\lambda^2} \frac{1 + \log t}{t}.$$

With stepsize $\gamma_t = \frac{C_2}{\sqrt{t}}$, the estimator verifies the following bound:

$$\mathbb{E}\left[\mathrm{KL}(\nu^*, \nu^t)\right] \leqslant \frac{C_3}{(\eta - \varepsilon)\lambda} \frac{2 + \log t}{\sqrt{t}}.$$

Simulations

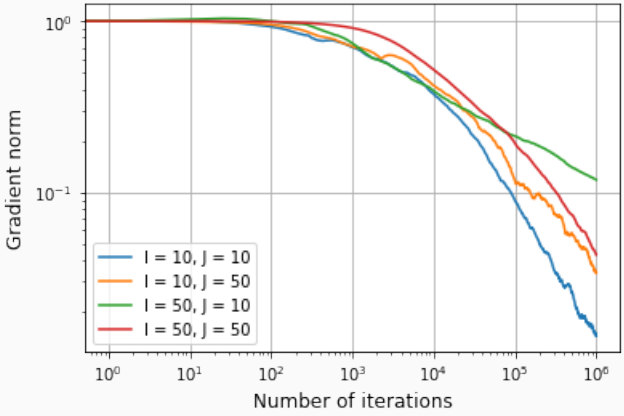


Figure 5: Convergence of the gradient norm for different dimensions.

Using for Wasserstein Barycenters

Wasserstein barycenter

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

Doubly regularized Wasserstein barycenter

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

Conclusion

Takeaways:

- Wasserstein estimators are "projections" according to Wasserstein distances,
- Two layers of entropic regularization are used here,
- It is then possible to compute stochastic gradients in $O(1)$ for this problem,
- The results are also valid for Wasserstein barycenters.

Thank you for your attention!