Learning with the Sinkhorn Loss

Aude Genevay

CEREMADE - Université Paris Dauphine
INRIA - Mokaplan project-team
DMA - Ecole Normale Supérieure

Modern Mathematical Methods for Data Analysis
Liège - June 2018

Joint work with M.Cuturi and G. Peyré

Recurrent issue in ML : Fitting data to a probabilistic model

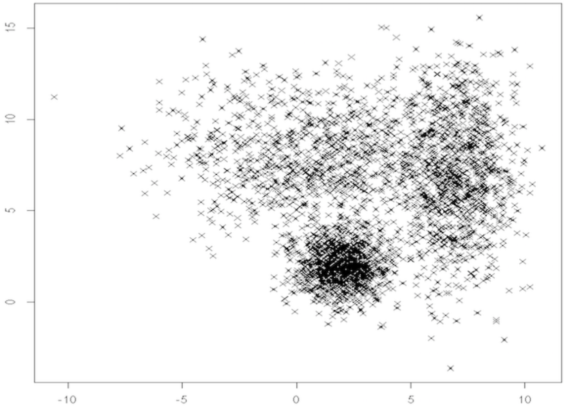


Figure 1: Data points in 2D

Recurrent issue in ML : Fitting data to a probabilistic model

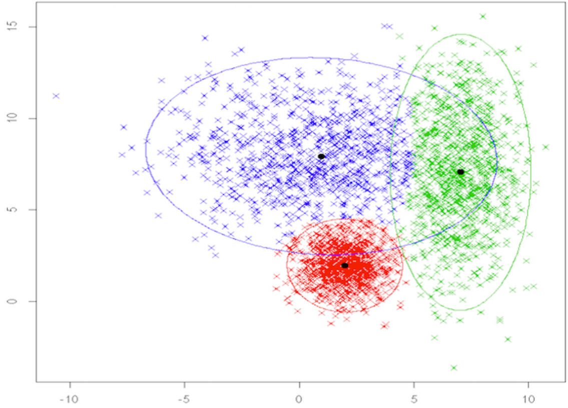


Figure 2: Gaussian Mixture Model

Density Fitting with MLE

- Observed dataset $(y_1, \ldots, y_n) \in \mathcal{X}$ (IID assumption)
- Empirical measure $\hat{\nu} = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$
- Parametric model $(\mu_\theta)_{\theta \in \Theta}$ measure with density $(f_\theta)_{\theta \in \Theta}$
- Goal : find $\hat{\theta} = \arg\min_{\theta \in \Theta} \mathcal{L}(\mu_\theta, \hat{\nu})$ where $\mathcal{L}$ is a loss on measures.
- **Maximum Likelikood Estimator**

$$\hat{\theta} \stackrel{\text{def.}}{=} \arg\min_{\theta \in \Theta} - \sum_{i=1}^{n} \log f(y_i \mid \theta)$$

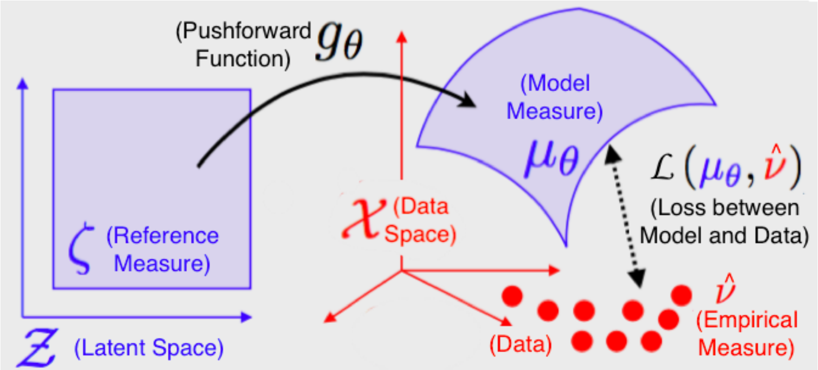


Figure 3: Illustration of Density Fitting on a Generative Model

Density Fitting for Generative Models I

- Parametric model : $\mu_\theta = g_{\theta\sharp}\zeta$
- ζ reference measure on (low dimensional) latent space $\mathcal{Z}$
- $g_\theta : \mathcal{Z} \to \mathcal{X}$ from latent space to data space
- Sampling procedure : $x \sim \mu_\theta$ obtained by $x = g_\theta(z)$ were $z \sim \zeta$
- Very popular topic in ML : image generation

Density Fitting for Generative Models II

- Generative Models usually supported on low dimensional manifolds (dim $\mathcal{Z} <$ dim $\mathcal{X}$)
- μ_θ doesn't have density wrt Lebesgue measure

$\Rightarrow$ **MLE can't be applied in this context!**

- 2 natural candidates emerge for $\mathcal{L}$
  - Maximum Mean Discrepency (based on Reproducing Kernel Hilbert Spaces) $\rightarrow$ Hilbertian norm
  - The Wasserstein Distance (based on Optimal Transport) $\rightarrow$ Non-Hilbertian distance

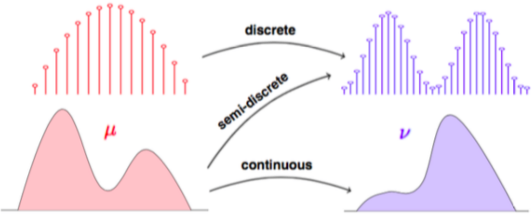
Maximum Mean Discrepency
Gretton et al. '12

- Consider Reproducing Kernel Hilbert Space $\mathcal{H}$ with kernel k
- $f \in \mathcal{H} \Rightarrow f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$

$$\begin{aligned} MMD_k(\mu, \nu) &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{\mu}[f(x)] - \mathbb{E}_{\nu}[f(y)] \\ &= \mathbb{E}_{\mu\otimes\mu}[k(x, x')] + \mathbb{E}_{\nu\otimes\nu}[k(y, y')] \\ &\quad -2\mathbb{E}_{\mu\otimes\nu}[k(x, y)] \end{aligned}$$

- Usual (positive definite) kernels
  - Gaussian kernel : $k(x, y) = \exp(\frac{\|x-y\|^2}{\sigma})$
  - Energy distance kernel : $k(x, y) = d(x, 0) + d(y, 0) - d(x, y)$

Optimal Transport I



- Optimal Transport : find coupling that minimizes total cost of moving μ to ν whith unit cost function **c**
- Constrained problem : coupling has fixed marginals
- Minimal cost of moving μ to ν(e.g. solution of the OT problem) is called the **Wasserstein distance** (it's an actual distance!)

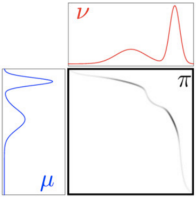
Optimal Transport II

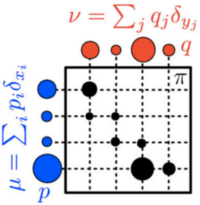
Cost $c(x, y)$ to move a unit of mass from x to y

Constrained set of couplings $\Pi(\mu, \nu)$ with marginals μ and ν

$$W(\mu, \nu) = \min_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y)$$

What's the coupling that minimizes the total cost?





Optimal Transport III

Main issues of Wasserstein distance :

- Computationally Expensive : need to solve LP (in discrete case)
- Poor Sample Complexity : $W(\mu, \hat{\mu}_n) \sim n^{-\frac{1}{d}}$
  $\rightarrow$ scales exponentially with dimension
  $\rightarrow$ need a lot of samples to get a good approximation of W

Entropy!

- Basically : Adding an entropic regularization smoothes the constraint
- Makes the problem easier :
  - yields an unconstrained dual problem
  - discrete case can be solved efficiently with alternate maximizations on the dual variables : Sinkhorn's algorithm (more on that later)
- For ML applications, regularized Wasserstein is better than standard one
- In high dimension, helps avoiding overfitting

Entropic Relaxation of OT
Cuturi '13

Add entropic Penalty to Kantorovitch formulation of OT

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\mu\otimes\nu) \qquad (\mathcal{P}_\varepsilon)$$

where

$$\mathrm{KL}(\pi|\mu\otimes\nu) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \big(\log\big(\frac{\mathrm{d}\pi}{\mathrm{d}\mu\mathrm{d}\nu}(x,y)\big) - 1\big)\mathrm{d}\pi(x,y)$$

Regularized loss :

$$W_{c,\varepsilon}(\mu,\nu) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi_\varepsilon(x,y)$$

where π_ε solution of $(\mathcal{P}_\varepsilon)$

Sinkhorn Divergences : interpolation between OT and MMD

Theorem

The Sinkhorn loss between two measures μ, ν is defined as:

$$\bar{W}_{c,\varepsilon}(\mu, \nu) = 2W_{c,\varepsilon}(\mu, \nu) - W_{c,\varepsilon}(\mu, \mu) - W_{c,\varepsilon}(\nu, \nu)$$

with the following limiting behavior in ε:

1. *as* $\varepsilon \to 0, \quad \bar{W}_{c,\varepsilon}(\mu, \nu) \to 2W_c(\mu, \nu)$
2. *as* $\varepsilon \to +\infty, \quad \bar{W}_{c,\varepsilon}(\mu, \nu) \to MMD_{-c}(\mu, \nu)$

Remark : Some conditions are required on c to get MMD distance when $\varepsilon \to \infty$. In particular, $c = \|\cdot\|_p, 0 < p < 2$ is valid.

Sample Complexity

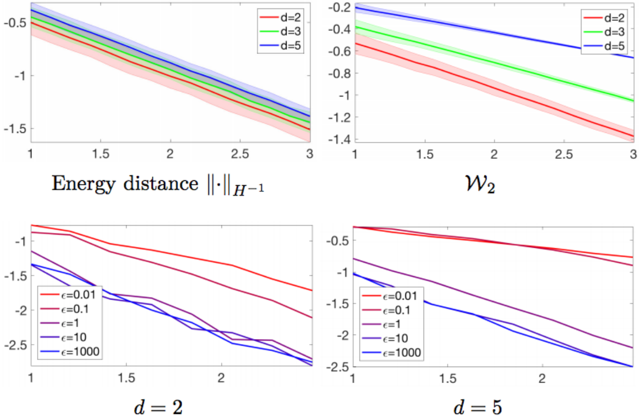
Sample Complexity of OT and MMD

Let μ a probability distribution on $\mathbb{R}^d$, and $\hat{\mu}_n$ an empirical measure from μ

$$
\begin{aligned}
W_c(\mu, \hat{\mu}_n) &= O(n^{-1/d}) \\
MMD(\mu, \hat{\mu}_n) &= O(n^{-1/2})
\end{aligned}
$$

$\Rightarrow$ the number n of samples you need to get a precision η on the Wassertein distance grows exponentially with the dimension d of the space!

Sample Complexity - Sinkhorn loss



Sample Complexity of Sinkhorn loss seems to improve as ε grows.

Plots courtesy of G. Peyré and M. Cuturi

Sample Complexity - Sinkhorn loss

Sample Complexity of Sinkhorn loss (conjecture)

Let μ, ν two probability distributions on $\mathbb{R}^d$, and $\hat{\mu}_n$, $\hat{\nu}_n$ their empirical measures

$$W_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - W_{c,\varepsilon}(\mu, \nu) \quad = \quad O(\varepsilon^{-d/2} n^{-1/2})$$

$\Rightarrow$ The $n^{-1/2}$ is obtained by proving that regularized potentials belong to a RKHS (Sobolev space W_s^2 with $s > \frac{d}{2}$)
$\Rightarrow$ Dependence on ε has to be confirmed - currently working on those bounds!

Density Fitting with Sinkhorn loss "Formally"

Solve $\min_\theta E(\theta)$

where $E(\theta) \stackrel{\text{def.}}{=} \bar{W}_{c,\varepsilon}(\mu_\theta, \nu)$

$\Rightarrow$ Issue : untractable gradient

Approximating Sinkhorn loss

- Rather than approximating the gradient approximate the loss itself
- Minibatches : $\hat{E}(\theta)$
  - sample $x_1, \ldots, x_m$ from μ_θ
  - use empirical Sinkhorn loss $\bar{W}_{c,\varepsilon}(\hat{\mu}_\theta, \hat{\nu})$ where $\hat{\mu}_\theta = \frac{1}{m}\sum_{i=1}^{m} \delta_{x_i}$
- Use L iterations of Sinkhorn's algorithm : $\hat{E}^{(L)}(\theta)$
  - compute L steps of the algorithm
  - use this as a proxy for $\bar{W}_{c,\varepsilon}(\mu_\theta, \nu)$

Sinkhorn's Algorithm

- State of the art solver for discrete regularized OT
- Two equivalent views
  - Alternate projections on the constraints of the primal
  - Alternate minimizations on the dual
- Iterates (a, b) : $\begin{cases} a \leftarrow \frac{1}{K(b \odot \nu)} \\ b \leftarrow \frac{1}{K^T(a \odot \mu)} \end{cases}$

  where $K \stackrel{\text{def.}}{=} \exp \frac{-\mathbf{c}}{\varepsilon}$ and $\odot$ is coordinatewise vector multiplication.
- Primal solution $\pi_\varepsilon = diag(a) K diag(b)$
- Linear convergence of the iterates to the optimizers
- Number of iterations needed for convergence increases when ε decreases

Computing the Gradient in Practice

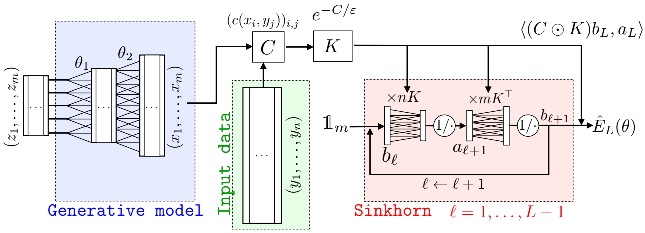


Figure 4: Scheme of the loss approximation

- Compute *exact* gradient of $\hat{E}^{(L)}(\theta)$ with autodiff
- Backpropagation through above graph
- Same computational cost as evaluation of $\hat{E}^{(L)}(\theta)$

Numerical Results on MNIST (L2 cost)

Figure 5: Samples from MNIST dataset

Numerical Results on MNIST (L2 cost)

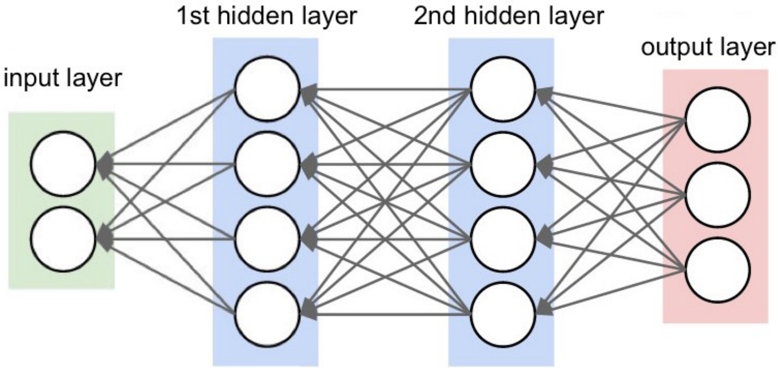
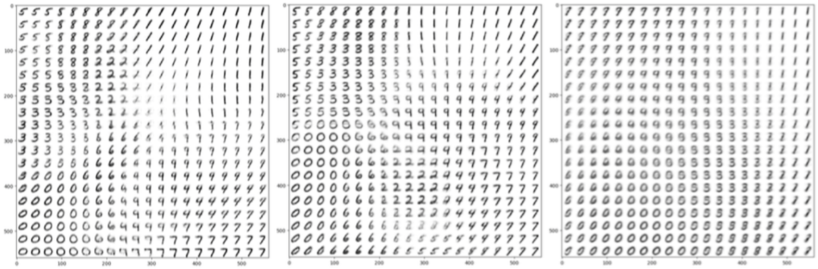


Figure 6: Fully connected NN with 2 hidden layers

Numerical Results on MNIST (L2 cost)



(a) $\varepsilon = 1, m = 200, L = 10$ (b) $\varepsilon = 10^{-1}, m = 200, L = 100$ (c) $\varepsilon = 10^{-1}, m = 10, L = 300$

Figure 7: Manifolds in the latent space for various parameters

Learning the cost
Li et al. '17, Bellemare et al. '17

- On complex data sets, choice of a good ground metric c is not trivial
- Use parametric cost function $c_\phi(x, y) = \|f_\phi(x) - f_\phi(y)\|_2^2$ (where $f_\phi : \mathcal{X} \to \mathbb{R}^d$)
- Optimization problem becomes minmax (like GANs)

$$min_\theta max_\phi \bar{W}_{c_\phi,\varepsilon}(\mu_\theta, \nu)$$

- Same approximations but alternate between updating the cost parameters ϕ and the measure parameters θ

Numerical Results on CIFAR (learning the cost)

Figure 8: Samples from CIFAR dataset

Numerical Results on CIFAR (learning the cost)

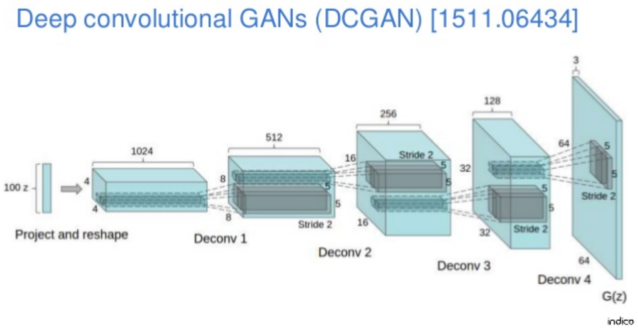
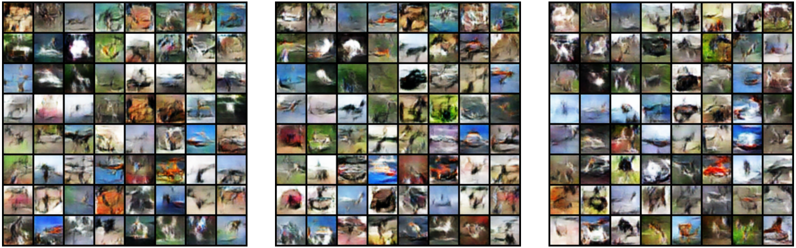


Figure 9: Fully connected NN with 2 hidden layers

Numerical Results on CIFAR (learning the cost)



(a) MMD (b) $\varepsilon = 1000$ (c) $\varepsilon = 10$

Figure 10: Samples from the generator trained on CIFAR 10 for MMD and Sinkhorn loss (coming from the same samples in the latent space)

Numerical Results on CIFAR (learning the cost)

Which image set is better? Not just about generating nice images, but more about capturing a high dimensional distribution...

$\rightarrow$ Hard to evaluate.

MMD	$\varepsilon = 100$	$\varepsilon = 10$	$\varepsilon = 1$
4.56 ± 0.07	4.81 ± 0.05	4.79 ± 0.13	4.43 ± 0.07

Table 1: Inception Scores

Conclusion

- Take Home message : Sinkhorn Divergences allow to interpolate between OT and MMD
- Future Work : Theory of Sinkhorn Divergences (positivity / sample complexity)