# Stochastic Methods for Optimal Transport and Applications in Machine Learning

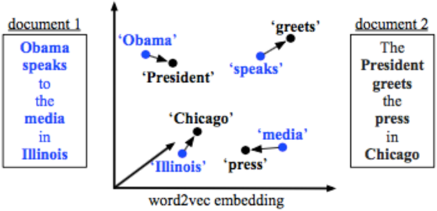
Aude Genevay

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

ISMP - July 2018

*Joint work with F. Bach, M.Cuturi, G. Peyré*

# Comparing High Dimensional Cloud Points




- Document $d_i$ = histogram of words
- Word $w_k$ = point in $\mathbb{R}^d$ for a certain embedding (usually learnt with neural networks, e.g. Word2Vec)
- Document $\sim$ weighted cloud of points in $\mathbb{R}^d \Rightarrow$ $d_i \sim \mu_i = \sum \alpha_{k,i} \delta_{w_k}$
- Distance between 2 documents $d_1$, $d_2$ is the optimal transport distance between the associated point clouds $\mu_1$, $\mu_2$.

# Fitting data to a probabilistic model

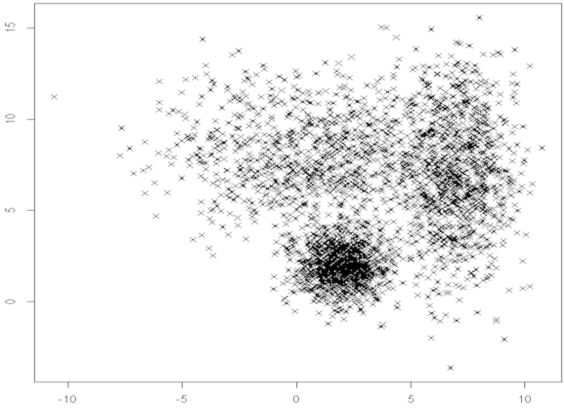



Figure 1: Data points in 2D

# Recurrent issue in ML : Fitting data to a probabilistic model

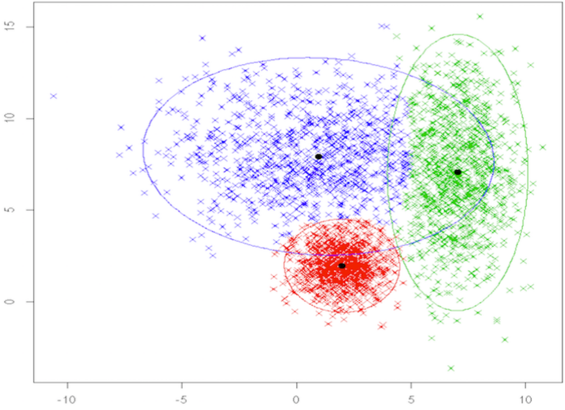



Figure 2: Gaussian Mixture Model

# Density Fitting with MLE

- Observed dataset $(y_1, \dots, y_n) \in \mathcal{X}$ (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**

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

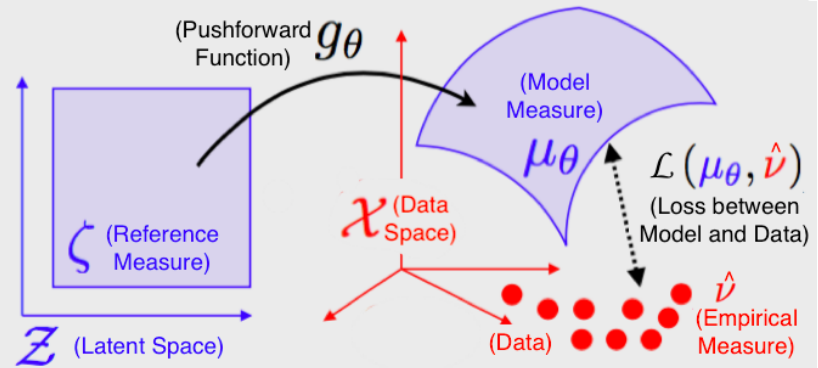



Figure 3: Illustration of Density Fitting on a Generative Model

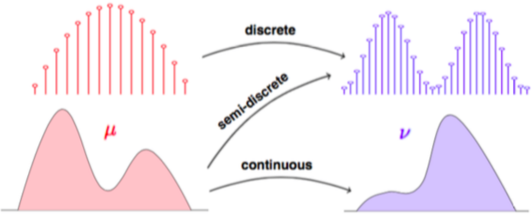
# Density Fitting for Generative Models I

Very popular topic in ML : image generation

- Parametric model : $\mu_\theta = g_{\theta\sharp}\zeta$
- $\zeta$ 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 $z \sim \zeta$
- dim $\mathcal{Z} < <$dim $\mathcal{X} \Rightarrow \mu_\theta$ doesn't have density wrt Lebesgue measure

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

# Optimal Transport I



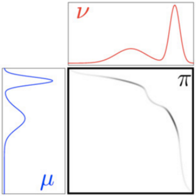

- Optimal Transport : find coupling that minimizes total cost of moving $\mu$ to $\nu$ whith unit cost function **c**
- Constrained problem : coupling has fixed marginals
- Minimal cost of moving $\mu$ to $\nu$(e.g. solution of the OT problem) is called the **Wasserstein distance** (it's an actual distance!)

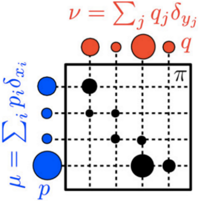
# Optimal Transport II

Cost $c(x, y)$ to move a unit of mass from $x$ to $y$
Constrained set of couplings $\Pi(\mu, \nu)$ with marginals $\mu$ and $\nu$

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

*What's the coupling that minimizes the total cost?*



# Kantorovitch Formulation of OT

The optimal overall cost for transporting $\mu$ to $\nu$ is given by

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

# Kantorovitch Formulation of OT

The optimal overall cost for transporting $\mu$ to $\nu$ is given by

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

where

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

# Kantorovitch Formulation of OT

The optimal overall cost for transporting $\mu$ to $\nu$ is given by

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

where

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

Adding an entropic regularization smoothes the constraint. In particular it yields an unconstrained dual problem.

# Dual formulation of OT

$$W(\mu, \nu) = \max_{u \in \mathcal{C}(\mathcal{X}), v \in \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y) \mathrm{d}\nu(y)(u, v) \tag{$\mathcal{D}_\varepsilon$}$$

under the constraint that

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, u(x) + v(y) \leq c(x, y)$$

# Dual formulation of OT (with entropy)

$$W_\varepsilon(\mu,\nu) = \max_{u\in\mathcal{C}(\mathcal{X}), v\in\mathcal{C}(\mathcal{Y})} \int_\mathcal{X} u(x)\mathrm{d}\mu(x) + \int_\mathcal{Y} v(y)\mathrm{d}\nu(y) - \iota^\varepsilon_{U_c}(u,v)$$

and the smoothed indicator is

$$\iota^\varepsilon_{U_c}(u,v) \stackrel{\text{def.}}{=} \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp\Big(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\Big)\mathrm{d}\mu(x)\mathrm{d}\nu(y)$$

# Semi-Dual formulation of OT (with entropy)

The dual problem is convex in $u$ and $v$. We fix $v$ and minimize over $u$.

## Semi-Dual formulation of OT (with entropy)

The dual problem is convex in $u$ and $v$. We fix $v$ and minimize over $u$. This yields :

$$u(x) \stackrel{\text{def.}}{=} -\varepsilon \log \left( \int_{\mathcal{Y}} \exp(\frac{v(y) - c(x,y)}{\varepsilon}) \mathrm{d}\nu(y) \right)$$

## Semi-Dual formulation of OT (with entropy)

The dual problem is convex in $u$ and $v$. We fix $v$ and minimize over $u$. This yields :

$$u(x) \stackrel{\text{def.}}{=} -\varepsilon \log \left( \int_{\mathcal{Y}} \exp\left(\frac{v(y) - c(x,y)}{\varepsilon}\right) \mathrm{d}\nu(y) \right)$$

Plugging back in the dual :

$$\begin{aligned} W_\varepsilon(\mu,\nu) &= \max_{v \in \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} -\varepsilon \log \left( \int_{\mathcal{Y}} \exp\left(\frac{v(y) - c(x,y)}{\varepsilon}\right) \mathrm{d}\nu(y) \right) \mathrm{d}\mu(y) \\ &\quad + \int_{\mathcal{Y}} v(y) \mathrm{d}\nu(y) - \varepsilon \\ &= \max_{v \in \mathcal{C}(\mathcal{Y})} \mathbb{E}_\mu \Big[ -\varepsilon \log \left( \int_{\mathcal{Y}} \exp\left(\frac{v(y) - c(x,y)}{\varepsilon}\right) \right) \\ &\quad + \int_{\mathcal{Y}} v(y) \mathrm{d}\nu(y) - \varepsilon \Big] \end{aligned}$$

We consider 2 frameworks :

- Semi-Discrete : $\mu$ is continuous and $\nu = \sum_{j=1}^{M} \nu_i \delta y_j$ The optimization problem is

$$\max_{v \in \mathbb{R}^M} \mathbb{E}_{\mu} \left[ -\varepsilon \log \left( \sum_{j=1}^{M} \exp\left( \frac{v(y_j) - c(x, y_j)}{\varepsilon} \right) \right) + \sum_{j=1}^{M} v(y_j) \nu_j - \varepsilon \right]$$

We consider 2 frameworks :

- Semi-Discrete : $\mu$ is continuous and $\nu = \sum_{j=1}^{M} \nu_i \delta y_j$ The optimization problem is

$$\max_{v \in \mathbb{R}^M} \mathbb{E}_{\mu} \left[ -\varepsilon \log \left( \sum_{j=1}^{M} \exp\left( \frac{v(y_j) - c(x, y_j)}{\varepsilon} \right) \right) + \sum_{j=1}^{M} v(y_j) \nu_j - \varepsilon \right]$$

- Discrete : $\mu = \sum_{i=1}^{N} \mu_i \delta x_i$ and $\nu = \sum_{j=1}^{M} \nu_i \delta y_j$ The optimization problem is

$$\max_{v \in \mathbb{R}^M} \sum_{i=1}^{N} \left[ -\varepsilon \log \left( \sum_{j=1}^{M} \exp\left( \frac{v(y_j) - c(x_i, y_j)}{\varepsilon} \right) \right) + \sum_{j=1}^{M} v(y_j) \nu_j - \varepsilon \right] \mu_i$$

## Stochastic Optimization

Computing the full gradient is

- Hard in the semi-discrete setting (even impossible if we don't know $\mu$ explicitly)

# Stochastic Optimization

Computing the full gradient is

- Hard in the semi-discrete setting (even impossible if we don't know $\mu$ explicitly)
- Very costly in the discrete case since we need to compute $N$ gradients and sum them.

The idea of stochastic optimization is to use approximate gradients so that each iteration is inexpensive.

# Stochastic Optimization I

- Goal : maximize $H_\varepsilon(v) = \mathbb{E}_\mu [h_\varepsilon(X, v)]$ over $v$ in $\mathbb{R}^M$.
- Standard gradient ascent :

$$v^{(k)} = v^{(k-1)} + \nabla_v H_\varepsilon(v^{(k-1)})$$

- The whole gradient $\nabla_v H_\varepsilon(v)$ is too costly/complicated to compute
- Idea : Sample $x$ from $\mu$ and use $\nabla_v h_\varepsilon(x, v)$ as a proxy for the full gradient in the gradient ascent.

# Stochastic Optimization II

**Algorithm 1** Averaged SGD

**Input:** $C$

**Output:** $v$

$v \leftarrow \mathbb{0}_M$, $\bar{v} \leftarrow v$

**for** $k = 1, 2, \ldots$ **do**

Sample $x_k$ from $\mu$

$v \leftarrow v + \frac{C}{\sqrt{k}} \nabla_v h_\varepsilon(x_k, v)$ (gradient ascent step)

$\bar{v} \leftarrow \frac{1}{k} v + \frac{k-1}{k} \bar{v}$ (averaging)

**end for**

- cost of each iteration $M$
- convergence rate $O(1/\sqrt(k))$

# Stochastic Optimization : Case of a Finite Sum I

In the specific case where $\mu$ is also a discrete measure, we are minimizing a finite sum of $N$ functionals :

$$\max_{v\in\mathbb{R}^M} \sum_{i=1}^{N} \left[ -\varepsilon \log \left( \sum_{j=1}^{M} \exp\left( \frac{v(y_j) - c(x_i, y_j)}{\varepsilon} \right) \right) + \sum_{j=1}^{M} v(y_j) \nu_j - \varepsilon \right] \mu_i$$

Variance reduction algorithms (e.g. SAGA) can be used to improve speed of convergence:

- cost of each iteration $M$
- convergence rate $O(1/k)$

# Numerical Results for Word Mover's Distance (Discrete OT)

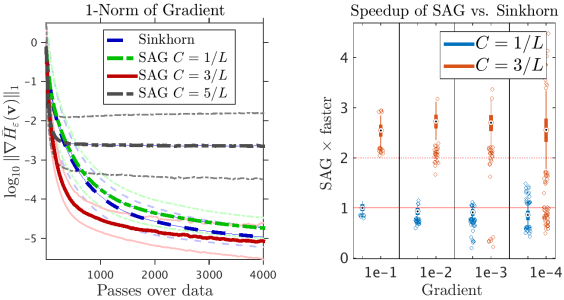
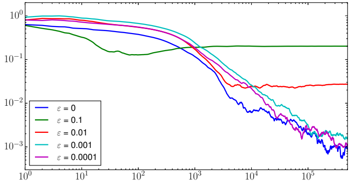
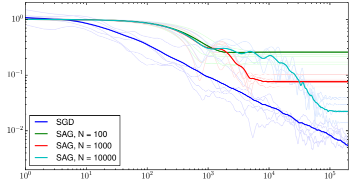



Figure 4: Results for the computation of 595 pairwise word mover's distances between 35 very large corpora of text, each represented as a cloud of $I = 20,000$ word embeddings.

# Numerical Results for Density Fitting (Semi-discrete OT)




(a) SGD




(b) SGD vs. SAG

Figure 5: (a) Effect of regularization parameter $\varepsilon$ (b) Effect of sampling (discrete algo) vs. using semi-discrete algo (blue)

# Dual Formulation as an Expectation

Recall the dual objective function to be maximized, for $\varepsilon > 0$

$$\begin{aligned} F_\varepsilon(u, v) &= \int_{\mathcal{X}} u(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\nu(y) \\ &\quad -\varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right)\mathrm{d}\mu(x)\mathrm{d}\nu(y) \end{aligned}$$

## Dual Formulation as an Expectation

Recall the dual objective function to be maximized, for $\varepsilon > 0$

$$\begin{aligned} F_\varepsilon(u, v) &= \int_{\mathcal{X}} u(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\nu(y) \\ &\quad -\varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right)\mathrm{d}\mu(x)\mathrm{d}\nu(y) \end{aligned}$$

Let $X \sim \mu$ and $Y \sim \nu$ be two independent random variables, we get

$$F_\varepsilon(u, v) = \mathbb{E}_{\mu\otimes\nu}\left[f_\varepsilon(X, Y, u, v)\right]$$

where $\forall \varepsilon > 0$,

$$f_\varepsilon(x, y, u, v) \stackrel{\text{def.}}{=} u(x) + v(y) - \varepsilon \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right).$$

# Reminder on RKHS I

We consider two reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}$ and $\mathcal{G}$ on $\mathcal{X}$ and on $\mathcal{Y}$, with kernels $\kappa$ and $\ell$.

## Properties of RKHS

(a) if $u \in \mathcal{H}$, then $u(x) = \langle u, \kappa(\cdot, x)\rangle_{\mathcal{H}}$
(b) $\kappa(x, x') = \langle \kappa(\cdot, x), \kappa(\cdot, x')\rangle_{\mathcal{H}}$.

## The Gaussian Kernel

For the Gaussian Kernel i.e. $\kappa(x, x') = \exp(\frac{\|x-x'\|^2}{2\sigma^2})$ the associated RKHS is dense in the space of continuous functions. This means that any continuous function can be approximated by a linear combination of Gaussian Kernels.

# Reminder on RKHS II

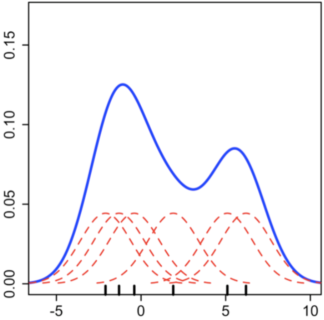



Figure 6: Approximation of a function by a sum of gaussian kernels. The choice of the bandwidth is crucial.

# Continuous OT I

$$f_\varepsilon(x, y, u, v) \overset{\text{def.}}{=} u(x) + v(y) - \varepsilon \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right).$$

Rewriting $u(x)$ and $v(y)$ as scalar products in $\mathcal{H}$ and $\mathcal{G}$ we get

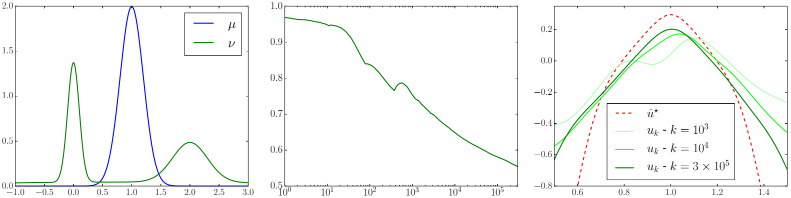
$$\begin{aligned} f_\varepsilon(x, y, u, v) &\overset{\text{def.}}{=} \langle u, \kappa(\cdot, x)\rangle_{\mathcal{H}} + \langle v, \ell(\cdot, y)\rangle_{\mathcal{G}} \\ &\quad - \varepsilon \exp\left(\frac{\langle u, \kappa(\cdot, x)\rangle_{\mathcal{H}} + \langle v, \ell(\cdot, y)\rangle_{\mathcal{G}} - c(x, y)}{\varepsilon}\right). \end{aligned}$$

we can apply the SGD algorithm in the RKHS :

$$(u_k, v_k) \overset{\text{def.}}{=} (u_{k-1}, v_{k-1}) + \frac{C}{\sqrt{k}} \nabla f_\varepsilon(x_k, y_k, u_{k-1}, v_{k-1}) \in \mathcal{H} \times \mathcal{G}, \tag{1}$$

where $(x_k, y_k)$ are i.i.d. samples from $\mu \otimes \nu$.

# Continuous OT : Numerical Results




(a) setting (b) convergence of $u_k$ (c) plots of $u_k$

Figure 7: (a) Plot of $\frac{\mathrm{d}\mu}{\mathrm{d}x}$ (blue) and $\frac{\mathrm{d}\nu}{\mathrm{d}x}$ (green). (b) Plot of $\|\mathbf{u}_k - \hat{\mathbf{u}}^\star\|_2 \,/\, \|\hat{\mathbf{u}}^\star\|_2$ as a function of $k$ with SGD in the RKHS, for regularized OT using $\varepsilon = 10^{-1}$. (c) Plot of the iterates $u_k$ for $k = 10^3, 10^4, 10^5$ and the proxy for the true potential $\hat{\mathbf{u}}^\star$, evaluated on a grid where $\mu$ has non negligible mass.

# Continuous OT : Theory in progress

We recently proved that the dual potentials are in a Sobolev ball (and thus bounded in a certain RKHS)

- We get convergence of kernel SGD in the continuous setting
- We can use standard results on RKHS to prove regularized OT has sample complexity in $O(n^{-1/2})$, similar to MMD / much better than standard OT

# Conclusion

- Dual formulations of OT can be rewritten as expectation maximization problems
- This allows the use of stochastic optimization methods
- Surpass Sinkhorn in the discrete setting (online method more efficient than batch)
- Tackle semi-discrete and continuous problems without requiring discretization