Optimal Transport
&
Applications to Data Science

Gonzalo Mena

Columbia University

Statistics Department

Jan 17th, 2018

This talk

- Increased interest in ML community to use optimal transport (OT) methods.

- NIPS Tutorial on OT.

- New Book: Computational Optimal Transport (Peyre & Cuturi), FREE online.

- Classic Book: Cédric Villani.
- Hundreds of research papers in the past 5 years.

- Good opportunity to create interactions between mathematicians/theorists and data scientists/ML researchers.

Current (past) statistical paradigm

Maximum Likelihood estimator (popularized by Fisher)

$$\{x_i\}_{i=1_N} \sim p_{\bar{\theta}}(x), i.i.d., \;\; \theta_{MLE} = \arg\max_\theta \prod_{i=1}^N p_\theta(x_i)$$

- Nice interpretation.
- Nice properties: sufficiency, efficiency, etc
- Can be thought as minimization of a discrepancy. If p_θ is density of α_θ and $p_{\bar{\theta}}$ of $\hat{\beta}$ with empirical $\beta_n = \frac{1}{N}\sum_{i=1}^N \delta(x_i)$

$$\min_\theta \mathcal{L}_{MLE}(\alpha_\theta, \beta_n) \to KL(\alpha_\theta, \beta)$$

$$\mathcal{L}_{MLE}(\alpha_\theta, \beta_n) \equiv -\sum_{i=1}^N \log p_\theta(x_i) \to KL(\alpha_\theta, \beta)$$

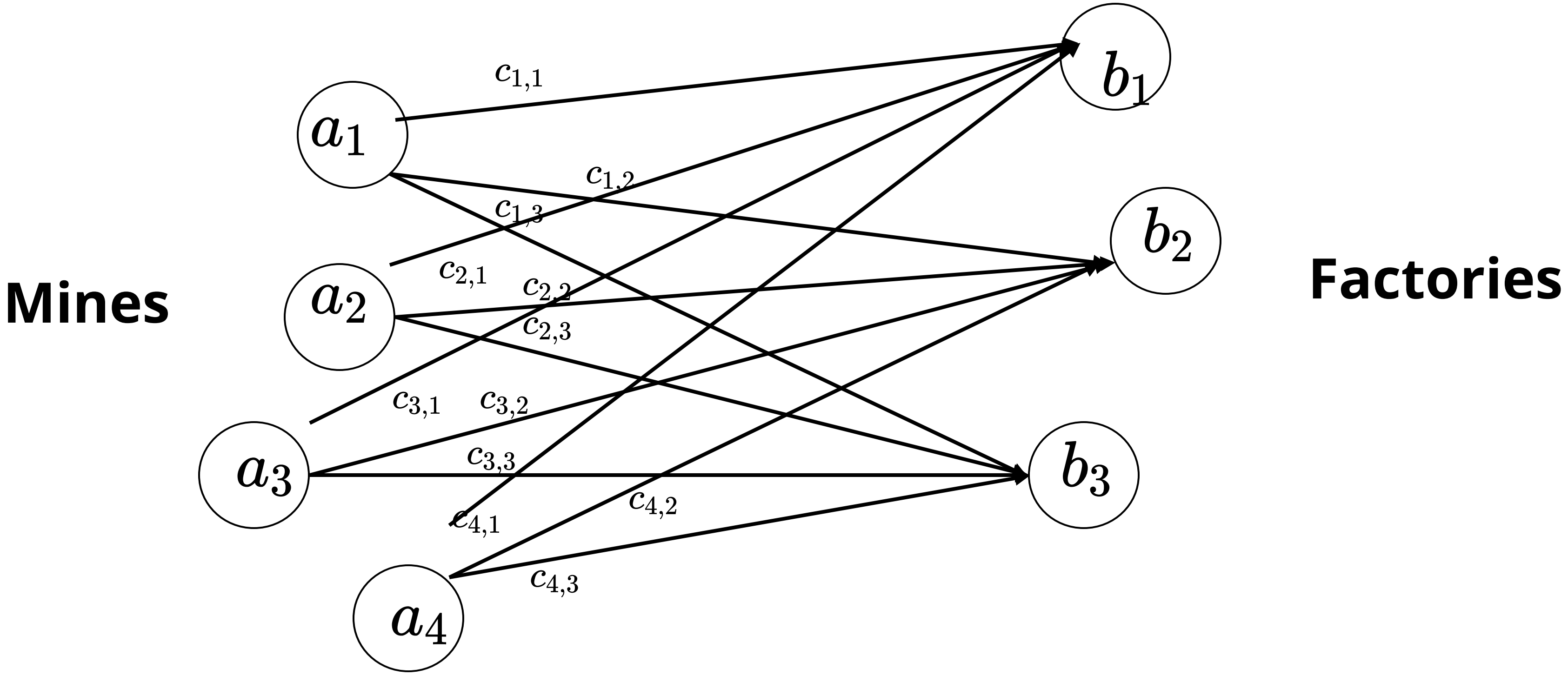
- It induces geometry, e.g. Fisher information metric, Amari 1985.
- However, it is not 'perfect'.

Optimal Transport

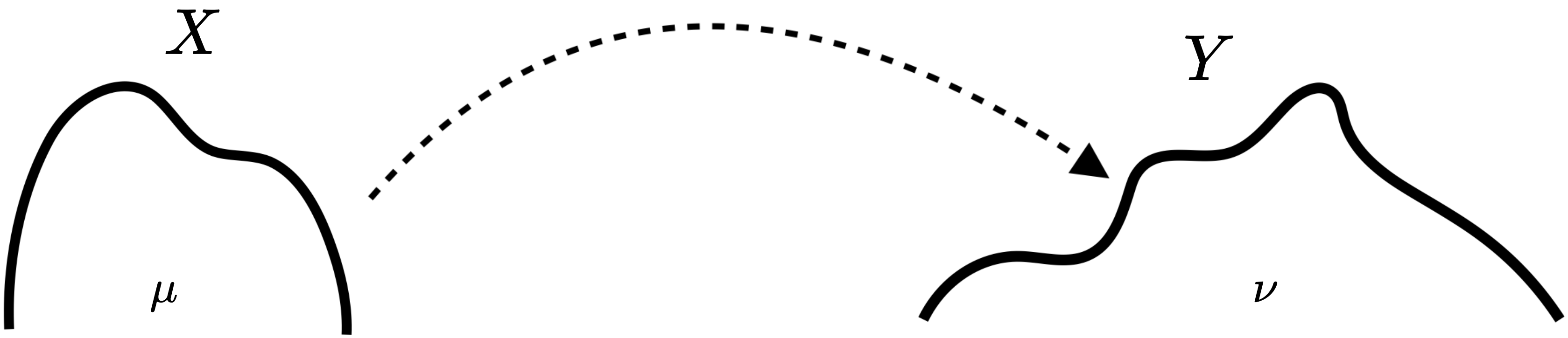
How to minimize the cost of transporting
all units in mines to factories?

$$\min p_{i,j}c_{i,j}$$

$$p_{i,j} \geq 0 \sum_i p_{i,j} = a_i, \sum_j p_{i,j} = b_j, \sum_i a_i = \sum_j b_j$$



Optimal Transport



Monge Formulation

$$\inf_T \int_X c(x, T(x)) d\mu(x)$$

$$\text{s.t. } \nu = T\#\mu \text{ i.e. } \nu[B] = \mu(T^{-1}(B))]$$

Kantorovich Formulation

$$\inf_\pi \int_{X\times Y} c(x, y) d\pi(x, y)$$

$$\text{s.t. } \pi \in \Pi[\mu, \nu], \text{ i.e. } \pi[X \times B] = \nu(B), \pi[A \times Y] = \mu(A)$$

Kantorovich Duality

Let X, Y be Polish spaces and $C : X \times Y \to \mathbb{R}_+ \cup \{+\infty\}$ be a lower-semi continuous cost function.

Define

$$I[\pi] = \int_{X \times Y} c(x,y)d\pi(x,y), \quad J(\phi,\psi) = \int_X \phi d_\mu + \int_Y \psi d\nu$$

And

$$\Phi_c = \{(\phi,\psi) \in L^1(d\mu) \times L^1(d\nu) : \phi(x) + \psi(y) \leq c(x,y)\}$$

Then,

$$\inf_{\Pi(\mu,\nu)} I[\pi] = \sup_{\Phi_c} J(\phi,\psi)$$

Legendre-Fenchel transform

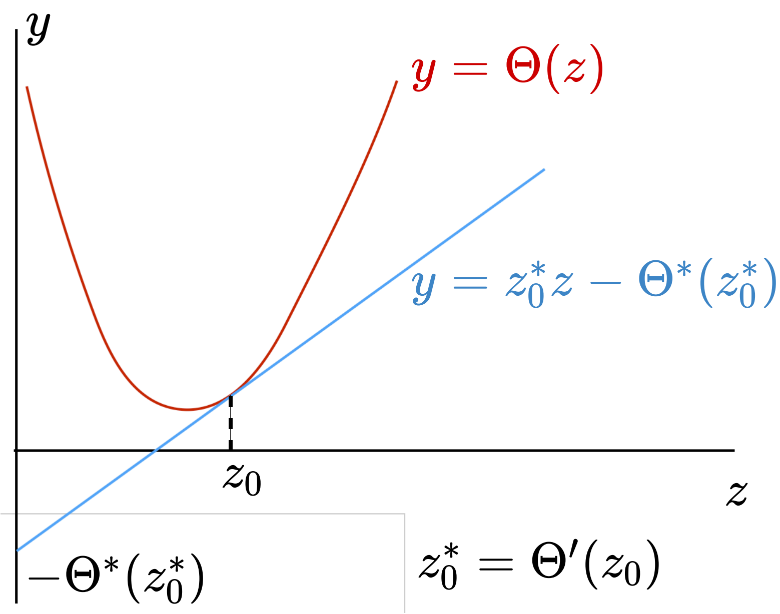
Let $\Theta(\cdot)$ be convex, on a n.v.s E

with values in $\mathbb{R} \cup \{+\infty\}$

For values in the topological dual E^*

define the L-F conjugate as:

$$\Theta^*(z^*) = \sup_{z \in E}[\langle z^*, z\rangle - \Theta(z)]$$



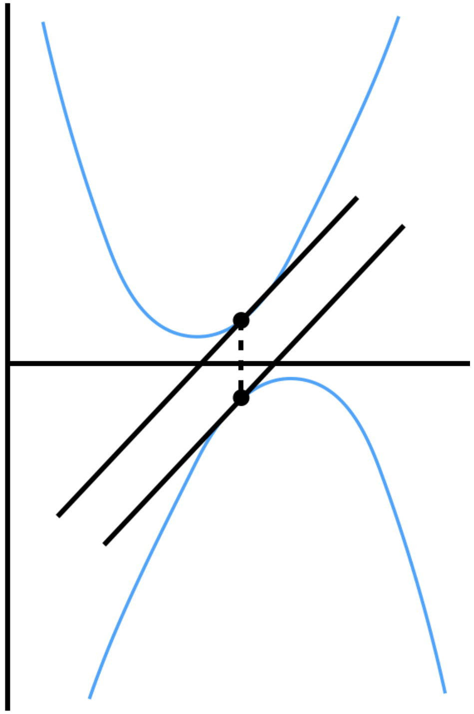
Fenchel-Rockafellar theorem

Let $\Theta(\cdot), \Xi(\cdot)$ be convex. If for some $z_0 \in E$

$\Theta(\cdot)$ is continuous at z_0 and $\Theta(z_0) < +\infty, \Xi(z_0) < +\infty$

Then,

$$\inf_{z \in E}[\Theta + \Xi] = \max_{z^* \in E^*}[-\Theta^*(-z^*) - \Xi^*(z^*)]$$



Proof sketch

1. Start with X, Y compact, c continuous and use $E = C_b(X \times Y)$ therefore $E^* = M(X \times Y)$ (regular radon measures). Show it holds there by Fenchel-Rockafellar theorem.

2. $\Pi(\mu, \nu)$ is compact; infimum is attained.

3. Relax compactness using tightness of the minimizer π^* and a truncation argument. Notice one can use $E = L^1(d\mu) \times L^1(d\nu)$.

Kantorovich-Rubinstein Theorem

Let $X = Y$ be a Polish space and $c(x, y) = d(x, y)$ a lower semi continuous metric.

Define

$$\mathcal{T}_d = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d(x, y) d\pi(x, y)$$

And

$$\|\phi\|_{Lip} \equiv \sup_{x \neq y} \frac{|\phi(x) - \phi(y)|}{d(x,y)}$$

Then,

$$\mathcal{T}_d = \sup\{\int_X \phi d(\mu - \nu);\ \phi \in \cap L^1(d|\mu - \nu|);\ ||\phi||_{Lip} \leq 1\}$$

Relevant case: $d(x, y) = \|x - y\|$.

Proof sketch

1. For ϕ bounded define the pair of c-concave conjugate pair (ϕ^c, ϕ^{cc}) as:

$$\phi^c(y) = \inf_{x\in X}[c(x,y) - \phi(x)], \phi^{cc}(y) = \inf_{y\in Y}[c(x,y) - \phi^c(y)]$$

2. Notice one can replace (ϕ, ψ) by (ϕ^{cc}, ϕ^c) and $\phi^c = -\phi^{cc}$

3. $$\sup_{\phi\in L^1(d\nu)} J(-\phi^d, \psi^d) = \sup_{\|\phi\|_{Lip}\leq 1} J(\phi, -\phi)$$

The case $c(x,y) = \|x-y\|^2$

Knott-Smith optimality criterion

$\pi \in \Pi(\mu,\nu)$ is optimal if there exists a convex lower semi-continuous $\phi(x)$ satisfying,

For $d\pi$ almost all (x,y), $\quad y \in \partial\phi(x)$

Moreover, (ϕ, ϕ^*) minimizes

$$\inf\{\textstyle\int_{\mathbb{R}^n} \phi d\mu + \int_{\mathbb{R}^n} \psi d\nu;\ \forall(x,y),\ x\cdot y \le \phi(x) + \psi(y)\}$$

Brenier's theorem

If μ does not give mass to Lebesgue-negligible sets then the unique optimal π satisfies.

$$d\pi(x,y) = d\mu(x)\delta[y = \nabla\phi(x)]$$

$\nabla\phi(x)$ is the unique gradient of a convex function s.t $\nu = \nabla\phi\#\mu$

The metric side of OT

Let X be a Polish space endowed with a distance d and $p \geq 0$

Define

$$\mathcal{T}_p = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d^p(x,y) d\pi(x,y)$$

Then, if $p \in [1, \infty), W_p = \mathcal{T}_p^{1/p}$ defines a metric on $\mathcal{P}_p(X)$ (p-moment bounded)

If $p \in [0, 1), W_p = \mathcal{T}_p$ defines a metric on $\mathcal{P}_p(X)$

Wasserstein distances are distances between distributions and metrize weak convergence.

Displacement interpolation

Time dependent (Monge) optimal transport

$$\inf\{\int_X C[(T_t x)]_{0\leq t\leq 1}\, d\mu(x); T_0 = Id, T_1\#\mu = \nu\}$$

C is consistent with original Monge if optimal T_t gives rise to optimal $T = T_1$. e.g., $c(x,y) = \inf\{C[(z_t)_{0\leq t\leq 1}], z_0 = x, z_1 = y\}$

$$C[(z_t)_{0\leq t\leq 1}] = \int_0^1 |\dot{z}_t|^2 dt \to c(x,y) = |x-y|^2$$

In this case **(McCann displacement interpolation)**

$$T_t(x) = (1-t)x + t\nabla\phi(x) \leftrightarrow \rho_t = [(1-t)Id + t\nabla\phi]\#\mu$$

$$W_2(\mu, \rho_t) = tW_2(\mu,\nu)$$

Constant velocity geodesic.

Computational Optimal transport

OT problem states: $d_c(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} E_\pi(c(x,y))$

In practice, transportation between discrete measures

$$\mu = \sum_{i=1}^n u_i \delta_{x_i}, \nu = \sum_{j=1}^m v_j \delta_{y_j}$$

The problem becomes

$$d_c(\mu,\nu) = \inf_{P \in U(u,v)} \langle P, C \rangle$$

$$\langle P, C \rangle = \sum_{i,j} P_{i,j} C_{i,j}, \quad C_{i,j} = c(x_i, y_j)$$

In the transportation polytope

$$U(u,v) = \{P \in \mathcal{R}_+^{n\times m};\ P1_m = u, P^T 1_n = v\}$$

Entropy regularized Optimal Transport. Why?

The discrete OT is solved by combinatorial optimization algorithm with $O(n^3 \log(n))$ Perhaps better approximate solutions?

The operator $C \to \arg\min_{P \in U(a,b)} \langle C, P \rangle$ is not differentiable. Too bad for automatic differentiation libraries.

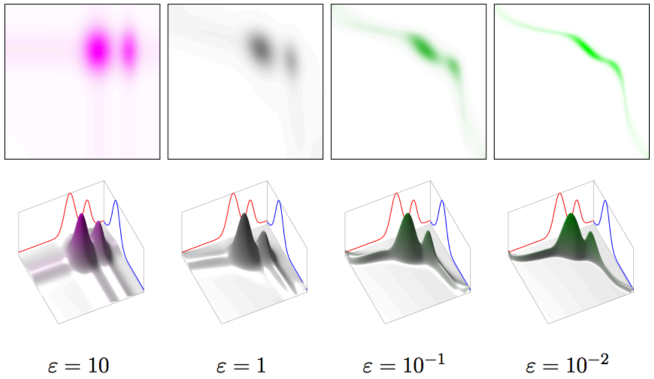
Entropy regularized Optimal Transport

Add entropy term $h(P) = -\sum_{i,j} P_{i,j} \log P_{i,j}$ Cuturi, 2012

$$d_c^\lambda(\mu,\nu) = \inf_{P\in U(u,v)} \langle P, C\rangle - \lambda h(P) \quad \textit{Cuturi function}$$
$$= \inf_{P\in U^\alpha(u,v)} \langle P, C\rangle, \quad U^\alpha(u,v) = \{P \in U(u,v) : KL(P\|uv^\top) \le \alpha\}$$

i.e., enforce contigency tables to be close to the independent

From Computational OT book



Sinkhorn algorithm

First order optimality condition

$$P_{i,j}^{\lambda} = \exp(-1/2 - \alpha_i/\lambda)\exp(-c_{i,j}/\lambda)\exp(-1/2 - \beta_j/\lambda)$$

with α, β Lagrange multipliers i.e.,

$$P^{\lambda} = diag(a)K^{\lambda}diag(b), \quad a, b > 0, K^{\lambda} = \exp(-C/\lambda)$$

By Sinkhorn's theorem, there is a unique $P^{\lambda} \in U(u, v)$ as above and can be obtained by Sinkhorn fixed point iterations (matrix scaling)

$$\boxed{a^{l+1} = \frac{u}{Kb^{l}}, b^{l+1} = \frac{v}{K^{\top}a^{l+1}}}$$

Other relations

Bregman projections

$$\mathcal{C}_u = \{P \in \mathcal{R}_+^{n\times m};\ P1_m = u\}, \mathcal{C}_v = \{P \in \mathcal{R}_+^{n\times m};\ P^\top 1_n = v\}$$

$$d_c^\lambda(\mu,\nu) = \inf_{P\in U(u,v)} KL(P\|K^\lambda)$$

$$P^{l+1} = Proj_{\mathcal{C}_u}^{KL}(P^l), P^{l+2} = Proj_{\mathcal{C}_v}^{KL}(P^{l+1})$$

Duality

$$d_c^\lambda(\mu,\nu) = \max_{f\in\mathbb{R}^n, g\in\mathbb{R}^m} \langle f,u\rangle + \langle g,v\rangle - \lambda\langle e^{f/\lambda}, Ke^{g/\lambda}\rangle$$

Convergence to actual solution. Cominetti & San Martin 1994

Reverse mode automatic differentiation

Solution can be written as $\langle diag(a^l)K^\lambda diag(b^l), C\rangle$

Differentiate this expression with respect to relevant parameters

The case of permutations

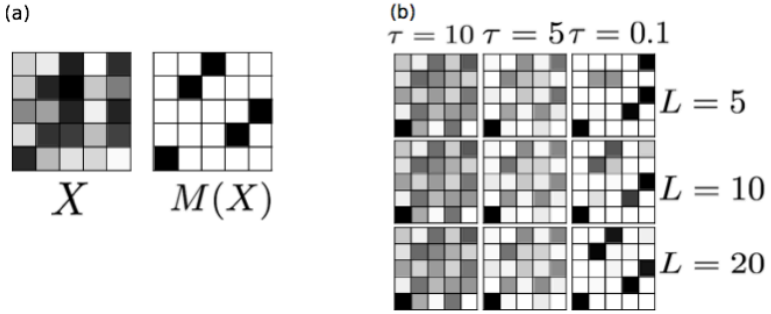
- If $m = n$ one solves a linear program on the Birkhoff polytope. $\mathcal{B}$
- Sinkhorn algorithm becomes iterative row and column scaling of a matrix, Sinkhorn operator $S(\cdot)$.
- Entropy regularization allow us to conceive soft matchings

$$S(C/\lambda) = \arg\max_{P \in \mathcal{B}} \langle P, C \rangle + \lambda h(P)$$

- In the limit become hard permutations

$$M(C) \equiv \arg\max_{P \in \mathcal{B}} \langle P, C \rangle = \lim_{\lambda \to 0^+} S(C/\lambda)$$

- Analogy with softmax



Mena et al, 2018

Complexity of Sinkhorn algorithm

Active area of research. Sinkhorn algorithm works in near-linear time

Theorem (Altschuler, Weed, Rigollet, 2017). Sinkhorn algorithm gives a solution $\hat{P}$ satisfying

$$\langle \hat{P}, C \rangle \leq \min_{P \in U(u,v)} \langle P, C \rangle + \epsilon$$

in $O(n^2 L^4 \log(n)^2 \epsilon^{-3})$ operations.

Here, $\epsilon = \lambda \log n, \epsilon' = \epsilon / (4 \|C\|_\infty) \quad \|C\|_\infty \leq L$

Sinkhorn iterations are defined with a tolerance ϵ'

Wasserstein Barycenters

$$\bar{\mu} = \arg\min_u \sum_{i=1}^N \lambda_i W_2^2(\mu, \nu_i)$$

Generalizes McCann displacement formula (Agueh and Carlier, 2010)

Relates to Multi-Marginal problem (Agueh and Carlier, 2010)

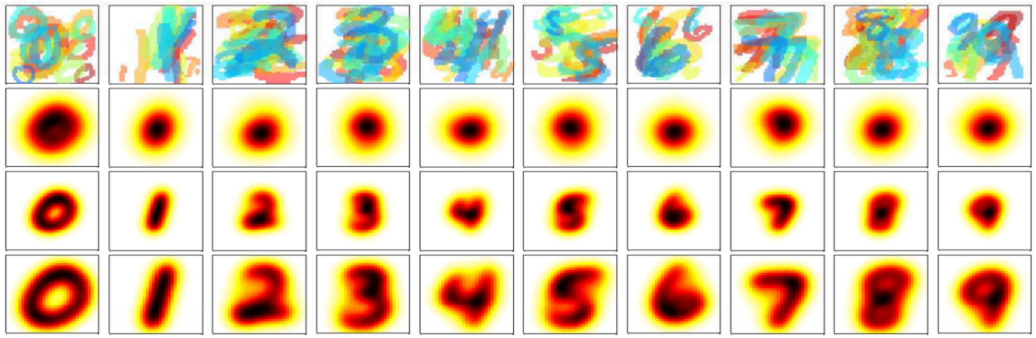
$$\inf_{\pi \in \Pi(\mu_1, \ldots, \mu_p)} \int_{\mathbb{R}^d} (\sum_{i=1}^p \frac{\lambda_i}{2} |x_i - T(x)|^2) d\pi(x_1, \ldots, x_p)$$

$$\bar{\mu} = (\sum_{i=1}^p T_i^j) \# \nu_j$$

Wasserstein barycenters

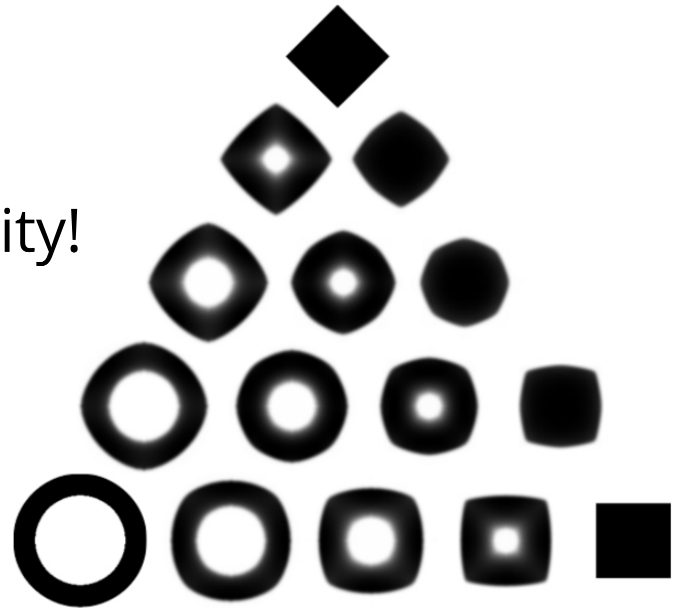
Cuturi and Doucet, 2014

Images as histograms



Benamou et al 2014

Use entropic regularization for scalability!



Wasserstein Principal Geodesic Analysis

Flectcher et al, 2004. Extension of PCA to the manifold setting

Algorithm 2: Principal Geodesic Analysis

Input: $x_1, \ldots, x_N \in M$

Output: Principal directions, $v_k \in T_\mu M$

Variances, $\lambda_k \in \mathbb{R}$

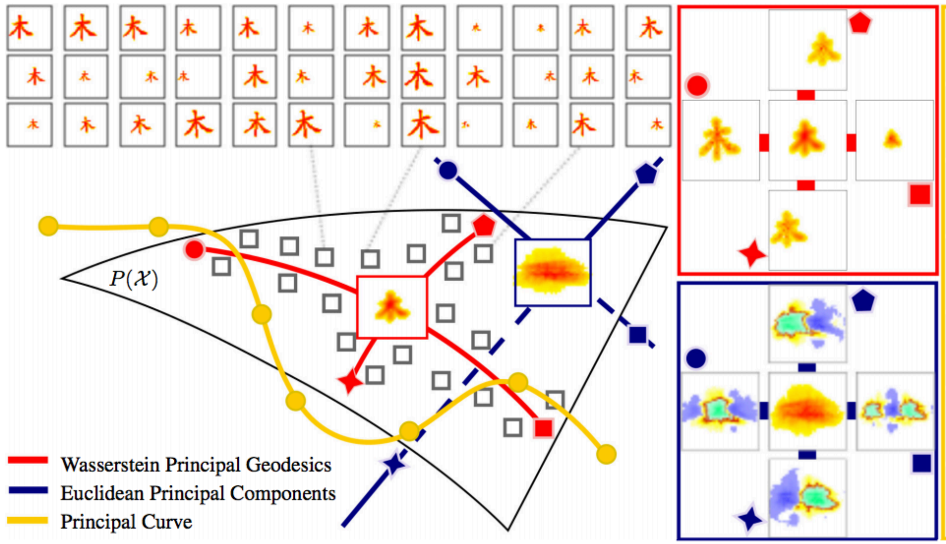
$\mu =$ intrinsic mean of $\{x_i\}$ (Algorithm 1)

$u_i = \mathrm{Log}_\mu(x_i)$

$\mathbf{S} = (1/N) \sum_{i=1}^{N} u_i u_i^T$

$\{v_k, \lambda_k\} =$ eigenvectors/eigenvalues of $\mathbf{S}$.

Scaled by Seguy and Cuturi, 2015



Wassersteing case

Bigot, Gouet, Klein, Lopez 2015.

Generative models

How to estimate θ in a model defined as follows: Sample from a reference (noise) distribution $z \sim \zeta \in \mathcal{P}(\mathcal{Z})$, then $x = h_\theta(z) \in X$

$h_\theta : \mathcal{Z} \to X$ is usually a complex nonlinear function between two high-dimensional spaces $\alpha_\theta = h_\theta \# \zeta$. Don't have a tractable density and cannot use MLE.

Instead, consider the problem(β is the true distribution of data.)

$$\min_\theta W_1(\alpha_\theta, \beta) = \min_\theta \inf_{P \in \Pi(\alpha_\theta, \beta)} E(\|x - y\|)$$

$$= \min_\theta \sup_{f \in Lip} E_{\alpha_\theta}(f(x)) - E_\beta(f(x))$$

Solve minimax problem using $\nabla_\theta W_1(\alpha_\theta, \beta) = E_{z \sim \zeta}(\nabla_\theta f(h_\theta(z))$

Wasserstein GAN, Arjovsky et al 2017

More generative models

Alternative formulation

Genevay et al 2017 uses primal formulation

$$\min_\theta W_1(\alpha_\theta, \beta) = \min_\theta \inf_{P \in \Pi(\alpha_\theta, \beta)} E(\|x - y\|)$$

Also, use entropy regularization to enable automatic differentiation.

Alternative distances, Integral probability metrics

$$\mathcal{E}(\alpha, \beta)^2 \equiv 2E_{x\sim\alpha, y\sim\beta}(\|x - y\|) - E_{x\sim\alpha, x'\sim\alpha}(\|x - x'\|) - E_{y\sim\beta, y'\sim\beta}(\|y - y'\|)$$

$$\mathcal{E}(\alpha, \beta)^2 = \sup_{f \in \mathcal{H}, ||f||_\mathcal{H} \leq 1} E_{\alpha_\theta}(f(x)) - E_\beta(f(x))$$

$$D_{TV}(\alpha, \beta) = \sup_{A \in \mathcal{F}} |\alpha(A) - \beta(A)| = \sup_{f \in C(X, [0,1])} E_{\alpha_\theta}(f(x)) - E_\beta(f(x))$$

More generative models

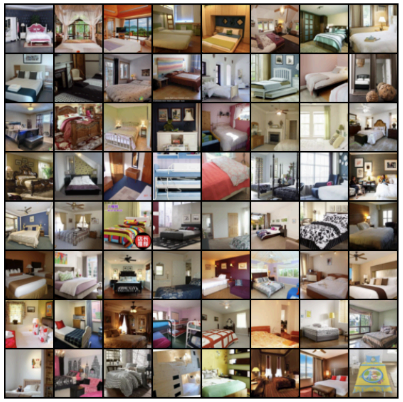
Sample complexity

Computations involve samples from the model and data.

$d =$ data dimension

$$E(|W_1(\alpha_n,\beta_n) - W_1(\alpha,\beta)|) = O(n^{-1/d})$$

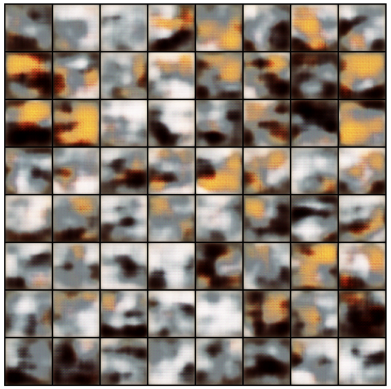
$$E(|\mathcal{E}_1^2(\alpha_n,\beta_n) - \mathcal{E}^2(\alpha,\beta)|) = O(n^{-1})$$



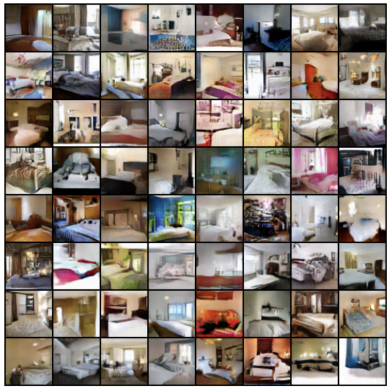
This would be catastrophic, but in practice it is not.

Why? study differences in induced geodesic structure

Bottou et al, 2017



Generated by the ED trained model



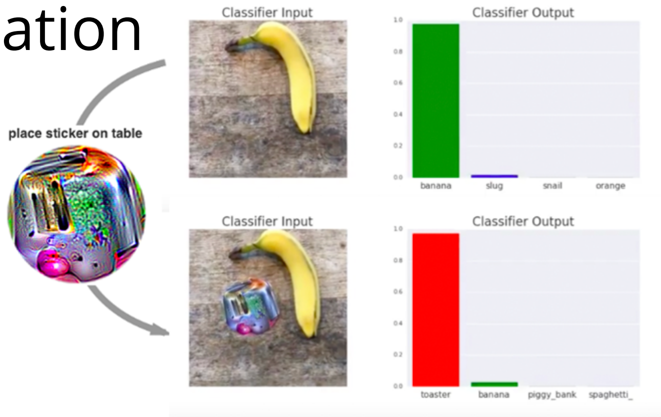
Generated by the WD trained model

Robustness to adversarial attacks

Goal: quantify robustness to adversarial examples and design robust algorithms

Classical approach: Empirical Risk Minimization

$$\min_\theta \frac{1}{N}\sum_{i=1}^N l(\theta, x_i) = \min_\theta E_{P_n}(l(\theta, x))$$



'Modern' approach: distributionally robust optimization

$$\min_\theta \sup_{P:W(P,P_n)\leq\rho} E_P(l(\theta, x))$$

Sinha,Namkook, Duchi 2017. Robust algorithm, and certificate For the true P_0

References

- Villani, Cédric. *Topics in optimal transportation*. No. 58. American Mathematical Soc., 2003.
- Peyré, Gabriel, and Marco Cuturi. *Computational Optimal Transport*. No. 2017-86. 2017.
- Cuturi, Marco. "Sinkhorn distances: Lightspeed computation of optimal transport." *Advances in neural information processing systems*. 2013
- Amari, Shun-ichi, and Hiroshi Nagaoka. *Methods of information geometry*. Vol. 191. American Mathematical Soc., 2007.
- Benamou, Jean-David, et al. "Iterative Bregman projections for regularized transportation problems." *SIAM Journal on Scientific Computing* 37.2 (2015): A1111-A1138.
- Altschuler, Jason, Jonathan Weed, and Philippe Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration." *arXiv preprint arXiv:1705.09634* (2017).
- Genevay, Aude, Gabriel Peyré, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences" *arXiv preprint arXiv:1706.00292* (2017).
- Mena, G., Belanger, D., Linderman, S., Snoek, J. Learning Latent Permutations with Gumbel-Sinkhorn Networks. ICLR 2018

References

- Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein gan." *arXiv preprint arXiv:1701.07875* (2017).
- Bottou, Leon, et al. "Geometrical Insights for Implicit Generative Modeling." *arXiv preprint arXiv:1712.07822* (2017).
- Cominetti, Roberto, and Jaime San Martín. "Asymptotic analysis of the exponential penalty trajectory in linear programming." *Mathematical Programming* 67.1-3 (1994): 169-187.
- Cuturi, Marco, and Arnaud Doucet. "Fast computation of Wasserstein barycenters." *International Conference on Machine Learning*. 2014.
- Agueh, Martial, and Guillaume Carlier. "Barycenters in the Wasserstein space." *SIAM Journal on Mathematical Analysis*43.2 (2011): 904-924.
- Bigot, Jérémie, et al. "Geodesic PCA in the Wasserstein space by convex PCA." *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 53. No. 1. Institut Henri Poincaré, 2017.
- Fletcher, P. Thomas, et al. "Principal geodesic analysis for the study of nonlinear statistics of shape." *IEEE transactions on medical imaging* 23.8 (2004): 995-1005.
- Sinha, Aman, Hongseok Namkoong, and John Duchi. "Certifiable Distributional Robustness with Principled Adversarial Training." *arXiv preprint arXiv:1710.10571* (2017).