Weierstrass Institute for

Applied Analysis and Stochastics

An Introduction to Optimal Transport

Pavel Dvurechensky

Partially based on joint work with Darina Dvinskikh, Alexander Gasnikov, Alexey Kroshnin, Mathias Liero, Nazarii Tupitsa, Cesar Uribe

HSE-Yandex autumn school on generative models, Moscow, 26 - 29 November, 2019

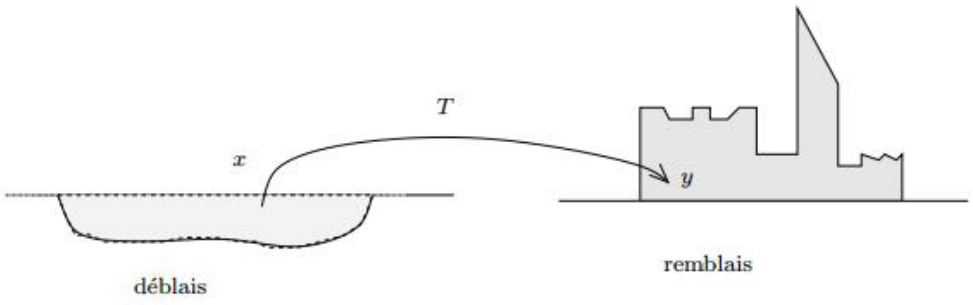
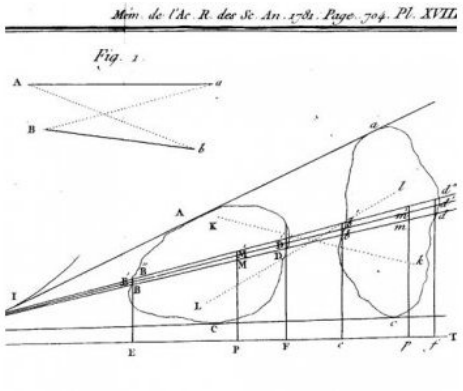


Fig. 3.1. Monge's problem of déblais and remblais



- $f(x), g(y) \geq 0$ s.t. $\int_{\mathbb{R}^d} f(x)dx = \int_{\mathbb{R}^d} g(y)dy = 1$ – mass distributions;
- $C(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_\infty$ – transportation cost function;
- Goal: transport map $T : \mathbb{R}^d \to \mathbb{R}^d$, s.t. $\forall A \subset \mathbb{R}^d$, $\int_A g(y)dy = \int_{T^{-1}(A)} f(x)dx$ minimizing

$$\int_{\mathbb{R}^d} C(x, T(x))f(x)dx.$$

G. Monge, Mémoire sur la théorie des déblais et des remblais, 1781.

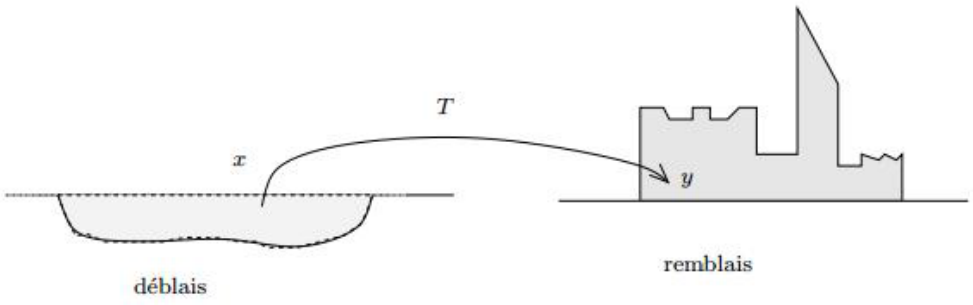
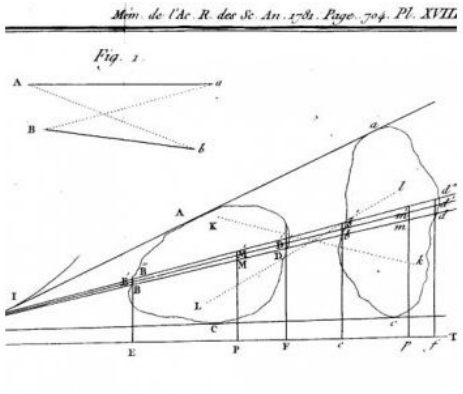


Fig. 3.1. Monge's problem of déblais and remblais



- (E, D) – metric space;
- $C(x, y) : E \times E \to \mathbb{R}_\infty$ – transportation cost function;
- $\mu, \nu \in \mathcal{P}_2(E)$ – measures to be transported;
- transport map $T : E \to E$, s.t. $\forall B$, $\mu(T^{-1}(B)) = \nu(B) \Longleftrightarrow \nu = T_{\#}\mu$.

$$\inf_T \int_E C(x, T(x))\mu(dx).$$

G. Monge, Mémoire sur la théorie des déblais et des remblais, 1781.

- **Filippo Santambrogio** Optimal Transport for Applied Mathematicians
- **Cedric Villani** Topics in Optimal Transport Optimal Transport: Old and New
- **Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré** Gradient Flows in Metric Spaces and in the Space of Probability Measures
- **Gabriel Peyré and Marco Cuturi** Computational Optimal Transport

- Highly non-linear constraint

$$\int_A g(y)dy = \int_{T^{-1}(A)} f(x)dx.$$

Change of variable leads to nonlinear PDE

$$g(T(x))\det(DT(x)) = f(x).$$

If $T(x) = \nabla u(x)$, we obtain a Monge-Ampere equation

$$\det D^2 u(x) = \frac{f(x)}{g(\nabla u(x))}.$$

NB: One of the ways to find the OT is to solve this equation.

- The mass can not be splitted and the solution does not exist in some situations

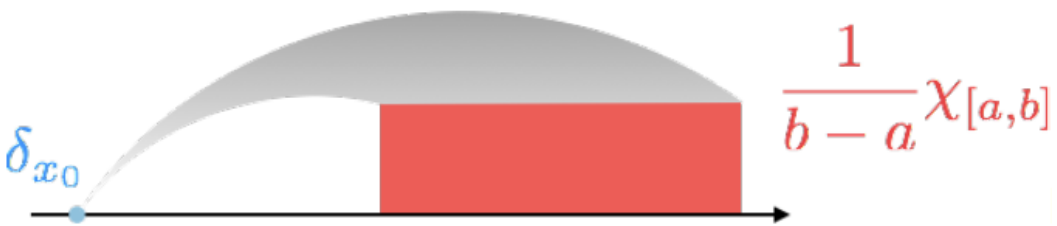


Image courtesy: Mathias Liero.

Instead of Transport map $T : E \to E$, consider transport plans $\pi \in \mathcal{P}(E \times E)$.

$\pi(x, y)$ – amount of mass transported from x to y.

Constraints become linear

$$\mathcal{U}(\mu, \nu) = \left\{ \pi \in \mathcal{P}(E \times E) : \int_E \pi(x, y) dy = \mu(x), \quad \int_E \pi(x, y) dx = \nu(y) \right\}.$$

$$\inf_{\pi \in \mathcal{U}(\mu,\nu)} \int_{E \times E} C(x, y) d\pi(x, y).$$

- Linear objective and linear constraints, but larger dimension.
- Feasible set is compact in weak* topology and solution exists.
- Probabilistic interpretation

$$\min_{\pi:(X,Y)\sim\pi} \mathbb{E}_\pi[C(X, Y)], \quad \text{s.t.} \quad X \sim \mu, \; Y \sim \nu.$$

L. Kantorovich, On the transfer of masses, 1942.

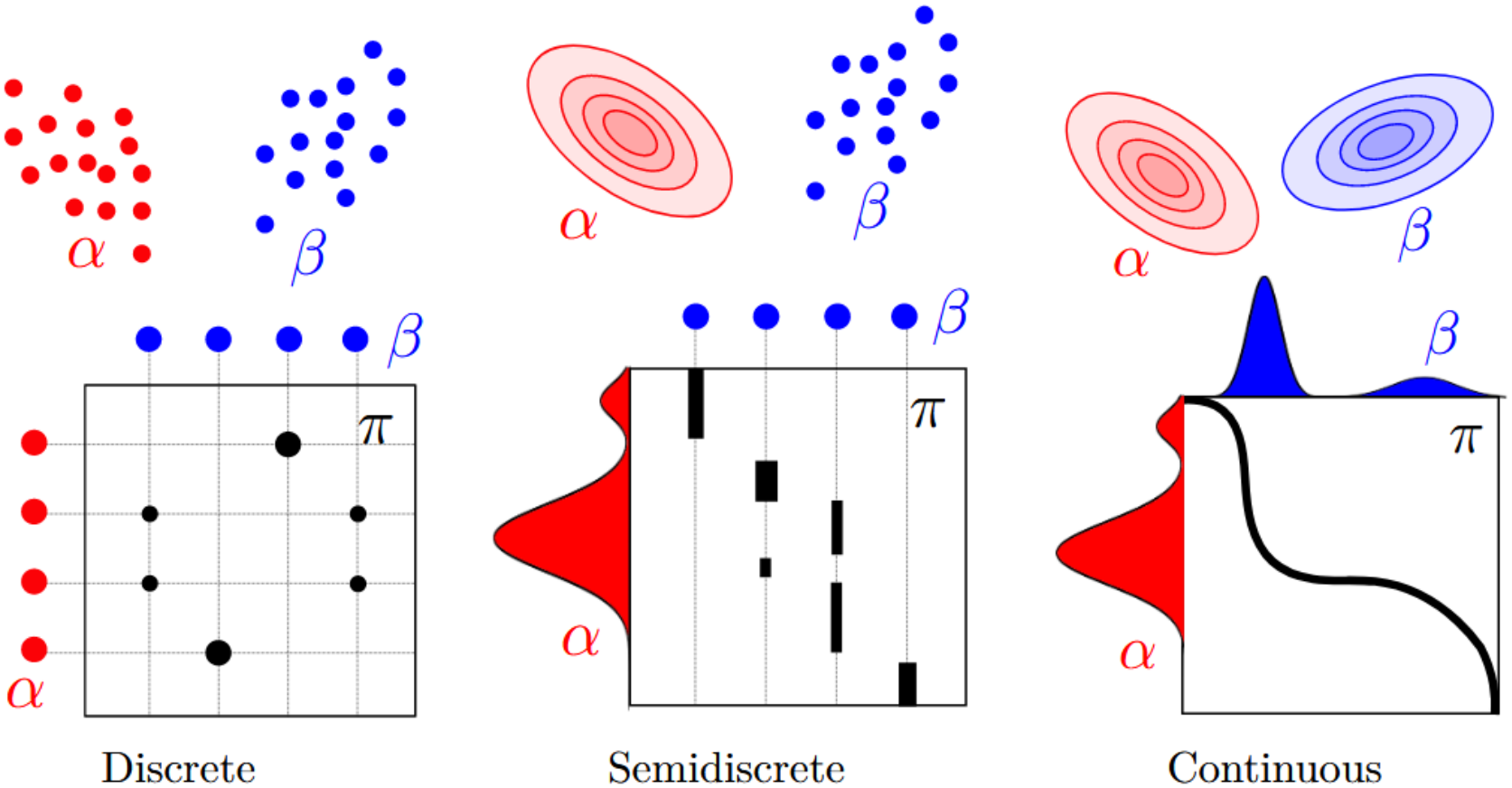
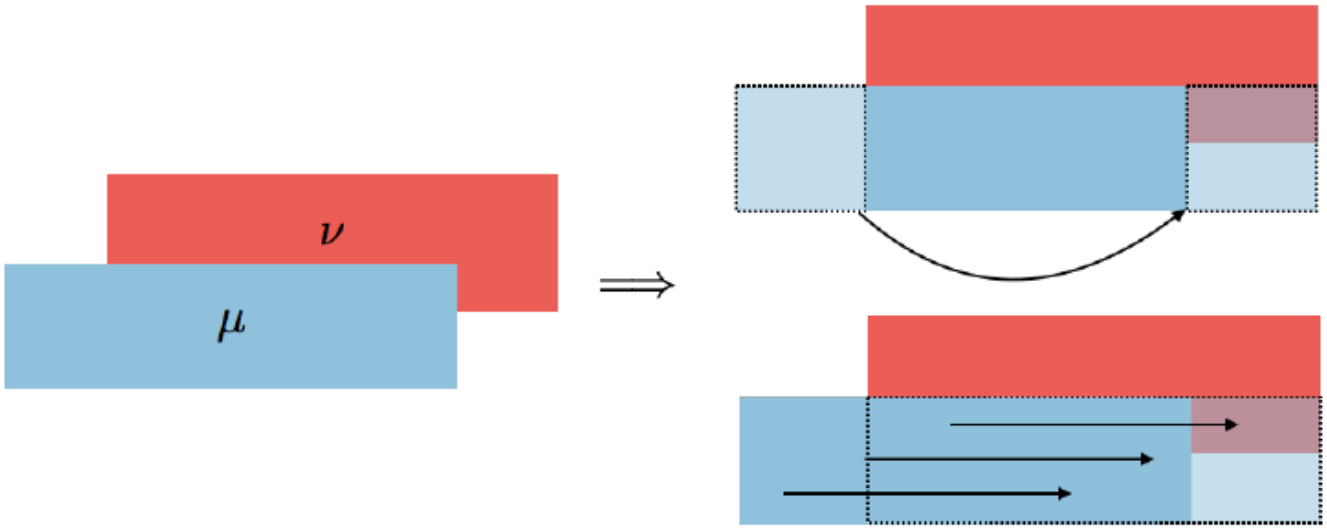


Image: Gabriel Peyré and Marco Cuturi. Computational Optimal Transport.

- Solves mass splitting problem in Monge formulation
- Monge transport maps are covered with transport plans of the form $\pi_T = (Id, T)_{\#}\mu$, i.e. $\pi_T(A \times B) = \mu(\{x \in A : T(x) \in B\})$.

book shifting for $C(x,y) = |x-y|$



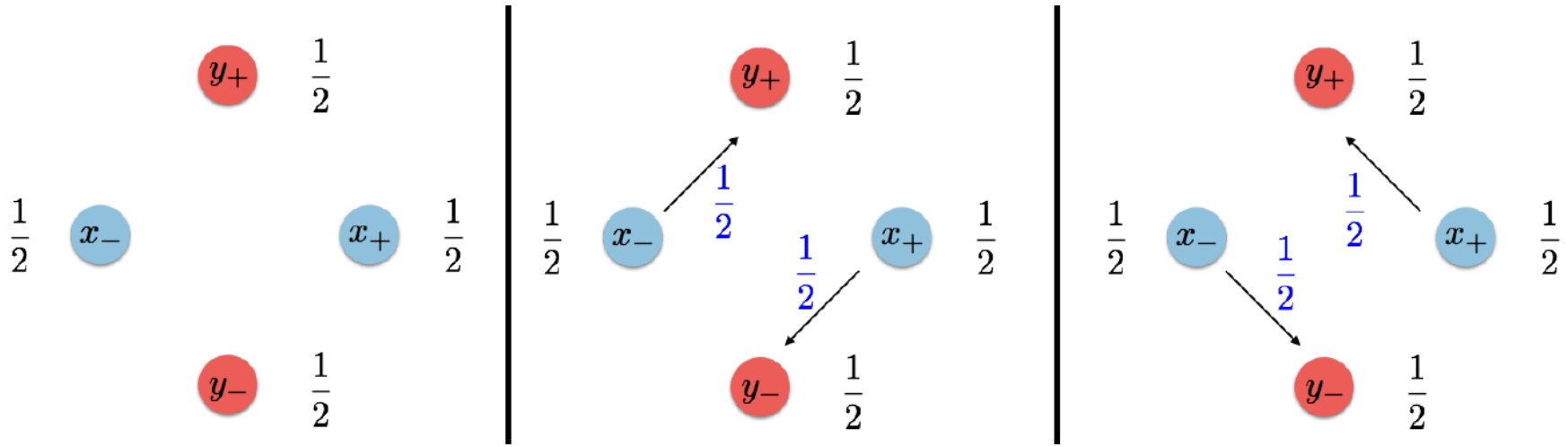


Image courtesy: Mathias Liero.

$E = \mathbb{R}^d$, cost function $C(x, y) = \|x - y\|^p$, $p \in [1, +\infty)$ leads to Monge-Kantorovich (Wasserstein) distances

$$(\mathcal{W}_p(\mu, \nu))^p = \inf_{\pi \in \mathcal{U}(\mu,\nu)} \int_{E \times E} \|x - y\|^p d\pi(x, y).$$

- $\mathcal{W}_p$ is a distance on the space of probability measures with finite p-th moments $\mathcal{P}_p(\mathbb{R}^d)$.
- Topology is equivalent to weak* convergence and convergence of p-th moments.
- Has Riemannian structure (Benamou-Brenier formulation, Otto calculus).
- Geodesic curves as shortest connecting curves between two probability measures.
- For $p = 1$ is known as Earth mover's distance.

$$(\mathcal{W}_p(\mu,\nu))^p = \inf_{\pi\in\mathcal{U}(\mu,\nu)} \int_{E\times E} \|x-y\|^p d\pi(x,y).$$

- Natural distance between probability measures and a tool for comparing them.
- Gives rise to a topological space with Riemannian structure.
- "Horizontal" distance as opposed to L^p distances

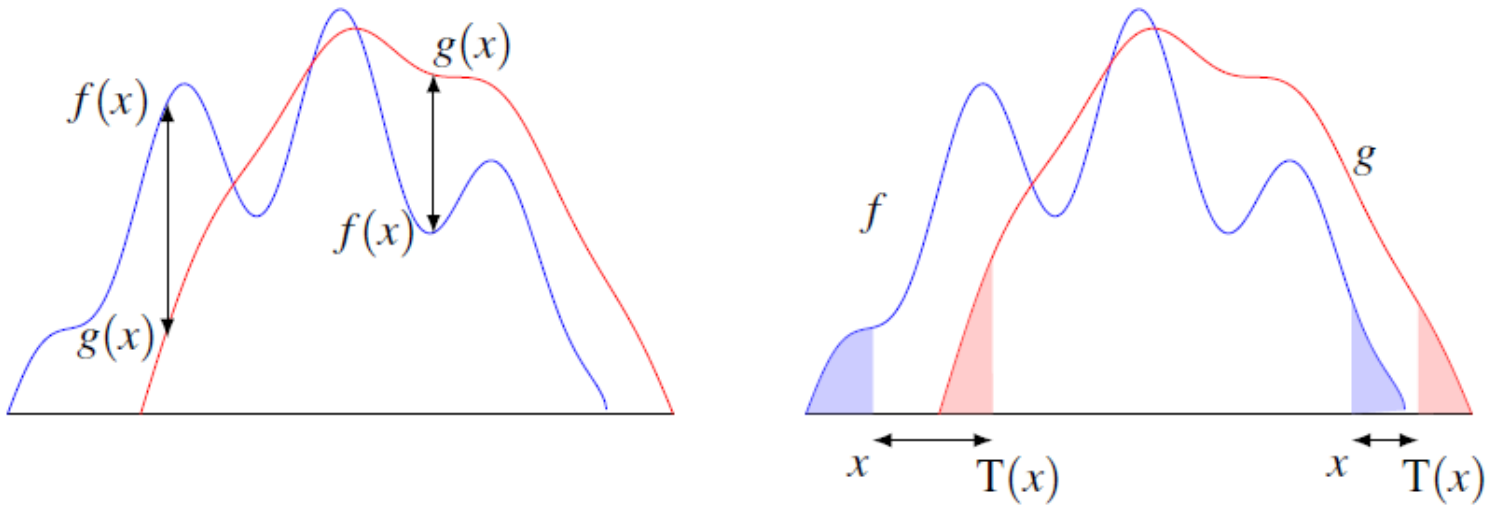


Image courtesy: Mathias Liero.

- In ML we need to find a probability measure which approximates the data distribution. We can use this distance as a regularizer.

$$\min_{\pi} \left\{ \int_{E\times E} C(x,y)d\pi(x,y) : \pi^1_{\#} = \mu, \pi^2_{\#} = \nu \right\}.$$

Lagrange multipliers $\xi : E \to \mathbb{R}$, $\eta : E \to \mathbb{R}$ (Kantorovich potentials)

$$\min_{\pi\in\mathcal{M}(E\times E)} \left\{ \int_{E\times E} C(x,y)d\pi(x,y) + \sup_{\xi,\eta} \left\{ \int_E \xi(x)d(\mu - \pi^1_{\#}) + \int_E \eta(y)d(\nu - \pi^2_{\#}) \right\} \right\}$$

$$= \sup_{\xi,\eta} \left\{ \int_E \xi(x)d\mu + \int_E \eta(y)d\nu + \inf_{\pi} \left\{ \int_{E\times E} (C(x,y) - \xi(x) - \eta(y))d\pi(x,y) \right\} \right\}$$

$$= \sup_{\xi,\eta} \left\{ \int_E \xi(x)d\mu + \int_E \eta(y)d\nu : C(x,y) - \xi(x) - \eta(y) \geq 0 \right\}$$

Economic interpretation: Outsourcing transport to a vendor.

- $\xi(x)$ vendor price for transportation of a unit mass from x;
- $\eta(y)$ vendor price for transportation of a unit mass to y;
- Vendor prices are reasonable if $\xi(x) + \eta(y) \leq C(x,y)$;
- Vendor maximizes the profit.

Theorem [Brenier, 1991]. Assume:

- Let $E = \mathbb{R}^d$, $C(x, y) = \|x - y\|^2$.
- One of the measures (say, μ) has a density w.r.t. Lebesgue measure.

Then:

- Optimal transport plan π in the Kantorovich formulation is unique and is supported on the graph $(x, T(x))$ of a transport map in the Monge formulation.
- $\pi = (Id, T)_{\#}\mu$, i.e.

$$\forall h \in \mathcal{C}(E \times E), \int_{E\times E} h(x, y)d\pi(x, y) = \int_E h(x, T(x))d\mu(x).$$

- T is uniquely defined as the gradient of a convex function φ, $T(x) = \nabla\varphi(x)$, where $\varphi(x)$ is unique (up to an additive constant) convex function s.t. $(\nabla\varphi)_{\#}\mu = \nu$.
- $\varphi(x) = \frac{\|x\|^2}{2} - \xi(x)$, ξ – optimal dual solution.

Assume that $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$, $\nu = \mathcal{N}(m_\nu, \Sigma_\nu)$ in $\mathbb{R}^d$. Let $A = \Sigma_\mu^{-\frac{1}{2}} \left(\Sigma_\mu^{\frac{1}{2}} \Sigma_\nu \Sigma_\mu^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_\mu^{-\frac{1}{2}}$. NB: $\Sigma_\mu = \Sigma_\nu \Longrightarrow A = I$.

$$T : x \to m_\nu + A(x - m_\mu) \text{ is s.t. } T_\# \rho_\mu = \rho_\nu.$$

$$T(x) = \nabla \left(\frac{1}{2} \langle x - m_\mu, A(x - m_\mu) \rangle + \langle m_\nu, x \rangle \right).$$

$$\mathcal{W}_2^2(\mu, \nu) = \|m_\mu - m_\nu\|^2 + \operatorname{tr}\left(\Sigma_\mu + \Sigma_\nu - 2(\Sigma_\mu^{1/2} \Sigma_\nu \Sigma_\mu^{1/2})^{1/2} \right).$$

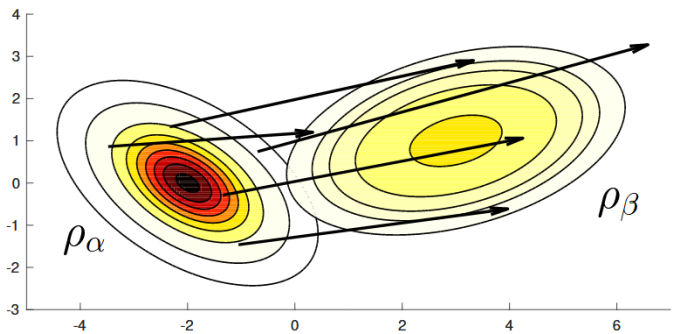


Image: Gabriel Peyré and Marco Cuturi. Computational Optimal Transport.

$$(\mathcal{W}_2(\mu_0,\mu_1))^2 = \min_{\pi}\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^2 d\pi(x,y) : \pi^1_{\#} = \mu_0, \pi^2_{\#} = \mu_1\right\}.$$

Benamou-Brenier "Dynamical" equivalent formulation: Consider connecting curves $t \to \mu_t$

$$(\mathcal{W}_2(\mu_0,\mu_1))^2 = \min_{\pi}\left\{\int_0^1\int_{\mathbb{R}^d}\|v_t(x)\|^2 d\mu_t dt : \mu_{t=0} = \mu_0, \mu_{t=1} = \mu_1, \right.$$
$$\left.\frac{d}{dt}\mu_t + \operatorname{div}(v_t\mu_t) = 0\right\}.$$

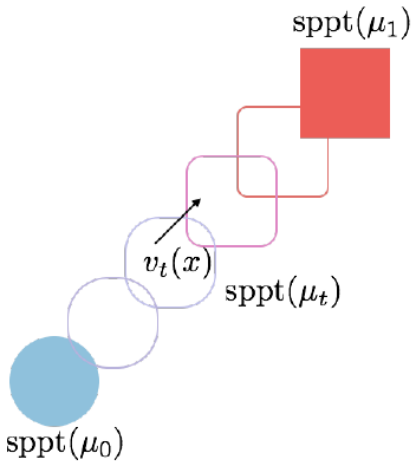
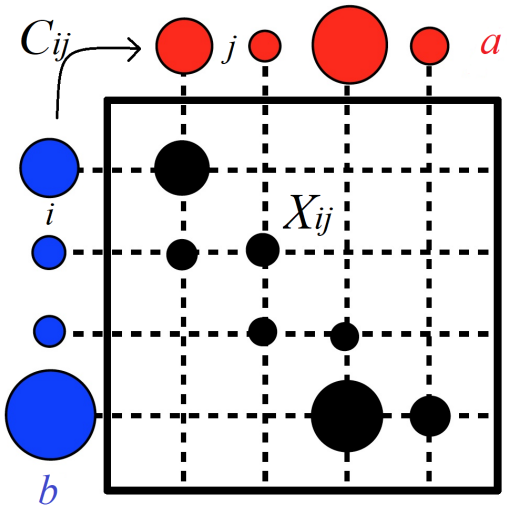


Image courtesy: Mathias Liero.

- $x_i \in \mathbb{R}^d$, $i = 1, ..., n$ – support of μ;
- $y_i \in \mathbb{R}^d$, $i = 1, ..., n$ – support of ν;
- $\mu = \sum_{i=1}^n a_i \delta(x_i), \quad a \in S_n(1)$;
- $\nu = \sum_{j=1}^n b_j \delta(y_j), \quad b \in S_n(1)$;
- $C_{ij} = C(x_i, y_j), \quad i, j = 1, ..., n$ – ground cost matrix;
- $X_{ij} = \pi(x_i, y_j), \quad i, j = 1, ..., n$ – transportation plan;



Optimal transport problem

$$\min_{X \in \mathcal{U}(a,b)} \langle C, X \rangle,$$

$$\mathcal{U}(a, b) := \{X \in \mathbb{R}_+^{n \times n} : X\mathbf{1} = a, \; X^T\mathbf{1} = b\}.$$

$$\min\left\{\langle C, X\rangle \ : X \in \mathbb{R}_+^{n\times n},\ X\mathbf{1} = a,\ X^T\mathbf{1} = b\right\}.$$

Lagrange multipliers $\xi, \eta \in \mathbb{R}^n$ (Kantorovich potentials)

$$\begin{aligned}
&\min_{X\in\mathbb{R}_+^{n\times n}} \left\{\langle C, X\rangle + \max_{\xi,\eta}\left\{\langle \xi, a - X\mathbf{1}\rangle + \langle \eta, b - X^T\mathbf{1}\rangle\right\}\right\} \\
&= \max_{\xi,\eta}\left\{\langle \xi, a\rangle + \langle \eta, b\rangle + \min_{X\in\mathbb{R}_+^{n\times n}}\left\{\langle C + \xi\mathbf{1}^T + \mathbf{1}\mu^T, X\rangle\right\}\right\} \\
&= \max_{\xi,\eta}\left\{\langle \xi, a\rangle + \langle \eta, b\rangle \ :\ C_{i,j} - \xi_i - \eta_j \geq 0\right\} \\
&= \max_{\xi}\left\{\langle \xi, a\rangle + \sum_{i=1}^n b_i \min_{j=1,\dots,n}\{C_{i,j} - \xi_i\}\right\}
\end{aligned} \tag{1}$$

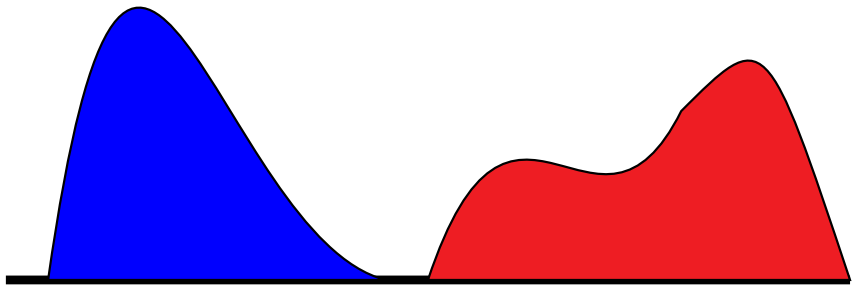
- $y_i \in \mathbb{R}^d$, $i = 1, ..., n$ – support of ν;
- $\nu = \sum_{j=1}^n b_j \delta(y_j), \quad b \in S_n(1)$;
- $C_j(x)$, $j = 1, ..., n$ – cost function;
- $\pi_j(x)$, $j = 1, ..., n$ – transportation plan;

$$\min_{\pi} \left\{ \sum_{j=1}^n \int_{\mathbb{R}^d} c_j(x) d\pi_j(x) : \sum_{j=1}^n \pi_j(x) = \mu, \int_{\mathbb{R}^d} d\pi_j(x) = b_j, \ j = 1, ..., n \right\}.$$

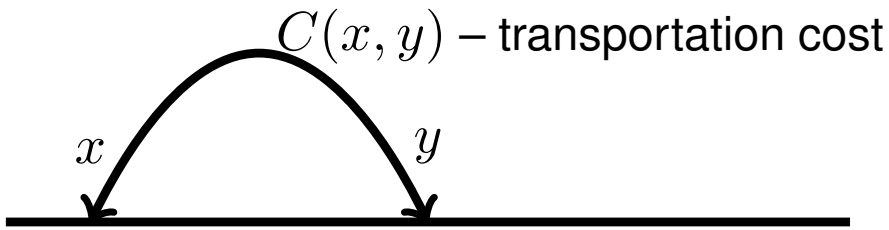
$$\min_{\pi}\left\{\sum_{j=1}^{n}\int_{\mathbb{R}^d} c_j(x)d\pi_j(x) : \sum_{j=1}^{n}\pi_j(x)=\mu,\ \int_{\mathbb{R}^d} d\pi_j(x)=b_j,\ j=1,...,n\right\}.$$

Lagrange multipliers $\xi:\mathbb{R}^d\to\mathbb{R}$, $\eta\in\mathbb{R}^n$ (Kantorovich potentials)

$$\begin{aligned}
&\min_{\pi}\left\{\sum_{j=1}^{n}\int_{\mathbb{R}^d} c_j(x)d\pi_j(x)\right.\\
&\qquad\left.+\sup_{\xi,\eta}\left\{\int_{\mathbb{R}^d}\xi(x)d\left(\mu-\sum_{j=1}^{n}\pi_j(x)\right)+\sum_{i=1}^{n}\eta_j\left(b_j-\int_{\mathbb{R}^d}d\pi_j(x)\right)\right\}\right\}\\
&=\sup_{\xi,\eta}\left\{\int_{\mathbb{R}^d}\xi(x)d\mu+\sum_{i=1}^{n}\eta_j b_j+\inf_{\pi}\left\{\sum_{j=1}^{n}\int_{\mathbb{R}^d}(C_j(x)-\xi(x)-\eta_j)d\pi_j(x)\right\}\right\}\\
&=\sup_{\eta\in\mathbb{R}^n}\left\{\int_{\mathbb{R}^d}\min_j\{C_j(x)-\eta_j\}d\mu+\sum_{i=1}^{n}\eta_j b_j\right\}\\
&=\sup_{\eta\in\mathbb{R}^n}\left\{\mathbb{E}_{X\sim\mu}\min_j\{C_j(X)-\eta_j\}+\sum_{i=1}^{n}\eta_j b_j\right\}
\end{aligned}$$



$\mathcal{P}(E)$ – space of measures on E



Goal: given an image, find similar in the database

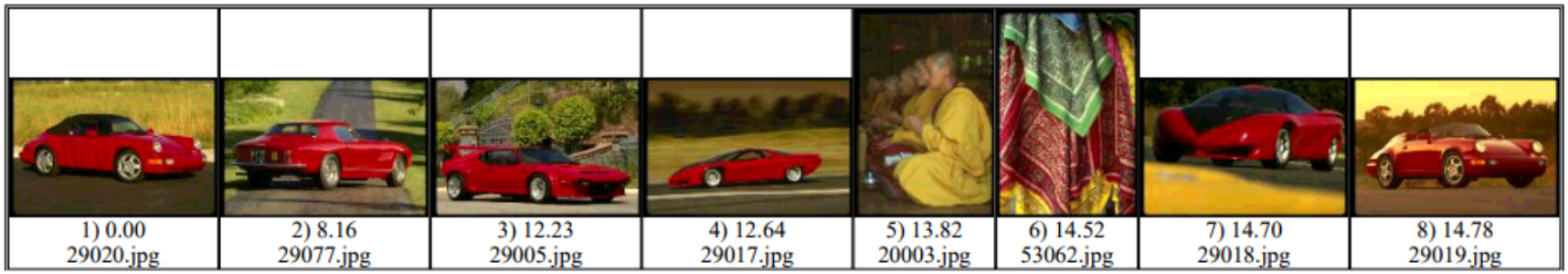
Basis space – CIELAB color space

Cost – Euclidean distance

(short Euclidean distances correlate strongly with human color discrimination performance)

Measures – histograms given by bins

		red			
		0-63	64-127	128-191	192-255
blue	0-63	43	78	18	0
	64-127	45	67	33	2
	128-191	127	58	25	8
	192-255	140	47	47	13



Rubner, Tomasi, Guibas. The earth mover's distance as a metric for image retrieval. 2000.

`https://en.wikipedia.org/wiki/Color_histogram`

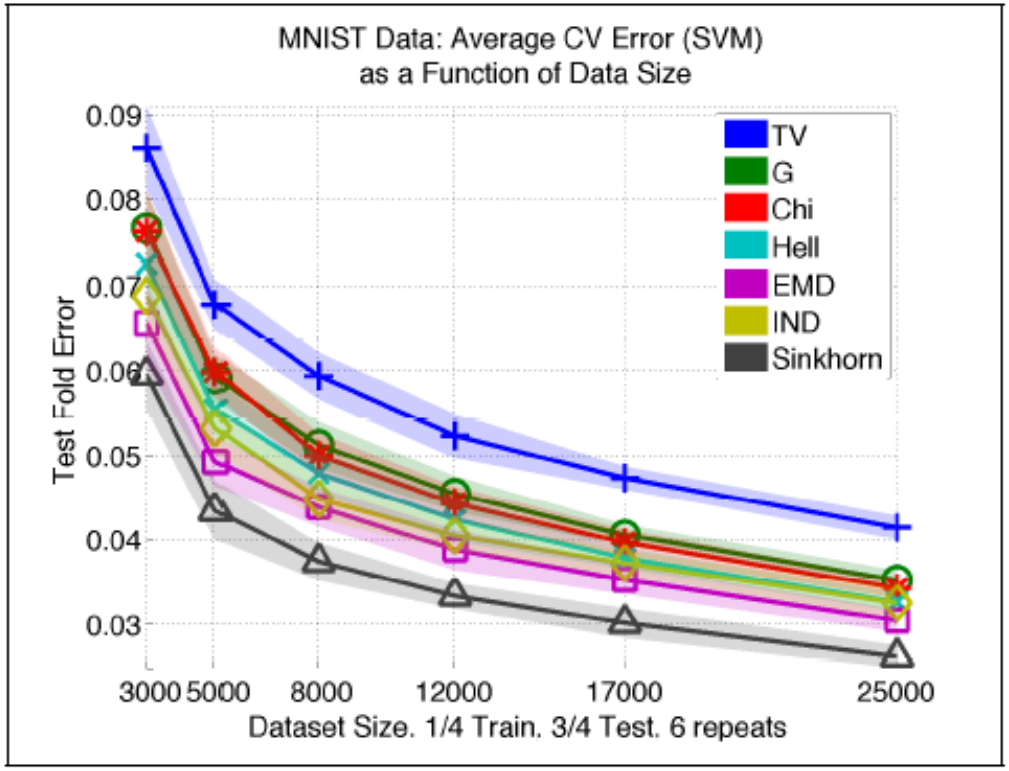
Goal: classify images from MNIST dataset 

Basis space – pixel grid

Cost – Squared Euclidean distance

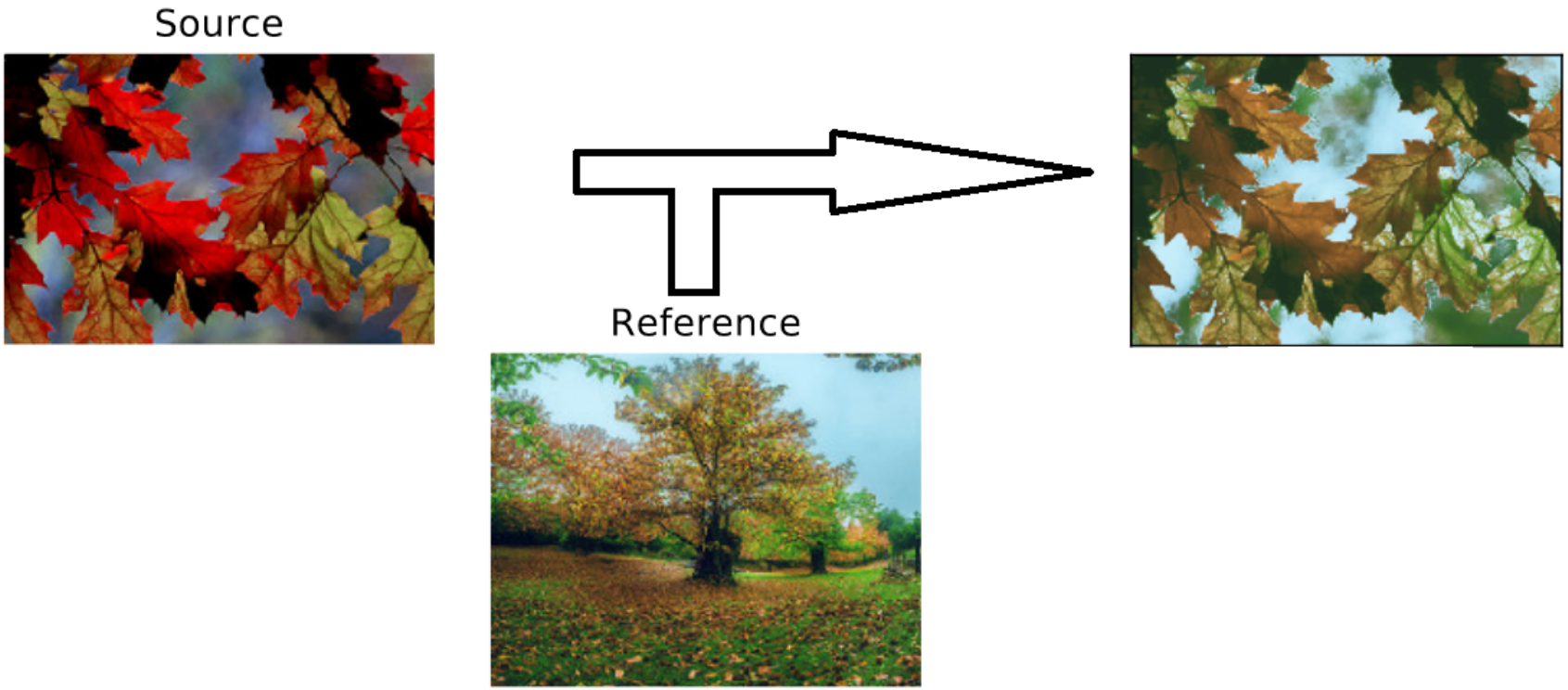
Measures – histograms of pixel intensities

Run standard SVM based on distance between images



Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. NIPS 2013.

Goal: transfer color from one image to another

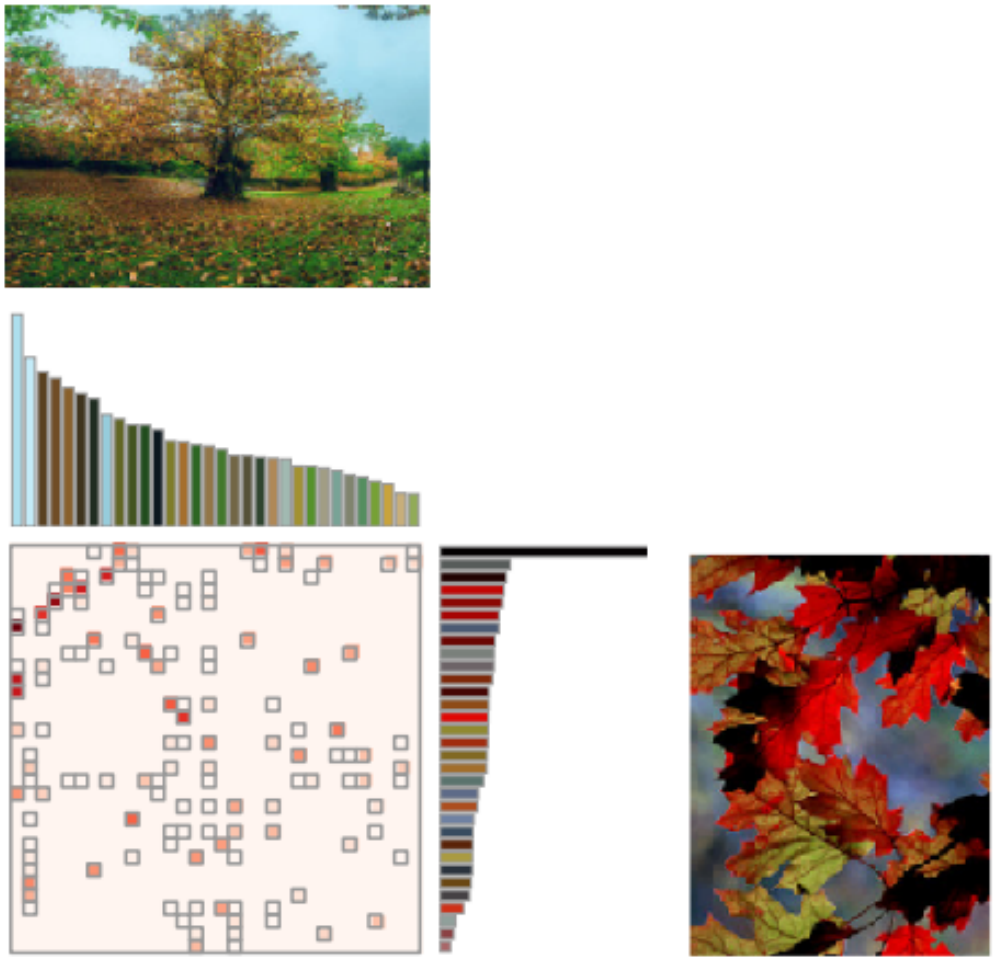


Blondel, Seguy, Rolet. Smooth and Sparse Optimal Transport. AISTATS 2018.

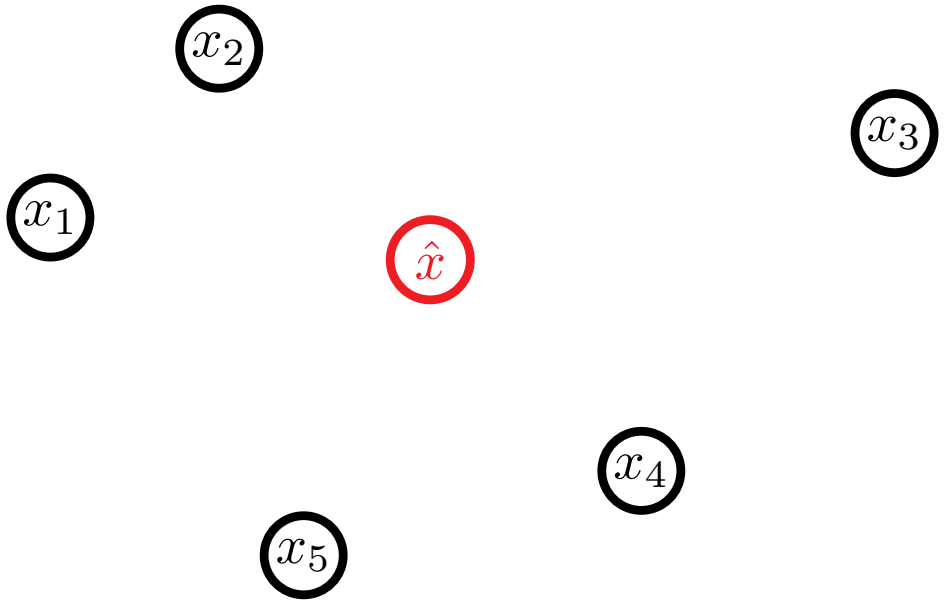
Basis space – RGB color space

Cost – Squared Euclidean distance

Measures – histograms given by clustering

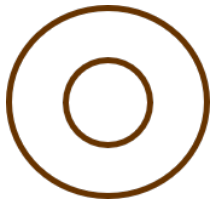


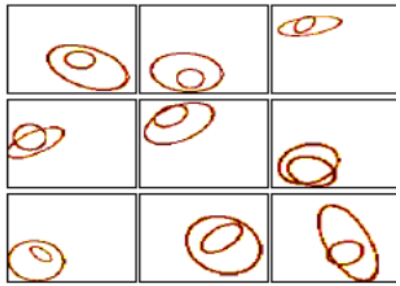
Blondel, Seguy, Rolet. Smooth and Sparse Optimal Transport. AISTATS 2018.



$$\hat{x} = \frac{1}{m}\sum_{i=1}^{m} x_i = \arg\min_x \frac{1}{m}\sum_{i=1}^{m} \|x - x_i\|_2^2.$$

Goal: reconstruct template from its random transformations



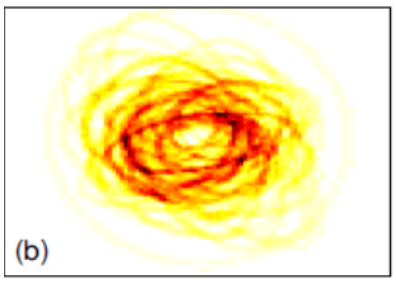


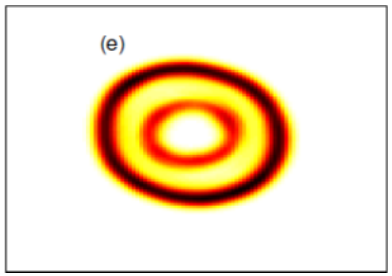
Basis space – pixel grid

Cost – Squared Euclidean distance

Measures – histograms given by intensity of pixels

$$\widehat{I} = \arg\min_{I} \frac{1}{m} \sum_{i=1}^{m} \mathrm{Dist}(I, I_i).$$

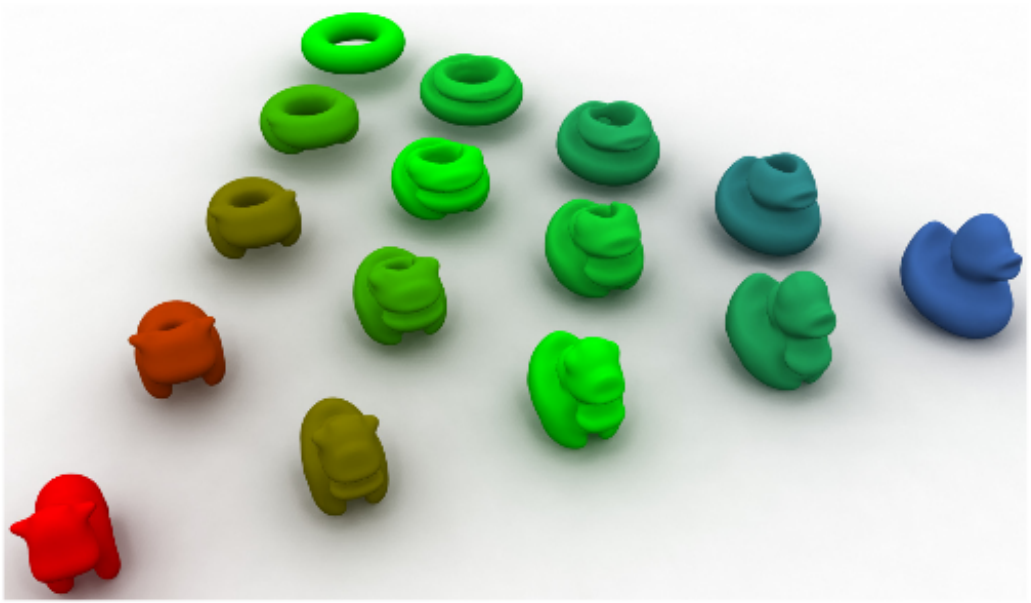




Euclidean distance OT distance

Cuturi, Doucet. Fast Computation of Wasserstein Barycenters. ICML 2014.

$$\mu_t = \arg\min_{\mu} \left((1-t)\widetilde{\mathcal{W}}_2^2(\mu_0, \mu) + t\widetilde{\mathcal{W}}_2^2(\mu, \mu_1) \right).$$



Solomon, de Goes, Peyré, Cuturi, Butscher, Nguyen, Du, Guibas Convolutional wasserstein distances: efficient optimal transportation on geometric domains. ACM Trans. Graph. 2015.

$$\text{Find} \quad \widehat{X} \in \mathcal{U}(a,b) \quad \text{s.t.} \quad \langle C, \widehat{X} \rangle \leq \min_{X \in \mathcal{U}(a,b)} \langle C, X \rangle + \varepsilon,$$

$$\mathcal{U}(a,b) := \{X \in \mathbb{R}_+^{n \times n} : \, X\mathbf{1} = a, \, X^T\mathbf{1} = b\}.$$

- Linear programming problem with complexity $O(n^3 \ln n)$ arithmetic operations [Pele & Werman, 2009].
- Widespread approach [Cuturi, 2013]. Solve by Sinkhorn's algorithm an entropy-regularized optimal transport problem

$$\min_{X \in \mathcal{U}(a,b)} \langle C, X \rangle + \gamma \langle X, \ln X \rangle.$$

- Complexity by accelerated gradient descent and by Sinkhorn's algorithm [D., Gasnikov, Kroshnin, 2018], resp.

$$\widetilde{O}\left(\frac{n^{2.5}}{\varepsilon}\right), \quad \widetilde{O}\left(\frac{n^2}{\varepsilon^2}\right).$$

Primal problem

$$\min_{X\in\mathcal{U}(a,b)} \langle C, X\rangle + \gamma\langle X, \ln X\rangle = \min_{X\in\mathcal{U}(a,b)} \gamma(-\langle X, \ln \mathrm{e}^{-\frac{C}{\gamma}}\rangle + \langle X, \ln X\rangle)$$
$$= \min_{X\in\mathcal{U}(a,b)} \gamma KL\left(X, \mathrm{e}^{-\frac{C}{\gamma}}\right).$$

$$\mathcal{U}(a,b) = \{X \in \mathbb{R}_+^{n\times n} : X\mathbf{1} = a,\ X^T\mathbf{1} = b\}$$

Dual problem

$$\max_{\xi,\eta} -\gamma \sum_{i,j=1}^{n} \exp\left(-\frac{1}{\gamma}(C_{ij} - \xi_i - \eta_j)\right) + \langle \xi, a\rangle + \langle \eta, b\rangle$$

Cf. with dual non-regularized problem

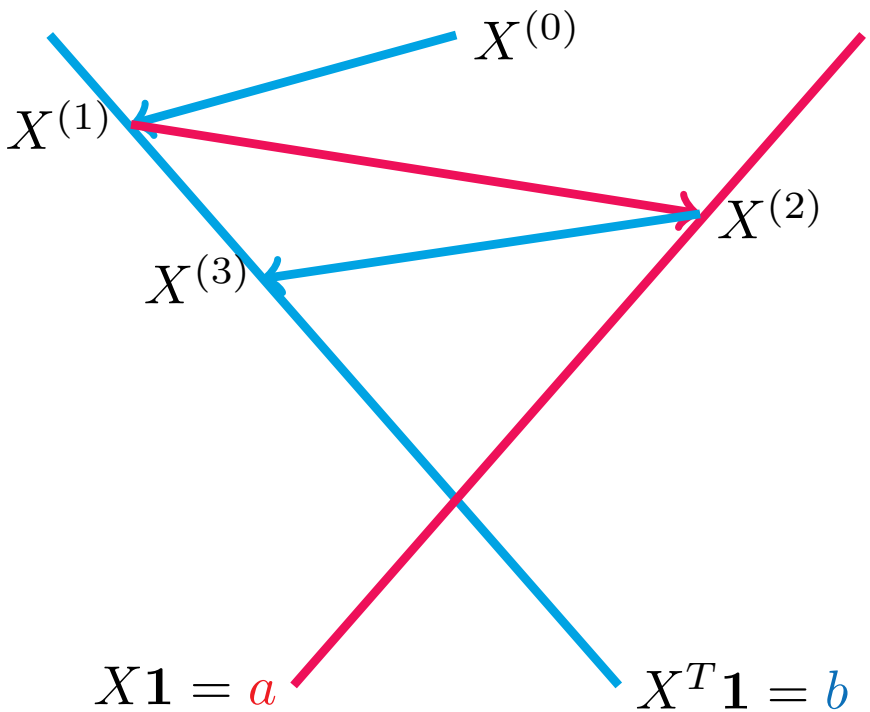
$$\max_{\xi,\eta}\{\langle \xi, a\rangle + \langle \eta, b\rangle | C_{ij} - \xi_i - \eta_j \geq 0\}$$

NB: Regularization introduces error $\gamma\langle X, \ln X\rangle \in [-\gamma \ln(n^2), 0] \Longrightarrow$ we need to take $\gamma = \Theta(\varepsilon/\ln n)$.

Primal problem $$\min_{X \in \mathbb{R}_+^{n \times n}} \left\{ KL\left(X, \mathrm{e}^{-\frac{C}{\gamma}}\right) | X\mathbf{1} = a,\ X^T\mathbf{1} = b \right\}.$$

Alternating minimization/projection $X^0 = \mathrm{e}^{-\frac{C}{\gamma}}$

$$X^{(k+1)} = \arg\min_{X: X\mathbf{1}=a} KL\left(X, X^{(k)}\right), \quad X^{(k+2)} = \arg\min_{X: X^T\mathbf{1}=b} KL\left(X, X^{(k+1)}\right)$$



Dual problem

$$\max_{\xi,\eta} -\gamma \sum_{i,j=1}^{n} \exp\left(-\frac{1}{\gamma}(C_{ij} - \xi_i - \eta_j)\right) + \langle \xi, a\rangle + \langle \eta, b\rangle$$

$$= \max_{\xi,\eta} -\gamma \left(\mathrm{e}^{\frac{\xi}{\gamma}}\right)^T \mathrm{e}^{-\frac{C}{\gamma}} \mathrm{e}^{\frac{\eta}{\gamma}} + \langle \xi, a\rangle + \langle \eta, b\rangle$$

Optimality conditions (gradient equal to 0)

$$\mathrm{diag}\left(\mathrm{e}^{\frac{\xi}{\gamma}}\right) \mathrm{e}^{-\frac{C}{\gamma}} \mathrm{e}^{\frac{\eta}{\gamma}} = a$$

$$\mathrm{diag}\left(\mathrm{e}^{\frac{\eta}{\gamma}}\right) \left(\mathrm{e}^{-\frac{C}{\gamma}}\right)^T \mathrm{e}^{\frac{\xi}{\gamma}} = b$$

Alternating minimization in ξ, η

$$\xi^{(k+1)} = \gamma \ln \frac{a}{\mathrm{e}^{-\frac{C}{\gamma}} \mathrm{e}^{\frac{\eta^{(k)}}{\gamma}}} \qquad \eta^{(k+1)} = \gamma \ln \frac{b}{\left(\mathrm{e}^{-\frac{C}{\gamma}}\right)^T \mathrm{e}^{\frac{\xi^{(k+1)}}{\gamma}}}.$$

$$\max_{\xi,\eta} -\gamma \left(\mathrm{e}^{\frac{\xi}{\gamma}} \right)^T \mathrm{e}^{-\frac{C}{\gamma}} \mathrm{e}^{\frac{\eta}{\gamma}} + \langle \xi, a \rangle + \langle \eta, b \rangle$$

$$= -\gamma \left(\min_{\xi,\eta} \left(\mathrm{e}^{\frac{\xi}{\gamma}} \right)^T \mathrm{e}^{-\frac{C}{\gamma}} \mathrm{e}^{\frac{\eta}{\gamma}} - \langle \frac{\xi}{\gamma}, a \rangle - \langle \frac{\eta}{\gamma}, b \rangle \right)$$

$$= -\gamma \left(\min_{u,v} \left(\mathrm{e}^{u} \right)^T \mathrm{e}^{-\frac{C}{\gamma}} \mathrm{e}^{v} - \langle u, a \rangle - \langle v, b \rangle \right)$$

$$= -\gamma \left(\min_{u,v} \psi(u,v) := \mathbf{1}^T B(u,v) \mathbf{1} - \langle u, a \rangle - \langle v, b \rangle \right),$$

where $K := e^{-C/\gamma}$ and $B(u,v) := \mathrm{diag}(e^u) K \, \mathrm{diag}(e^v)$

$$\xi^{(k+1)} = \gamma \ln \left(a \Big/ \left(K \, \mathrm{e}^{\frac{\eta^{(k)}}{\gamma}} \right) \right) \qquad \eta^{(k+1)} = \gamma \ln \left(b \Big/ \left(K^T \, \mathrm{e}^{\frac{\xi^{(k+1)}}{\gamma}} \right) \right).$$

$$u^{(k+1)} = \ln \left(a \Big/ \left(K \, \mathrm{e}^{v^{(k)}} \right) \right) \qquad v^{(k+1)} = \ln \left(b \Big/ \left(K^T \, \mathrm{e}^{u^{(k+1)}} \right) \right).$$

$$\tilde{u}^{(k+1)} = a \Big/ K \tilde{v}^{(k)} \qquad \tilde{v}^{(k+1)} = b \Big/ K^T \, \tilde{u}^{(k+1)} .$$

Primal problem $\min_{X \in \mathcal{U}(a,b)} \langle C, X \rangle + \gamma \langle X, \ln X \rangle,$

Dual problem $\min_{u,v \in \mathbb{R}^n} \left\{ \psi(u,v) := \mathbf{1}^T B(u,v) \mathbf{1} - \langle u, a \rangle - \langle v, b \rangle \right\},$

where $K := e^{-C/\gamma}$ and $B(u,v) := \operatorname{diag}(e^u) K \operatorname{diag}(e^v)$

Sinkhorn's algorithm

repeat
  if $k \bmod 2 = 0$ **then**
    $u_{k+1} = u_k + \ln(a/(B(u_k, v_k)\mathbf{1}))$, $v_{k+1} = v_k$
  else
    $v_{k+1} = v_k + \ln(b/(B(u_k, v_k)^T \mathbf{1}))$, $u_{k+1} = u_k$
  end if
  $k = k + 1$
until $\|B(u_k, v_k)\mathbf{1} - a\|_1 + \|B(u_k, v_k)^T \mathbf{1} - b\|_1 \leq \varepsilon'$

Bounds for the iterates and optimal solution [D., Gasnikov, Kroshnin, 2018]

Denote $R := -\ln\left(\nu \min_{i,j}\{a^i, b^j\}\right)$, $\nu := \min_{i,j} K^{ij} = e^{-\|C\|_\infty/\gamma}$. Then $\max_i u_k^i - \min_i u_k^i \leq R$ and the same bounds hold for v_k, u^*, v^*.

Objective residual and constraints feasibility [D., Gasnikov, Kroshnin, 2018]

Denote $\widetilde{\psi}(u, v) := \psi(u, v) - \psi(u^*, v^*)$. Then $\widetilde{\psi}(u_k, v_k) \leq R\left(\|B(u_k, v_k)\mathbf{1} - a\|_1 + \|B(u_k, v_k)^T\mathbf{1} - b\|_1\right)$.

Sinkhorn's convergence rate [D., Gasnikov, Kroshnin, 2018]

Sinkhorn's algorithm requires no more than

$$k \leq 2 + \frac{4R}{\varepsilon'}$$

iterations to find $B(u_k, v_k)$ s.t. $\|B(u_k, v_k)\mathbf{1} - a\|_1 + \|B(u_k, v_k)^T\mathbf{1} - b\|_1 \leq \varepsilon'$.

$$\psi(u_k, v_k) - \psi(u_{k+1}, v_{k+1})$$
$$= \langle \mathbf{1}, B_k \mathbf{1} \rangle - \langle \mathbf{1}, B_{k+1} \mathbf{1} \rangle + \langle u_{k+1} - u_k, a \rangle + \langle v_{k+1} - v_k, b \rangle$$
$$= \langle a, u_{k+1} - u_k \rangle = \langle a, \ln a - \ln(B_k \mathbf{1}) \rangle = KL(a \| B_k \mathbf{1})$$

$$\widetilde{\psi}(u_k, v_k) - \widetilde{\psi}(u_{k+1}, v_{k+1}) = KL\left(a \| B_k \mathbf{1}\right)$$
$$\geq \frac{1}{2} \left\| B_k \mathbf{1} - r \right\|_1^2 \geq \max \left\{ \frac{\widetilde{\psi}(u_k, v_k)^2}{2R^2}, \frac{(\varepsilon')^2}{2} \right\}.$$

$$\frac{\widetilde{\psi}(u_{k+1}, v_{k+1})}{2R^2} \leq \frac{\widetilde{\psi}(u_k, v_k)}{2R^2} - \left(\frac{\widetilde{\psi}(u_k, v_k)}{2R^2} \right)^2 \leq \frac{1}{k + \ell},$$

where $\ell = \frac{2R^2}{\widetilde{\psi}(u_1, v_1)}$. Thus $k \leq 1 + \frac{2R^2}{\widetilde{\psi}(u_k, v_k)} - \frac{2R^2}{\widetilde{\psi}(u_1, v_1)}$.

$$\widetilde{\psi}(u_{k+m}, v_{k+m}) \leq \widetilde{\psi}(u_k, v_k) - \frac{(\varepsilon')^2 m}{2}, \quad k, m \geq 0.$$

Require: Accuracy ε.

1: Set $\gamma = \frac{\varepsilon}{4 \ln n}$, $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.

2: Define $(\tilde{r}, \tilde{c}) = \left(1 - \frac{\varepsilon'}{8}\right)\left((a, b) + \frac{\varepsilon'}{n(8-\varepsilon')}(\mathbf{1}, \mathbf{1})\right)$.
NB: $\min_{i,j}\{\tilde{r}^i, \tilde{c}^j\} \geq \varepsilon'/(8n)$.

3: Calculate $B(u_k, v_k)$ by Sinkhorn's algorithm with marginals $\tilde{r}$, $\tilde{c}$ and accuracy $\varepsilon'/2$.

4: Find $\widehat{X}$ as the projection of $B(u_k, v_k)$ on $\mathcal{U}(a, b)$ by Algorithm 2 in [Altschuler et.al.,2017].

Complexity of OT by Sinkhorn [D., Gasnikov, Kroshnin, 2018]

Algorithm outputs $\widehat{X} \in \mathcal{U}(a, b)$ s.t. $\langle C, \widehat{X}\rangle \leq \min_{X \in \mathcal{U}(a,b)} \langle C, X\rangle + \varepsilon$ in

$$O\left(\frac{n^2 \|C\|_\infty^2 \ln n}{\varepsilon^2}\right) \quad \textit{arithmetic operations.}$$

Require: Matrix Y to be projected, $\mathcal{U}(a, b)$.

1: Set $X = \mathrm{diag}(x)$, where $x_i = \min\{a_i/[Y\mathbf{1}]_i, 1\}$.

2: Set $Y' = XY$.

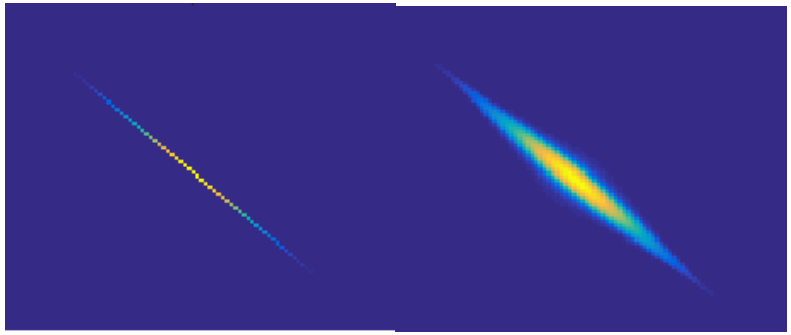
3: Set $X = \mathrm{diag}(y)$, where $y_j = \min\{b_j/[(Y')^T\mathbf{1}]_j, 1\}$.

4: Set $Y'' = Y'X$.

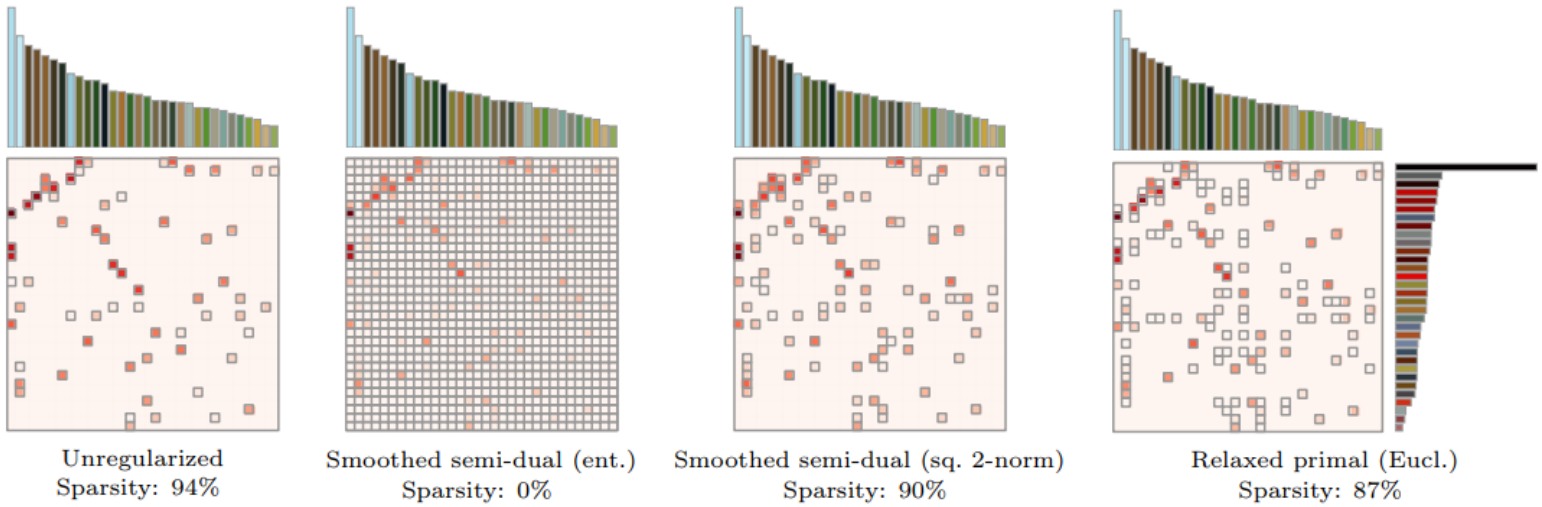
5: Set $\mathrm{err}_a = a - Y''\mathbf{1}$, $\mathrm{err}_b = b - (Y'')^T\mathbf{1}$.

6: Output $\widehat{X} = Y'' + \mathrm{err}_a\mathrm{err}_b^T/\|\mathrm{err}_a\|_1$.

- Blurring in the transportation plan.



- Dense transportation plan.



Lower image: Blondel et al., 2017

- Better than $O(n^3 \ln n)$ (LP solver) and $O(n^2 \ln n/\varepsilon^2)$ (Sinkhorn's algorithm) complexity bound.
- Flexibility w.r.t. the choice of the regularizer, e.g. squared Euclidean norm instead of the entropy.

$$\min_{x \in Q \subseteq E} \{f(x) : Ax = c\},$$

where

- E – finite-dimensional real vector space;
- Q – simple closed convex set;
- $A : E \to H, b \in H$;
- $f(x)$ is γ-strongly convex on Q w.r.t $\|\cdot\|_E$. i.e. for all $x, y \in Q$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\gamma}{2}\|x - y\|_E^2.$$

To obtain entropy-regularized optimal transport problem, set

- $E = \mathbb{R}^{n^2}, H = \mathbb{R}^{2n}, \|\cdot\|_E = \|\cdot\|_1, Q = S_{n^2}(1)$;
- $f(x) = \langle C, X \rangle + \gamma \langle X, \ln X \rangle$;
- $\{x : Ax = c\} = \{X : X\mathbf{1} = a,\ X^T\mathbf{1} = b\}$.

$$\min_{X \in \mathcal{U}(a,b)} \langle C, X \rangle + \gamma \mathcal{R}(X),$$

$$\mathcal{U}(a,b) := \{X \in \mathbb{R}_+^{n \times n} : \, X\mathbf{1} = a, \, X^T\mathbf{1} = b\}.$$

- Entropy regularization: $f(X) = \langle C, X \rangle + \gamma \langle X, \ln X \rangle$ is strongly convex w.r.t. $\| \cdot \|_1$ on $S_{n^2}(1)$.
- Squared Euclidean norm: $f(X) = \langle C, X \rangle + \gamma \|X\|_2^2$ is strongly convex w.r.t. the squared Euclidean norm.

$$\min_{x\in Q}\{f(x) : Ax = c\} = \min_{x\in Q}\left\{f(x) + \max_{\lambda\in H^*}\langle\lambda, Ax - c\rangle\right\}$$
$$= \max_{\lambda\in H^*}\left\{-\langle\lambda, c\rangle + \min_{x\in Q}\{f(x) + \langle\lambda, Ax\rangle\}\right\}$$

Dual problem

$$\min_{\lambda\in H^*}\left\{\varphi(\lambda) := \langle\lambda, c\rangle + \max_{x\in Q}\{-f(x) - \langle\lambda, Ax\rangle\}\right\}.$$

$$\nabla\varphi(\lambda) = c - Ax(\lambda), \quad x(\lambda) := \arg\max_{x\in Q}\{-f(x) - \langle\lambda, Ax\rangle\}.$$

NB: $\nabla\varphi(\lambda)$ is Lipschitz-continuous

$$\varphi(\lambda) \le \varphi(\zeta) + \langle\nabla\varphi(\zeta), \lambda - \zeta\rangle + \frac{\|A\|^2_{E\to H}}{2\gamma}\|\lambda - \zeta\|^2_{H,*}.$$

$$\min_{X \in \mathcal{U}(a,b)} \langle C, X \rangle + \gamma \langle X, \ln X \rangle$$

$$\begin{aligned} \mathcal{U}(a,b) &= \{X \in \mathbb{R}_+^{n \times n} : \; X\mathbf{1} = a, \; X^T\mathbf{1} = b\} \\ &= \{X \in S_{n^2}(1) : \; X\mathbf{1} = a, \; X^T\mathbf{1} = b\} \end{aligned}$$

Dual problem

$$\max_{\xi,\eta} -\gamma \ln \sum_{i,j=1}^{n} \exp\left(-\frac{1}{\gamma}(C_{ij} - \xi_i - \eta_j)\right) + \langle \xi, a \rangle + \langle \eta, b \rangle$$

Cf. with dual non-regularized problem

$$\max_{\xi,\eta} \{\langle \xi, a \rangle + \langle \eta, b \rangle | C_{ij} - \xi_i - \eta_j \geq 0\}$$

$$\max_{\xi,\eta} \{\langle \xi, a \rangle + \langle \eta, b \rangle + \min_{i,j} \{C_{ij} - \xi_i - \eta_j\}\}$$

$$\min_{x \in Q \subseteq E} \{f(x) : Ax = c\},$$

- Function f is γ-strongly convex.
- The problem

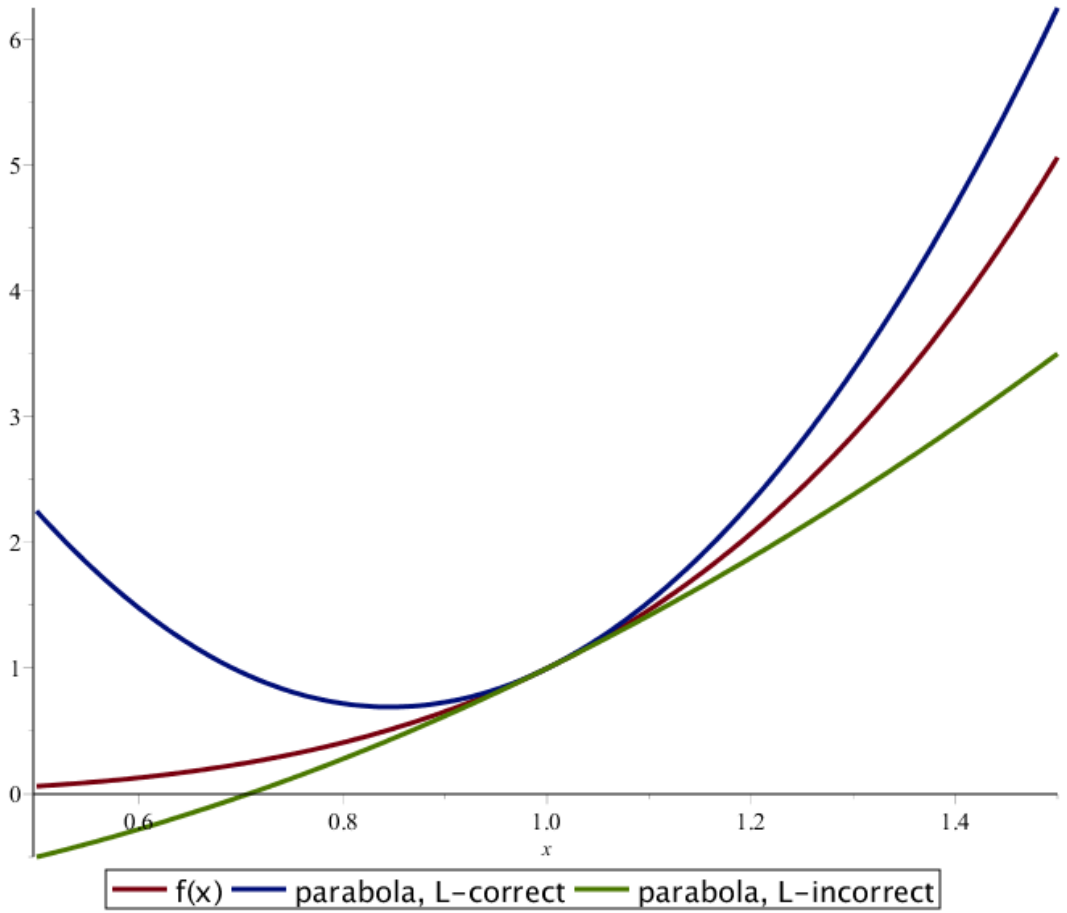
$$\max_{x \in Q} \left(-f(x) - \langle A^T \lambda, x \rangle\right)$$

  is simple: has a closed form solution or can be solved very fast up to the machine precision.
- The dual problem has a solution λ^* and there exist some $R > 0$ such that

$$\|\lambda^*\|_2 \leq R < +\infty.$$

NB: R is used *only in the convergence analysis*, but not in the algorithm itself.

$$\varphi(\lambda) \leq \varphi(\zeta) + \langle \nabla\varphi(\zeta), \lambda - \zeta \rangle + \frac{L}{2}\|\lambda - \zeta\|_{H,*}^2.$$



Require: Accuracy $\varepsilon_f, \varepsilon_{eq} > 0$, initial estimate L_0 s.t. $0 < L_0 < 2L$.

1: Set $i_0 = k = 0$, $M_{-1} = L_0$, $\beta_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.

2: **repeat** {Main iterate}

3: **repeat** {Line search}

4: Set $M_k = 2^{i_k - 1} M_k$, find α_{k+1} s.t. $\beta_{k+1} := \beta_k + \alpha_{k+1} = M_k \alpha_{k+1}^2$. Set $\tau_k = \alpha_{k+1} / \beta_{k+1}$.

5: $\lambda_{k+1} = \tau_k \zeta_k + (1 - \tau_k)\eta_k$.

6: [Update momentum] $\zeta_{k+1} = \zeta_k - \alpha_{k+1} \nabla \varphi(\lambda_{k+1})$.

7: [Gradient step] $\eta_{k+1} = \tau_k \zeta_{k+1} + (1 - \tau_k)\eta_k \sim$ $\eta_{k+1} = \lambda_{k+1} - \frac{1}{M_k} \nabla \varphi(\lambda_{k+1})$.

8: **until**

$$\varphi(\eta_{k+1}) \leq \varphi(\lambda_{k+1}) + \langle \nabla \varphi(\lambda_{k+1}), \eta_{k+1} - \lambda_{k+1} \rangle + \frac{M_k}{2} \|\eta_{k+1} - \lambda_{k+1}\|_2^2.$$

9: [Primal update] $\hat{x}_{k+1} = \tau_k x(\lambda_{k+1}) + (1 - \tau_k)\hat{x}_k$.

10: Set $i_{k+1} = 0$, $k = k + 1$.

11: **until** $f(\hat{x}_{k+1}) + \varphi(\eta_{k+1}) \leq \varepsilon_f$, $\|A\hat{x}_{k+1} - b\|_2 \leq \varepsilon_{eq}$.

Ensure: $\hat{x}_{k+1}, \eta_{k+1}$.

Assume that the objective in the primal problem is γ-strongly convex and that the dual solution λ^* satisfies $\|\lambda^*\|_2 \leq R$. Then, for $k \geq 1$, the points $\hat{x}_k$, η_k in our Algorithm satisfy

$$f(\hat{x}_k) - f^* \leq f(\hat{x}_k) + \varphi(\eta_k) \leq \frac{16\|A\|^2_{E\to H} R^2}{\gamma k^2} = O\left(\frac{1}{k^2}\right),$$

$$\|A\hat{x}_k - b\|_2 \leq \frac{16\|A\|^2_{E\to H} R}{\gamma k^2} = O\left(\frac{1}{k^2}\right),$$

$$\|\hat{x}_k - x^*\|_E \leq \frac{8}{k}\frac{\|A\|_{E\to H} R}{\gamma} = O\left(\frac{1}{k}\right),$$

where x^* and f^* are respectively an optimal solution and the optimal value in the primal problem.

1. Given a target accuracy $\varepsilon > 0$, choose $\gamma = \frac{\varepsilon}{5 \ln n}$ and apply our Algorithm to entropy-regularized OT problem

$$\min_{X \in \mathcal{U}(r,c)} \langle C, X \rangle + \gamma \langle X, \ln X \rangle$$

with stopping criterion

$$f(\widehat{X}_k) + \varphi(\eta_k) \leq \frac{\varepsilon}{10}, \quad \|\widehat{X}_k \mathbf{1} - r\|_1 + \|\widehat{X}_k^T \mathbf{1} - c\|_1 \leq \frac{\varepsilon}{10\|C\|_\infty}.$$

Output: $\widehat{X}_k$. NB: Can happen that $\widehat{X}_k \notin \mathcal{U}(r,c)$.

2. Round $\widehat{X}_k$ to $\widehat{X} \in \mathcal{U}(r,c)$ by Algorithm 2 in [Altschuler et.al.,2017].

Complexity theorem [D., Gasnikov, Kroshnin, 2018]

Total number of a.o. to obtain $\widehat{X}$ s.t. $\langle C, \widehat{X} \rangle \leq \min_{X \in \mathcal{U}(r,c)} \langle C, X \rangle + \varepsilon$ is

$$O\left(\min \left\{ \frac{n^{9/4} \sqrt{\|C\|_\infty R \ln n}}{\varepsilon}, \frac{n^2 \ln n \|C\|_\infty R}{\varepsilon^2} \right\} \right).$$

- Sinkhorn's algorithm, [Altschuler, Weed, Rigollet, 2017]

$$O\left(\frac{n^2\|C\|_\infty^3 \ln n}{\varepsilon^3}\right).$$

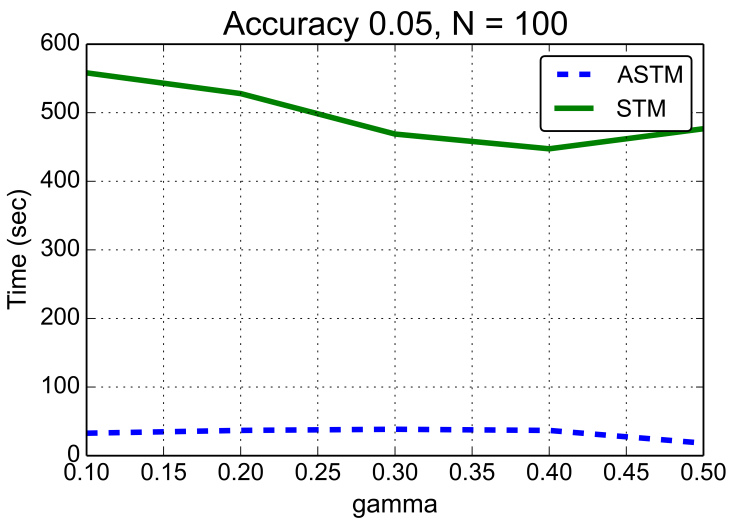
- Sinkhorn's algorithm, [D., Gasnikov, Kroshnin, 2018]

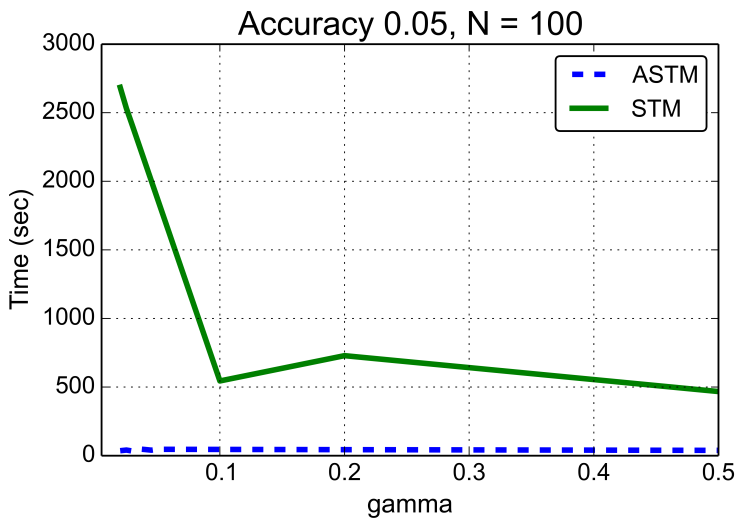
$$O\left(\frac{n^2\|C\|_\infty^2 \ln n}{\varepsilon^2}\right).$$

- It can be shown [Guminov, 2019] that $R \leq \|C\|_\infty \sqrt{n}$. Then for Accelerated Gradient Descent, [D., Gasnikov, Kroshnin, 2018]

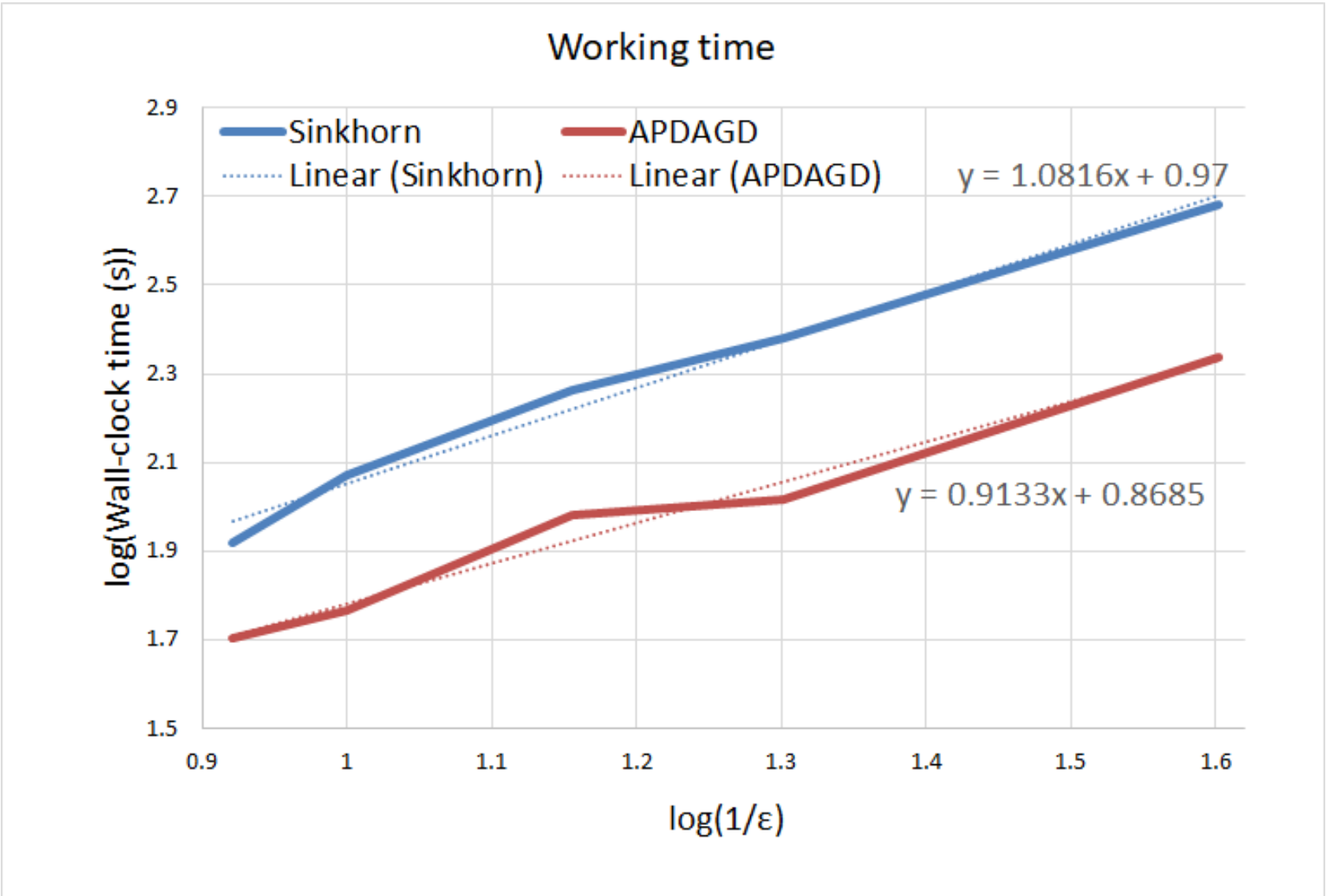
$$O\left(\frac{n^{2.5}\|C\|_\infty \sqrt{\ln n}}{\varepsilon}\right).$$

Adaptive vs non-adaptive algorithm

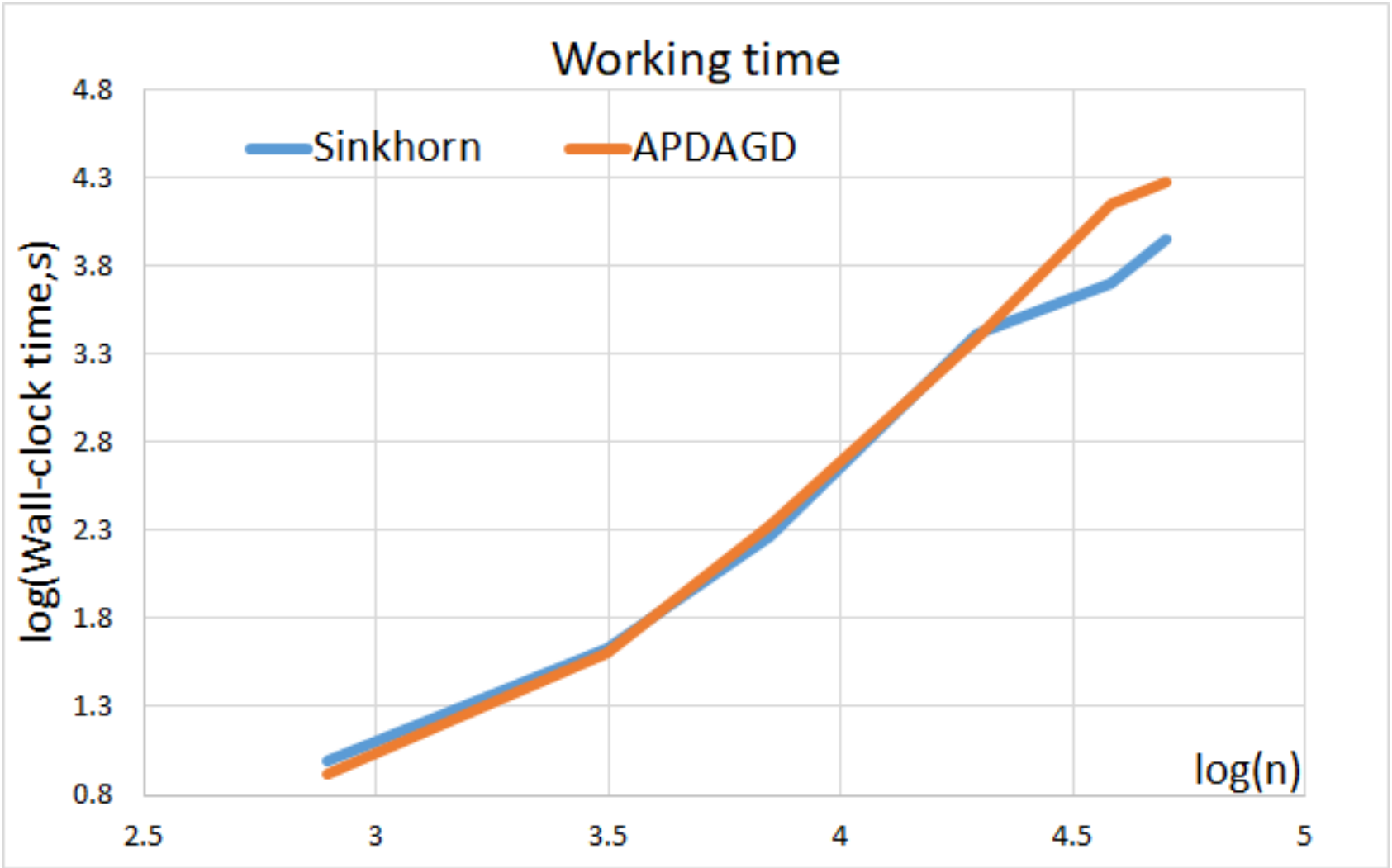




Images: S. Omelchenko



MNIST dataset, average in 10 randomly chosen images.



MNIST dataset, average in 5 randomly chosen and scaled images,

$n \in [28^2 = 784, 224^2 = 50176], \varepsilon = 0.1.$

Histograms

- Discrete-discrete case. Generate two random vectors $a, b \in S_n(1)$.
- Quasi Semi-discrete case. Generate $a \in S_n(1)$ and an empirical counterpart of 1D Gaussian distribution.
- Quasi Continuous case. Generate two empirical counterpart of 1D Gaussian distribution.

  Case 1. Equal variances, different expectation. Case 2. Different variances, different expectation.
- MNIST dataset.

  Case 1. Image of 1 and 3. Case 2. Image of 1 and 7.

Cost

- $C_{ij} = \|x_i - x_j\|_2$ (Wasserstein - 1)
- $C_{ij} = \|x_i - x_j\|_2^2$ (Wasserstein - 2)

- Standard LP solver `https://www.cvxpy.org/`. Interior point methods $O(n^3)$ complexity. Allows also to construct dual variables.
  - Apply directly to the linear program.
  - Apply to the entropy-regularized problem.
  - Apply to the squared-Euclidean norm regularized problem.
- Sinkhorn's method $K = \exp(-C/\gamma)$.

$$\tilde{u}^{(k+1)} = a \Big/ K\tilde{v}^{(k)} \qquad \tilde{v}^{(k+1)} = b \Big/ K^T \tilde{u}^{(k+1)} .$$

$\tilde{u}^{(k+1)} \exp(-C/\gamma)\tilde{v}^{(k+1)}$ – approximation for the transport plan

Exercise: How to find the approximation for the distance?

- Transport plan.
- Dual variables.
- Distance.

1 Introduction

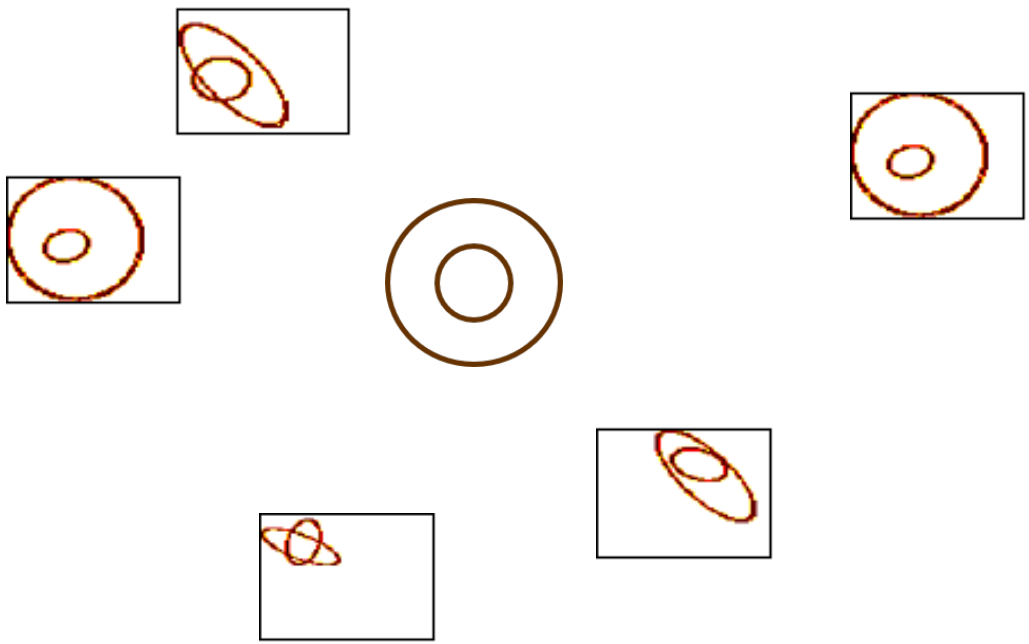
2 Application examples

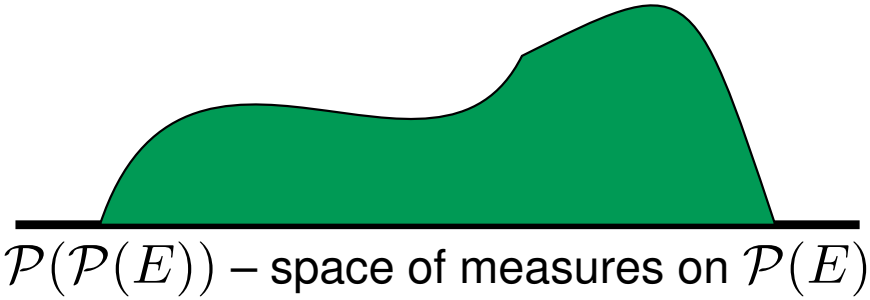
3 Numerical methods for OT distance

4 OT barycenters

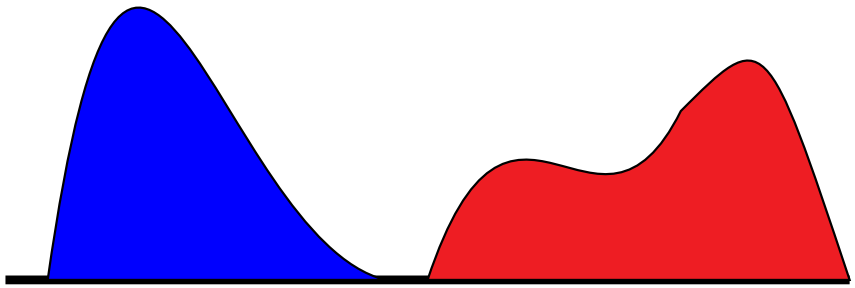
- Iterative Bregman Projections
- Accelerated gradient method
- Stochastic accelerated gradient method

Barycenter

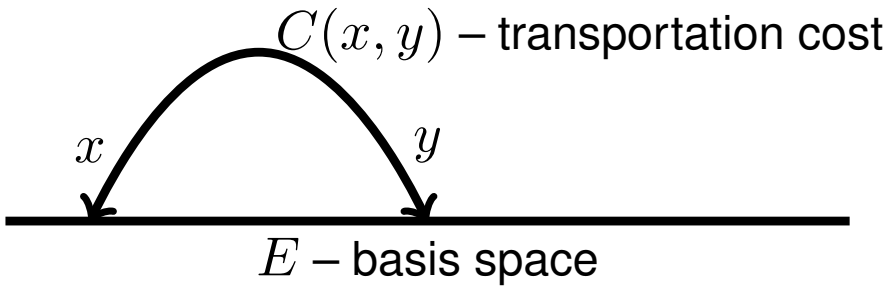




$\mathcal{P}(\mathcal{P}(E))$ – space of measures on $\mathcal{P}(E)$



$\mathcal{P}(E)$ – space of measures on E

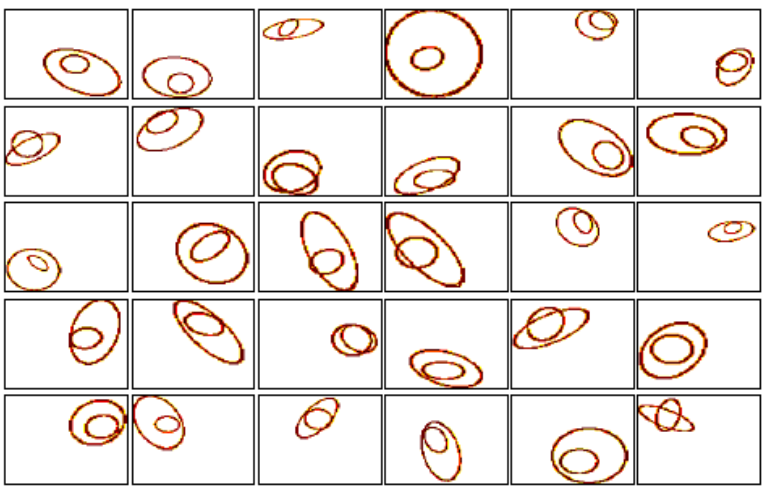


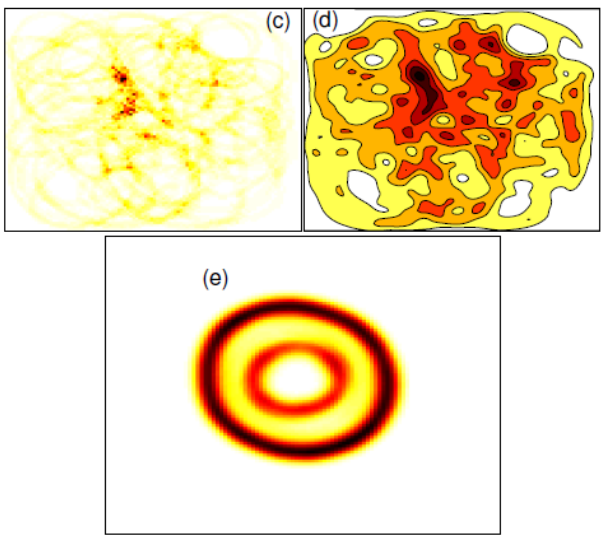
E – basis space

$$\hat{\nu} = \arg \min_{\nu \in \mathcal{P}_2(\Omega)} \frac{1}{m} \sum_{i=1}^{m} \mathcal{W}_p^p(\mu_i, \nu),$$

where $\mathcal{W}_p(\mu, \nu)$ is the Wasserstein distance between measures μ and ν on Ω.

WB is efficient in machine learning problems with geometric data, e.g. template image reconstruction from random sample:





Images from [Cuturi & Doucet, 2014]

Assume

- T – bounded random admissible transformation with finite moment $\mathbb{E}T$
- $T_i, i = 1, ..., m$ – random sample of realizations of T
- $\mu_i = (T_i)_{\#}\mu$ – random sample of measures, where μ is compactly supported.
- $\hat{\nu} = \arg\min_{\nu \in \mathcal{P}_2(\Omega)} \frac{1}{m} \sum_{i=1}^{m} \mathcal{W}_2^2(\mu_i, \nu)$

Then

- If $\mathbb{E}T = Id$, then $\mathcal{W}_2^2(\hat{\nu}, \mu) \to 0$ a.s. as $m \to \infty$.
- If $\|T - Id\|_{L^2} \leq M$ a.s., then

$$\mathbb{P}(\mathcal{W}_2(\hat{\nu}, \mu) \geq \varepsilon) \leq 2 \exp\left(-m \frac{\varepsilon^2}{M^2(1 + c\varepsilon/M)}\right).$$

Boissard, Le Gouic, Loubes. Distribution's template estimate with Wasserstein metrics. 2015

p-Kantorovich-Wasserstein distance $\mathcal{W}_p(a,b)$

$$(\mathcal{W}_p(a,b))^{1/p} = \min_{X \in \mathcal{U}(a,b)} \langle C, X \rangle,$$

$$\mathcal{U}(a,b) = \{X \in \mathbb{R}_+^{n \times n} : X\mathbf{1} = a,\ X^T\mathbf{1} = b\}.$$

Given a set of m measures $b_i \in S_n(1)$, $i = 1, ..., m$, their Wasserstein barycenter is

$$\hat{a} = \arg \min_{a \in S_n(1)} \frac{1}{m} \sum_{i=1}^{m} (\mathcal{W}_p(a, b_i))^p = \arg \min_{a \in S_n(1)} \frac{1}{m} \sum_{i=1}^{m} \min_{X_i \in \mathcal{U}(a,b_i)} \langle C, X_i \rangle$$

$$= \arg \min_{\substack{a \in S_n(1) \\ X_i \in \mathbb{R}_+^{n \times n} \\ X_i\mathbf{1} = a,\ X_i^T\mathbf{1} = b_i}} \frac{1}{m} \sum_{i=1}^{m} \langle C, X_i \rangle.$$

Large-scale linear program of dimension $mn^2 + n$.

For simplicity, we omit p below.

Main question: How much work is it needed to find their barycenter $\hat{a} \in S_n(1)$ with accuracy ε?

$$\frac{1}{m}\sum_{l=1}^{m}\mathcal{W}(\hat{a}, b_l) - \min_{a \in S_n(1)} \frac{1}{m}\sum_{l=1}^{m}\mathcal{W}(a, b_l) \leq \varepsilon$$

Challenges:

- Fine discrete approximation for ν and $\mu \Rightarrow$ large n,
- Large amount of data $\Rightarrow$ large m,
- Data produced and stored distributedly (e.g. produced by a network of sensors),
- Possibly continuous measures μ_i.

γ-regularized Monge-Kantorovich (Wasserstein) distance (Sinkhorn's distance) $\mathcal{W}_\gamma(a,b)$

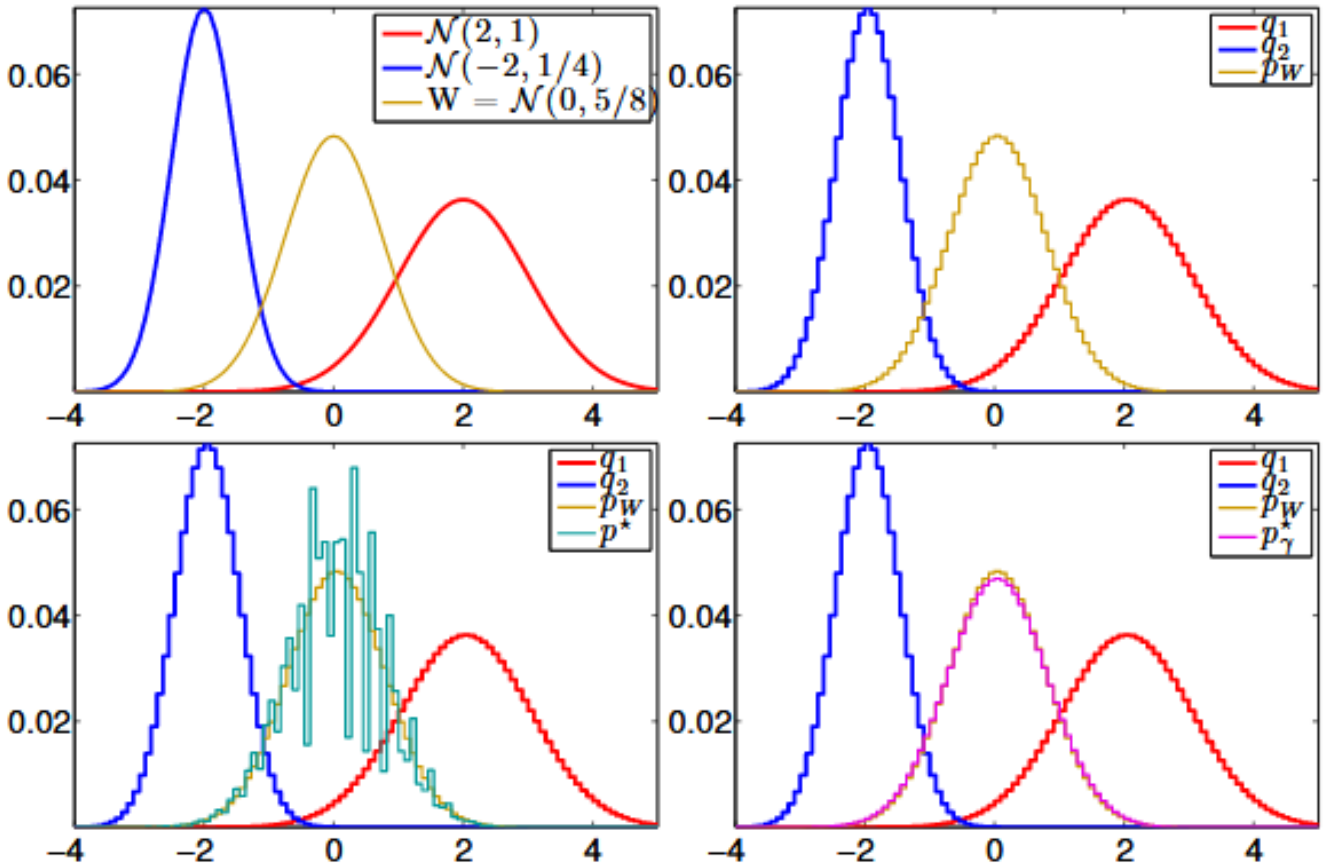
$$\mathcal{W}_\gamma(a,b) = \min_{X \in \mathcal{U}(a,b)} \langle C, X\rangle + \gamma\langle X, \ln X\rangle,$$

$$\mathcal{U}(a,b) = \{X \in \mathbb{R}_+^{n\times n} : X\mathbf{1} = a,\ X^T\mathbf{1} = b\}.$$

Given a set of m measures $b_i \in S_n(1)$, $i = 1, ..., m$, their regularized Wasserstein barycenter is

$$\hat{a}_\gamma = \arg\min_{a \in S_n(1)} \frac{1}{m}\sum_{i=1}^{m} \mathcal{W}_\gamma(a, b_i)$$

Cuturi, Doucet. Fast Computation of Wasserstein Barycenters. ICML 2014.



Cuturi, Peyré. A smoothed dual approach for variational Wasserstein problems, 2015

$$\hat{a}_\gamma = \arg \min_{a \in S_n(1)} \frac{1}{m} \sum_{i=1}^{m} \mathcal{W}_\gamma(a, b_i), \quad \gamma \geq 0.$$

Algorithms for barycenter

- Sinkhorn + Gradient Descent [Cuturi, Doucet, ICML'14]
- Iterative Bregman Projections [Benamou et al., SIAM J Sci Comp'15]
- (Accelerated) Gradient Descent [Cuturi, Peyre, SIAM J Im Sci'16; Dvurechensky et al, NeurIPS'18; Uribe et al., CDC'18].
- Stochastic Gradient Descent [Staib et al., NeurIPS'17; Claici, Chen, Solomon, ICML'18]

Question of complexity was open.

How to choose γ?

By the duality

$$\mathcal{W}_\gamma(a,b) = \max_{\xi,\eta} -\gamma \sum_{i,j=1}^{n} \exp\left(-\frac{1}{\gamma}(C_{ij} - \xi_i - \eta_j)\right) + \langle \xi, a\rangle + \langle \eta, b\rangle.$$

By the Demyanov-Danskin theorem

$$\nabla_a \mathcal{W}_\gamma(a,b) = \xi^*.$$

$$\nabla_a \sum_{i=1}^{m} \mathcal{W}_\gamma(a,b_i) = \sum_{i=1}^{m} \xi_i^*$$

Idea

1. Given current approximation $a_\gamma^{(k)}$, find $(\xi_i^*)^{(k)}$.
2. Make a gradient descent step

$$a_\gamma^{(k+1)} = a_\gamma^{(k)} - \frac{\alpha_k}{m} \nabla_a \sum_{i=1}^{m} \mathcal{W}_\gamma(a_\gamma^{(k)}, b_i) = a_\gamma^{(k)} - \frac{\alpha_k}{m} \nabla_a \sum_{i=1}^{m} (\xi_i^*)^{(k)}$$

Cuturi, Doucet. Fast Computation of Wasserstein Barycenters. ICML 2014.

$$\min_{a \in S_n(1)} \frac{1}{m} \sum_{i=1}^{m} \mathcal{W}_\gamma(a, b_i) = \min_{a \in S_n(1)} \frac{1}{m} \sum_{i=1}^{m} \min_{X_i \in \mathcal{U}(a, b_i)} \left\{ \langle C, X_i \rangle + \gamma \langle X_i, \ln X_i \rangle \right\}.$$

$$X_i \in \mathcal{U}(a, b_i), i = 1, \ldots, m \Longleftrightarrow X_i \in \mathbb{R}_+^{n \times n};\ X_i^T \mathbf{1} = b_i,\ X_i \mathbf{1} = X_{i+1} \mathbf{1},\ i = 1, \ldots$$

$$\min_{\substack{X_i \in \mathbb{R}_+^{n \times n} \\ X_i^T \mathbf{1} = b_i,\ X_i \mathbf{1} = X_{i+1} \mathbf{1}, \\ i = 1, \ldots, m}} \frac{1}{m} \sum_{i=1}^{m} \left\{ \langle X_i, C \rangle + \gamma \langle X_i, \ln X_i \rangle \right\} = \min_{\boldsymbol{X} \in \mathcal{C}_1 \cap \mathcal{C}_2} \frac{1}{m} \sum_{i=1}^{m} KL\left(X_i, \mathrm{e}^{-\frac{C}{\gamma}} \right),$$

where

$$\mathcal{C}_1 = \left\{ \boldsymbol{X} = [X_1, \ldots, X_m] : \forall i\ X_i^T \mathbf{1} = b_i \right\},$$

$$\mathcal{C}_2 = \left\{ \boldsymbol{X} = [X_1, \ldots, X_m] : \exists a \in S_n(1) \quad \forall i \quad X_i \mathbf{1} = a \right\}.$$

Benamou, Carlier, Cuturi, Peyré. Iterative Bregman Projections for Regularized Transportation Problems, 2015

$$\min_{\boldsymbol{X} \in \mathcal{C}_1 \cap \mathcal{C}_2} \frac{1}{m} \sum_{i=1}^{m} KL\left(X_i, \mathrm{e}^{-\frac{C}{\gamma}}\right),$$

where

$$\mathcal{C}_1 = \left\{\boldsymbol{X} = [X_1, \ldots, X_m] : \forall i \; X_i^T \mathbf{1} = b_i\right\},$$
$$\mathcal{C}_2 = \left\{\boldsymbol{X} = [X_1, \ldots, X_m] : \exists a \in S_n(1) \quad \forall i \quad X_i \mathbf{1} = a\right\}.$$

Alternating minimization $\boldsymbol{X}^0 = [\mathrm{e}^{-\frac{C}{\gamma}}, ..., \mathrm{e}^{-\frac{C}{\gamma}}]$

$$\boldsymbol{X}^{(k+1)} = \arg\min_{\boldsymbol{X} \in \mathcal{C}_1} \sum_{i=1}^{m} KL\left(X_i, X_i^{(k)}\right), \; \boldsymbol{X}^{(k+2)} = \arg\min_{\boldsymbol{X} \in \mathcal{C}_2} \sum_{i=1}^{m} KL\left(X_i, X_i^{(k+1)}\right).$$

Benamou, Carlier, Cuturi, Peyré. Iterative Bregman Projections for Regularized Transportation Problems, 2015

$$\min_{\substack{X_i \in \mathbb{R}_+^{n\times n} \\ X_i^T \mathbf{1} = b_i,\ X_i \mathbf{1} = X_{i+1}\mathbf{1}, i=1,\dots,m}} \frac{1}{m}\sum_{i=1}^{m}\{\langle X_i, C\rangle + \gamma\langle X_i, \ln X_i\rangle\}$$

Dual problem:

$$\min_{\substack{\mathtt{u},\mathtt{v} \\ \frac{1}{m}\sum_{l=1}^{m} v_l = 0}} f(\mathtt{u},\mathtt{v}) := \frac{1}{m}\sum_{l=1}^{m}\Big\{\langle \mathbf{1}, B_l(u_l, v_l)\mathbf{1}\rangle - \langle u_l, b_l\rangle\Big\},$$

$\mathtt{u} = [u_1, \dots, u_m]$, $\mathtt{v} = [v_1, \dots, v_m]$, $u_l, v_l \in \mathbb{R}^n$,

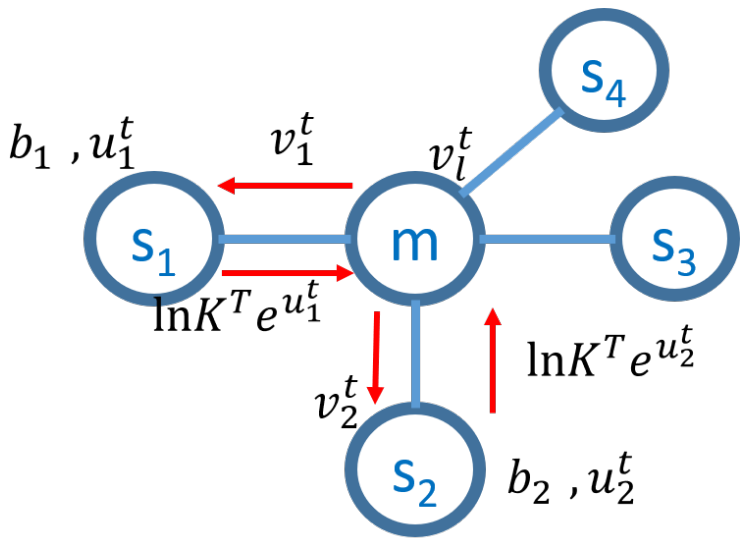
$B_l(u_l, v_l) := \operatorname{diag}(e^{u_l}) \exp(-C/\gamma) \operatorname{diag}(e^{v_l})$, $K = \exp(-C/\gamma)$.

IBP is equivalent to alternating minimization for the dual problem.

- $u_l^{t+1} := \ln b_l - \ln K e^{v_l^t}$, $\mathtt{v}^{t+1} := \mathtt{v}^t$
- $v_l^{t+1} := \frac{1}{m}\sum_{k=1}^{m} \ln K^T e^{u_k^t} - \ln K^T e^{u_l^t}$, $\mathtt{u}^{t+1} := \mathtt{u}^t$
- $\hat{a} := \frac{1}{\sum_{l=1}^{m}\langle \mathbf{1}, B_l(u_l, v_l)\mathbf{1}\rangle} \sum_{l=1}^{m} B_l(u_l, v_l)\mathbf{1}$

Kroshnin, Tupitsa, Dvinskikh, D., Gasnikov, Uribe. On the complexity of approximating Wasserstein barycenters, ICML 2019.

- To find an ε approximation of the γ-regularized WB, Iterative Bregman Projections (IBP) needs $\frac{1}{\gamma\varepsilon}$ iterations. (cf. $\frac{1}{\gamma\varepsilon}$ for the Sinkhorn's algorithm)
- Setting $\gamma = \Theta\left(\varepsilon/\ln n\right)$ allows to find an ε-approximation for the *non-regularized* WB with arithmetic operations complexity $\frac{mn^2}{\varepsilon^2}$. (cf. $\frac{n^2}{\varepsilon^2}$ for the Sinkhorn's algorithm).
- IBP can be implemented distributedly in a centralized architecture (master/slaves).

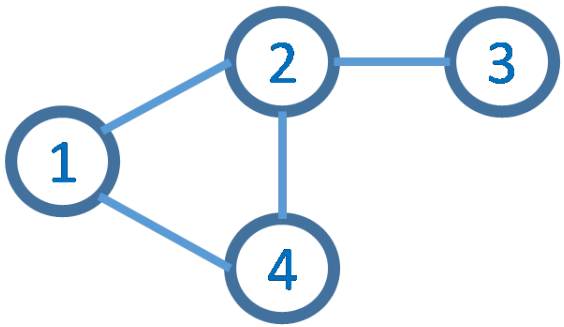


Kroshnin, Tupitsa, Dvinskikh, D., Gasnikov, Uribe. On the complexity of approximating Wasserstein barycenters, ICML 2019.

$$\min_{a\in S_n(1)} \frac{1}{m}\sum_{i=1}^{m} \mathcal{W}_\gamma(a, b_i) = \min_{\substack{a_1=\ldots=a_m \\ a_1,\ldots,a_m\in S_n(1)}} \frac{1}{m}\sum_{i=1}^{m} \mathcal{W}_\gamma(a_i, b_i)$$

Define symmetric p.s.d. matrix $\bar{L}$ s.t. $\mathrm{Ker}(\bar{\mathrm{L}}) = \mathrm{span}(\mathbf{1})$.

Example: graph Laplace matrix



$$L = \begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 1 & 0 \\ -1 & -1 & 0 & 2 \end{pmatrix}$$

Let $\mathsf{a} = [a_1, ..., a_m]$, $L = \bar{L} \otimes I_n$. Then $a_1 = ... = a_m \iff L\mathsf{a} = 0$.

Equivalent form of the WB problem

$$\max_{a_1,\ldots,a_m\in S_n(1)} -\frac{1}{m}\sum_{i=1}^{m} \mathcal{W}_\gamma(a_i, b_i) \quad \text{s.t.} \quad L\mathsf{a} = 0.$$

Uribe, Dvinskikh, D., Gasnikov, Nedić. Distributed Computation of Wasserstein Barycenters over Networks, CDC 2018; D., Dvinskikh, Gasnikov, Uribe, Nedić Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters, NeurIPS 2018

$$\max_{a_1,\dots,a_m \in S_n(1)} \left\{ -\frac{1}{m} \sum_{i=1}^{m} \mathcal{W}_\gamma(a_i, b_i) + \min_{\boldsymbol{\lambda} \in \mathbb{R}^{mn}} \sum_{i=1}^{m} \langle \lambda_i, [L\mathsf{a}]_i \rangle \right\}$$

$$= \min_{\boldsymbol{\lambda} \in \mathbb{R}^{mn}} \max_{a_1,\dots,a_m \in S_n(1)} \sum_{i=1}^{m} \left(\langle a_i, [L\boldsymbol{\lambda}]_i \rangle - \frac{1}{m} \mathcal{W}_\gamma(a_i, b_i) \right)$$

Dual problem

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^{mn}} \Phi(\boldsymbol{\lambda}) := \frac{1}{m} \sum_{i=1}^{m} \mathcal{W}^*_{\gamma, b_i} \left(m \, [L\boldsymbol{\lambda}]_i \right),$$

where $\mathcal{W}^*_{\gamma,b}$ is the Fenchel-Legandre conjugate for $\mathcal{W}_\gamma(\cdot, b)$

$$\mathcal{W}^*_{\gamma,b}(s) := \max_{a \in S_n(1)} \{ \langle a, s \rangle - \mathcal{W}_\gamma(a, b) \}.$$

$$a(s) = \arg \max_{a \in S_n(1)} \{ \langle a, s \rangle - \mathcal{W}_\gamma(a, b) \}.$$

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^{mn}} \Phi(\boldsymbol{\lambda}) := \frac{1}{m}\sum_{i=1}^{m} \mathcal{W}_\gamma^*\left(m\left[L\boldsymbol{\lambda}\right]_i\right)$$

By the chain rule

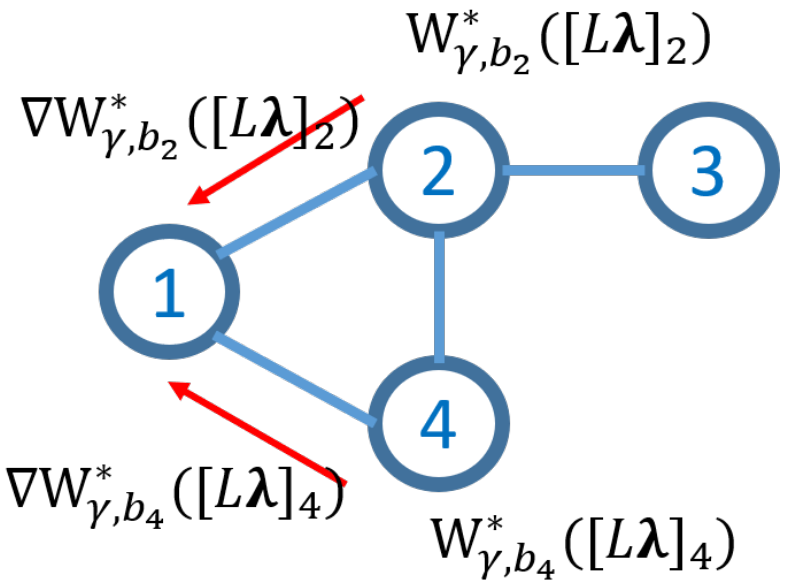
$$\left[\nabla\Phi(\boldsymbol{\lambda})\right]_i = \sum_{j=1}^{m} \left[L\right]_{ij} \nabla\mathcal{W}_{\gamma,b_j}^*\left(\left[L\boldsymbol{\lambda}\right]_j\right)$$

Gradient descent step

$$\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)} - \alpha\nabla\Phi(\boldsymbol{\lambda}^{(k)})$$

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} - \alpha\sum_{j=1}^{m} \left[L\right]_{ij} \nabla\mathcal{W}_{\gamma,b_j}^*\left(\left[L\boldsymbol{\lambda}^{(k)}\right]_j\right)$$

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} - \alpha \sum_{j=1}^{m} [L]_{ij} \, \nabla \mathcal{W}^*_{\gamma,b_j} \left(\left[L\boldsymbol{\lambda}^{(k)} \right]_j \right).$$



$$L = \begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 1 & 0 \\ -1 & -1 & 0 & 2 \end{pmatrix}$$

Boyd, Parikh, Chu, Peleato, and Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. 2011.

Jakovetić, Moura, Xavier. Linear convergence rate of a class of distributed augmented Lagrangian algorithms. 2015.

We run (accelerated) gradient descent for the dual

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} - \alpha \sum_{j=1}^{m} [L]_{ij} \, \nabla \mathcal{W}^*_{\gamma, b_j} \left(\left[L\boldsymbol{\lambda}^{(k)} \right]_j \right).$$

and average the obtained primal information,

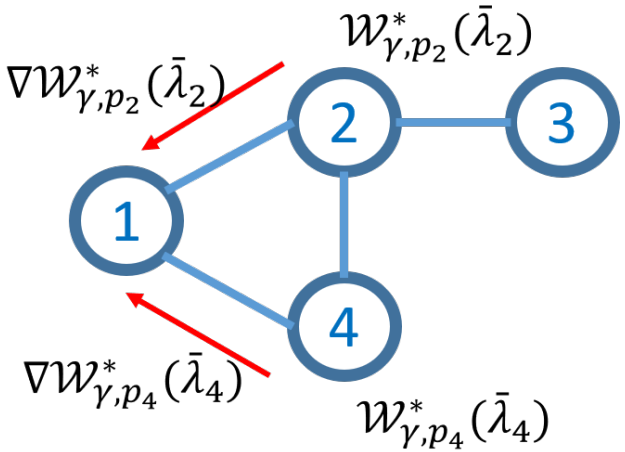
$a(s) = \arg\max_{a \in S_n(1)} \{\langle a, s \rangle - \mathcal{W}_\gamma(a, b)\}.$

$$\hat{a}_i^{(k)} = \frac{1}{k+1} \sum_{\ell=0}^{k} a([L\boldsymbol{\lambda}^{(\ell)}]_i).$$

Convergence rate [Kroshnin, Dvinskikh, D., Gasnikov, Tupitsa, Uribe, 2019]

$$\frac{1}{m} \sum_{i=1}^{m} \mathcal{W}_\gamma(\hat{a}_i^{(k)}, b_i) - \frac{1}{m} \sum_{i=1}^{m} \mathcal{W}_\gamma(a_i^*, b_i) = O\left(\frac{\sqrt{n}}{\gamma k^2}\right), \quad \|L\hat{\mathtt{a}}^{(k)}\|_2 = O\left(\frac{\sqrt{n}}{\gamma k^2}\right)$$

- To find an ε approximation of the γ-regularized WB, accelerated gradient method (AGM) needs $\sqrt{\frac{n}{\gamma\varepsilon}}$ iterations. (cf. with $\sqrt{\frac{n}{\gamma\varepsilon}}$ iterations for OT distance.)
- Setting $\gamma = \Theta\left(\varepsilon/\ln n\right)$ allows to find an ε-approximation for the *non-regularized* WB with arithmetic operations complexity $\frac{mn^{2.5}}{\varepsilon}$. (cf. with $\frac{n^{2.5}}{\varepsilon}$ a.o. for OT distance.)
- AGM can be implemented distributedly in a *decentralized* architecture.



Kroshnin, Dvinskikh, D., Gasnikov, Tupitsa, Uribe. On the Complexity of Approximating Wasserstein Barycenter, ICML 2019

- Iterative Bregman Projections algorithms with $\gamma = \frac{\varepsilon}{4 \ln n}$ requires

$$O\left(\frac{mn^2}{\varepsilon^2}\right) \text{ a.o.}$$

to obtain $\hat{a}$ s.t.

$$\frac{1}{m}\sum_{i=1}^{m} \mathcal{W}(\hat{a}, b_i) - \frac{1}{m}\sum_{i=1}^{m} \mathcal{W}(a^*, b_i) \leq \varepsilon$$

- Accelerated gradient descent with $\gamma = \frac{\varepsilon}{4 \ln n}$ requires

$$O\left(\frac{mn^{2.5}}{\varepsilon}\right) \text{ a.o.}$$

to obtain $\hat{a}_i$, $i = 1, ..., m$ s.t.

$$\frac{1}{m}\sum_{i=1}^{m} \mathcal{W}(\hat{a}_i, b_i) - \frac{1}{m}\sum_{i=1}^{m} \mathcal{W}(a^*, b_i) \leq \varepsilon, \quad \|L\hat{\mathsf{a}}\|_2 \leq \varepsilon.$$

Kroshnin, Dvinskikh, D., Gasnikov, Tupitsa, Uribe. On the Complexity of Approximating Wasserstein Barycenter, ICML 2019

Given: A) Probability measure μ with density $q(y)$ on $\mathcal{Y}$.

B) Discrete probability measure $\nu = \sum_{i=1}^{n} p_i \delta(z_i)$ with finite support given by points $z_1, \dots, z_n \in \mathcal{Z}$.

Regularized Wasserstein distance in semi-discrete setting:

$$\mathcal{W}_\gamma(\mu, \nu) = \min_{\pi \in \mathcal{U}(\mu,\nu)} \left\{ \sum_{i=1}^{n} \int_{\mathcal{Y}} c_i(y) \pi_i(y) dy + \gamma KL(\pi|\xi) \right\},$$

$c_i(y) = c(z_i, y)$ – cost function, $KL(\pi|\xi) = \sum_{i=1}^{n} \int_{\mathcal{Y}} \pi_i(y) \log\left(\frac{\pi_i(y)}{\xi}\right) dy$, ξ – uniform distribution on $\mathcal{Y} \times \mathcal{Z}$, the set of admissible coupling measures π is defined as follows

$$\mathcal{U}(\mu, \nu) = \left\{ \pi : \sum_{i=1}^{n} \pi_i(y) = q(y), y \in \mathcal{Y}, \int_{\mathcal{Y}} \pi_i(y) dy = p_i, \forall\, i = 1, \dots, n \right\}.$$

$$\hat{a}_\gamma = \arg \min_{a \in S_1(n)} \sum_{i=1}^{m} \mathcal{W}_{\gamma,\mu_i}(a),$$

where we fixed the support $z_1, \ldots, z_n \in \mathcal{Z}$ of the barycenter ν and characterize it by the vector $a \in S_n(1)$, i.e., $\nu = \sum_{i=1}^{n} [a]_i \delta(z_i)$ and $\mathcal{W}_{\gamma,\mu}(p) := \mathcal{W}_\gamma(\mu, \nu)$.

Equivalent form

$$\min_{\substack{a_1 = \ldots = a_m \\ a_1, \ldots, a_m \in S_1(n)}} \sum_{i=1}^{m} \mathcal{W}_{\gamma,\mu_i}(a_i).$$

Primal-Dual pair of problems

$$\max_{a_1,\dots,a_m \in S_1(n), L\mathtt{a}=0} -\sum_{i=1}^{m} \mathcal{W}_{\gamma,\mu_i}(a_i), \quad \min_{\boldsymbol{\lambda}\in\mathbb{R}^{mn}} \sum_{i=1}^{m} \mathcal{W}^*_{\gamma,\mu_i}([L\boldsymbol{\lambda}]_i),$$

where $\mathtt{a} = [a_1^T, \cdots, a_m^T]^T \in \mathbb{R}^{mn}$, $\boldsymbol{\lambda} = [\lambda_1^T, \cdots, \lambda_m^T]^T \in \mathbb{R}^{mn}$.

Lemma

Given a measure μ with density $q(y)$ on a metric space $\mathcal{Y}$, the Fenchel-Legendre dual function for $\mathcal{W}_{\gamma,\mu}(p)$ is

$$\begin{aligned}\mathcal{W}^*_{\gamma,\mu}(\bar{\lambda}) &= \int \gamma \log\left(\frac{1}{q(y)}\sum_{l=1}^{n}\exp\left(\frac{[\bar{\lambda}]_l - c_l(y)}{\gamma}\right)\right) q(y)dy \\ &= \mathbb{E}_{Y\sim\mu}\gamma\log\left(\frac{1}{q(Y)}\sum_{l=1}^{n}\exp\left(\frac{[\bar{\lambda}]_l - c_l(Y)}{\gamma}\right)\right),\end{aligned}$$

and its gradient is $1/\gamma$ Lipschitz-continuous w.r.t. 2-norm.

D., Dvinskikh, Gasnikov, Uribe, Nedić Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters, 2018

Lemma [D., Dvinskikh, Gasnikov, Uribe, Nedić, 2018]

$\nabla \mathcal{W}^*_\gamma(\boldsymbol{\lambda})$ is $\lambda_{\max}(L)/\gamma$-Lipschitz-continuous w.r.t. 2-norm.

If its stochastic approximation is defined as

$$[\widetilde{\nabla} \mathcal{W}^*_\gamma(\boldsymbol{\lambda})]_i = \sum_{j=1}^m L_{ij} \widetilde{\nabla} \mathcal{W}^*_{\gamma,\mu_j}(\bar{\lambda}_j),\ i = 1, ..., m, \text{ with}$$

$$\widetilde{\nabla} \mathcal{W}^*_{\gamma,\mu_j}(\bar{\lambda}_j) = \frac{1}{M} \sum_{r=1}^M p_j(\bar{\lambda}_j),\ j = 1, ..., m$$

$$[p_j(\bar{\lambda}_j)]_l = \frac{\exp(([\bar{\lambda}_j]_l - c_l(Y_r^j))/\gamma)}{\sum_{\ell=1}^n \exp(([\bar{\lambda}_j]_\ell - c_\ell(Y_r^j))/\gamma)},\ j = 1, ..., m,\ l = 1, ..., n,$$

M – batch size, $\bar{\lambda}_j := [L\boldsymbol{\lambda}]_j$, $Y_1^j, ..., Y_r^j$ is a sample from the measure μ_j.

Then $\mathbb{E}_{Y_r^j \sim \mu_j, j=1,...,m, r=1,...,M} \widetilde{\nabla} \mathcal{W}^*_\gamma(\boldsymbol{\lambda}) = \nabla \mathcal{W}^*_\gamma(\boldsymbol{\lambda})$ and

$$\mathbb{E}_{Y_r^j \sim \mu_j, j=1,...,m, r=1,...,M} \|\widetilde{\nabla} \mathcal{W}^*_\gamma(\boldsymbol{\lambda}) - \nabla \mathcal{W}^*_\gamma(\boldsymbol{\lambda})\|_2^2 \leq \lambda_{\max}(L)/M,\ \boldsymbol{\lambda} \in \mathbb{R}^{mn}.$$

Primal Problem

$$\min_{x \in Q \subseteq E} \{f(x) : Ax = c\}.$$

Dual problem

$$\min_{\lambda \in H^*} \left\{ \varphi(\lambda) := \langle \lambda, c \rangle + \max_{x \in Q} \{-f(x) - \langle \lambda, Ax \rangle\} \right\}.$$

$$\nabla \varphi(\lambda) = c - Ax(\lambda), \quad x(\lambda) := \arg\max_{x \in Q} \{-f(x) - \langle \lambda, Ax \rangle\}.$$

NB: $\nabla \varphi(\lambda)$ is Lipschitz-continuous

$$\varphi(\lambda) \leq \varphi(\zeta) + \langle \nabla \varphi(\zeta), \lambda - \zeta \rangle + \frac{\kappa}{2} \|\lambda - \zeta\|_{H,*}^2.$$

$$\mathbb{E}_\xi \nabla \Phi(\lambda, \xi) = \nabla \varphi(\lambda), \quad \mathbb{E}_\xi \|\nabla \Phi(\lambda, \xi) - \nabla \varphi(\lambda)\|_2^2 \leq \sigma^2.$$

$$\varphi(\lambda) \leq \varphi(\zeta) + \langle \nabla\varphi(\zeta), \lambda - \zeta \rangle + \frac{\kappa}{2}\|\lambda - \zeta\|^2_{H,*}.$$

$$\eta_{k+1} = \arg\min_{\lambda} \varphi(\zeta_k) + \langle \nabla\varphi(\zeta_k), \lambda - \zeta_k \rangle + \frac{\kappa}{2}\|\lambda - \zeta_k\|^2_{H,*}.$$

Batch of size r

$$\nabla^r \Phi(\lambda, \{\xi\}^r) = \frac{1}{r} \sum_{i=1}^{r} \nabla \Phi(\lambda, \xi_i).$$

$$\mathbb{E}_{\{\xi\}^r} \nabla^r \Phi(\lambda, \{\xi\}^r) = \nabla \varphi(\lambda), \quad \mathbb{E}_{\{\xi\}^r} \|\nabla^r \Phi(\lambda, \{\xi\}^r) - \nabla \varphi(\lambda)\|_2^2 \le \frac{\sigma^2}{r}.$$

$$\begin{aligned}
\varphi(\lambda) &\le \varphi(\zeta) + \langle \nabla \varphi(\zeta), \lambda - \zeta \rangle + \frac{\kappa}{2} \|\lambda - \zeta\|_{H,*}^2 \\
&= \varphi(\zeta) + \langle \nabla^r \Phi(\zeta, \{\xi\}^r), \lambda - \zeta \rangle + \langle \nabla \varphi(\zeta) - \nabla^r \Phi(\zeta, \{\xi\}^r), \lambda - \zeta \rangle + \frac{\kappa}{2} \|\lambda - \zeta\|_{H,*}^2 \\
&\le \varphi(\zeta) + \langle \nabla^r \Phi(\zeta, \{\xi\}^r), \lambda - \zeta \rangle + \|\nabla \varphi(\zeta) - \nabla^r \Phi(\zeta, \{\xi\}^r)\|_H \|\lambda - \zeta\|_{H,*} \\
&\quad + \frac{\kappa}{2} \|\lambda - \zeta\|_{H,*}^2 \\
&\le \varphi(\zeta) + \langle \nabla^r \Phi(\zeta, \{\xi\}^r), \lambda - \zeta \rangle + \frac{1}{2\kappa} \|\nabla \varphi(\zeta) - \nabla^r \Phi(\zeta, \{\xi\}^r)\|_H^2 \\
&\quad + \frac{\kappa}{2} \|\lambda - \zeta\|_{H,*} + \frac{\kappa}{2} \|\lambda - \zeta\|_{H,*}^2 \\
&\le \varphi(\zeta) + \langle \nabla^r \Phi(\zeta, \{\xi\}^r), \lambda - \zeta \rangle + \frac{2\kappa}{2} \|\lambda - \zeta\|_{H,*}^2 + \frac{1}{2\kappa} \|\nabla \varphi(\zeta) - \nabla^r \Phi(\zeta, \{\xi\}^r)\|_H^2.
\end{aligned}$$

$$\eta_{k+1} = \arg\min_{\lambda} \varphi(\zeta_k) + \langle \nabla^r \Phi(\zeta_k, \{\xi\}^r), \lambda - \zeta_k \rangle + \frac{\kappa}{2}\|\lambda - \zeta_k\|_{H,*}^2.$$

$$\mathbb{E}\varphi(\eta_{k+1}) \leq \varphi(\zeta_k) + \langle \nabla\varphi(\zeta_k), \eta_{k+1} - \zeta_k \rangle + \frac{2\kappa}{2}\|\eta_{k+1} - \zeta_k\|_{H,*}^2 + \frac{\sigma^2}{\kappa r}.$$

Require: Each agent $i \in V$ is assigned its measure μ_i.

All agents set $[\bar{\boldsymbol{\eta}}_0]_i = [\bar{\boldsymbol{\zeta}}_0]_i = [\bar{\boldsymbol{\lambda}}_0]_i = \mathbf{0} \in \mathbb{R}^n$, $C_0 = \alpha_0 = 0$ and N

For each agent $i \in V$:

for $k = 0, \dots, N-1$ **do**

  Find α_{k+1} as the largest root of the equation

  $C_{k+1} := C_k + \alpha_{k+1} = 2L\alpha_{k+1}^2,$

  $\tau_{k+1} = \alpha_{k+1}/C_{k+1}.$

  Set $M_{k+1} = \max\{1, \lceil \gamma C_{k+1}/(\alpha_{k+1}\varepsilon)\rceil\}$

  $[\bar{\boldsymbol{\lambda}}_{k+1}]_i = \tau_{k+1}[\bar{\boldsymbol{\zeta}}_k]_i + (1-\tau_{k+1})[\bar{\boldsymbol{\eta}}_k]_i$

  Generate M_{k+1} samples $\{Y_r^i\}_{r=1}^{M_{k+1}}$ from the measure μ_i and set $\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_i}([\bar{\boldsymbol{\lambda}}_{k+1}]_i)$.

  Share $\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_i}([\bar{\boldsymbol{\lambda}}_{k+1}]_i)$ with $\{j \mid (i,j) \in E\}$

  $[\bar{\boldsymbol{\zeta}}_{k+1}]_i = [\bar{\boldsymbol{\zeta}}_k]_i - \alpha_{k+1}\sum_{j=1}^m L_{ij}\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_j}([\bar{\boldsymbol{\lambda}}_{k+1}]_j)$

  $[\bar{\boldsymbol{\eta}}_{k+1}]_i = \tau_{k+1}[\bar{\boldsymbol{\zeta}}_{k+1}]_i + (1-\tau_{k+1})[\bar{\boldsymbol{\eta}}_{k+1}]_i$

  Set $[\hat{\mathrm{a}}_{k+1}]_i = \tau_{k+1}a_i([\bar{\boldsymbol{\lambda}}_{k+1}]_i) + (1-\tau_{k+1})[\hat{\mathrm{a}}_k]_i.$

end for

Ensure: $\hat{\mathrm{a}}_N$.

Theorem [D., Dvinskikh, Gasnikov, Uribe, Nedić, 2018]

Algorithm after $N = \sqrt{16\lambda_{max}(L)R^2/(\varepsilon\gamma)}$ iterations returns an approximation $\hat{\mathsf{a}}_N$ for the barycenter, which satisfies

$$\sum_{i=1}^{m} \mathcal{W}_{\gamma,\mu_i}(\mathbb{E}[\hat{\mathsf{a}}_N]_i) - \sum_{i=1}^{m} \mathcal{W}_{\gamma,\mu_i}([\mathsf{a}^*]_i) \leq \varepsilon, \quad \|L\mathbb{E}\hat{\mathsf{a}}_N\|_2 \leq \varepsilon/R.$$

Moreover, the total complexity in arithmetic operations is

$$O\left(n \cdot \max\left\{\frac{\lambda_{\max}(L)R^2}{\varepsilon^2}, \sqrt{\frac{\lambda_{\max}(L)R^2}{\varepsilon\gamma}}\right\}\right).$$

If all μ_i's are discrete, one can apply a deterministic algorithm with complexity

$$O\left(n^2\sqrt{\frac{\lambda_{\max}(L)R^2}{\varepsilon\gamma}}\right).$$

If $n > \frac{1}{\varepsilon}$, our new approach is better.

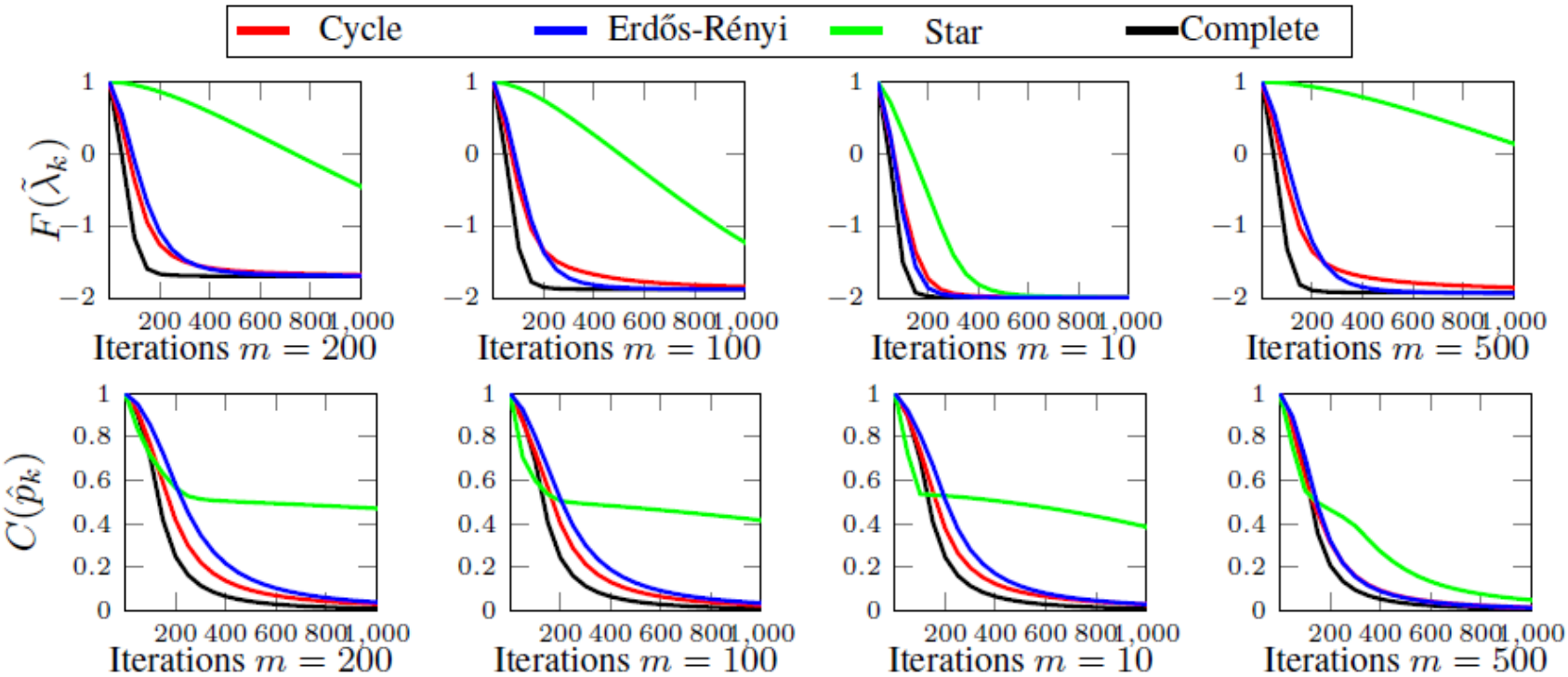
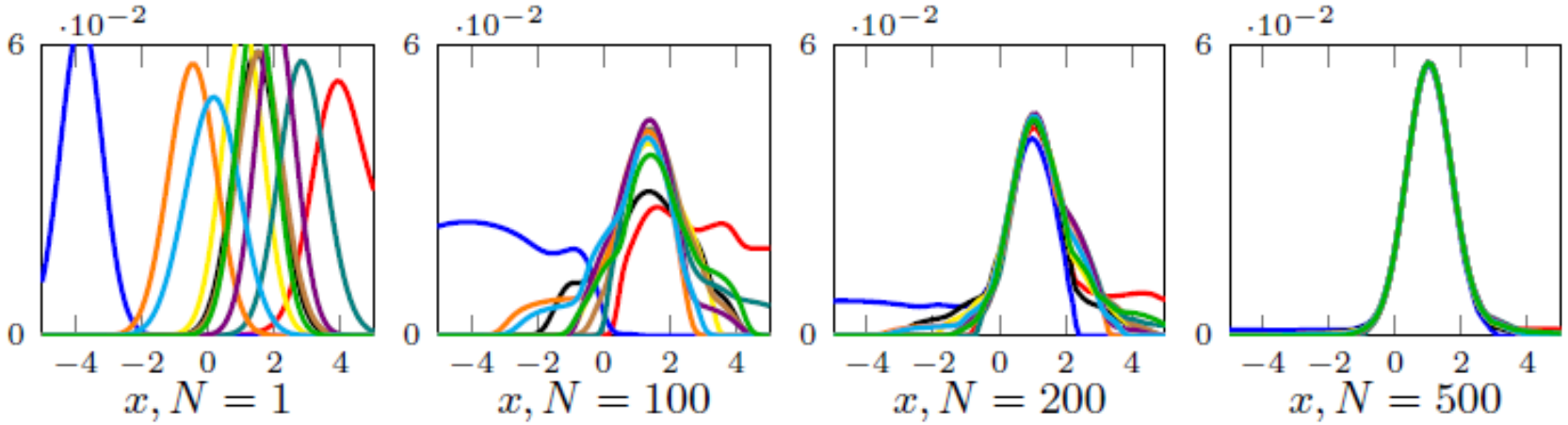


Figure: Dual function value and distance to consensus for graphs with $200, 100, 10, 500$ agents, $M_k = 100$ and $\gamma = 0.1$.



(a) Local Gaussian Distributions

Figure: Local barycenter generated by the Algorithm for a set of 10 agents over an Erdös-Rényi random graph at different iteration numbers. Each agent can access private realizations from a Gaussian random variable.

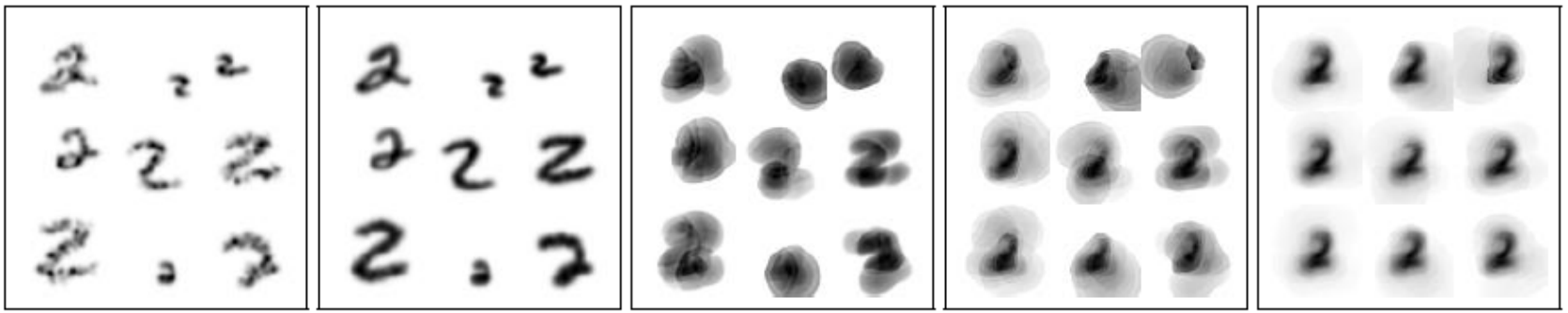
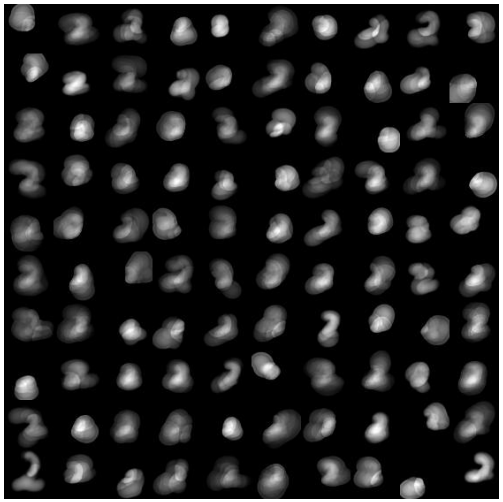


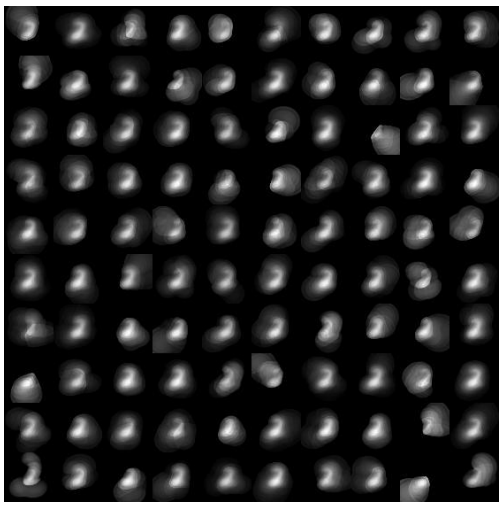
Figure: Wasserstein Barycenter of a subset of images of the digit 2 from the MNIST dataset. Each block shows a subset of 9 randomly selected local barycenters, generated by our Algorithm at different time instances. The 9 agents are selected from a network of 500 agents on an Erdös-Rényi random graph.

Experiments on MNIST dataset

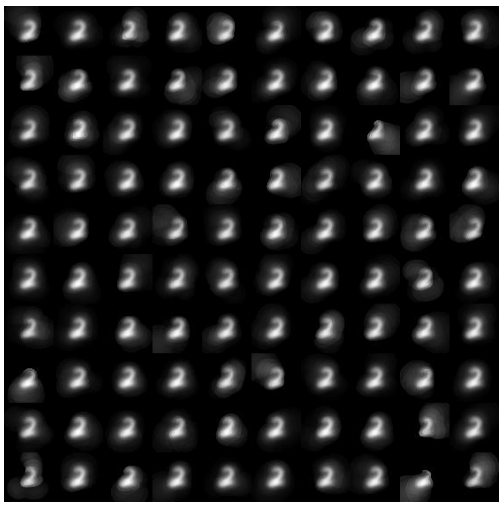
$k = 0$



$k = 10$



$k = 20$



$k = 30$

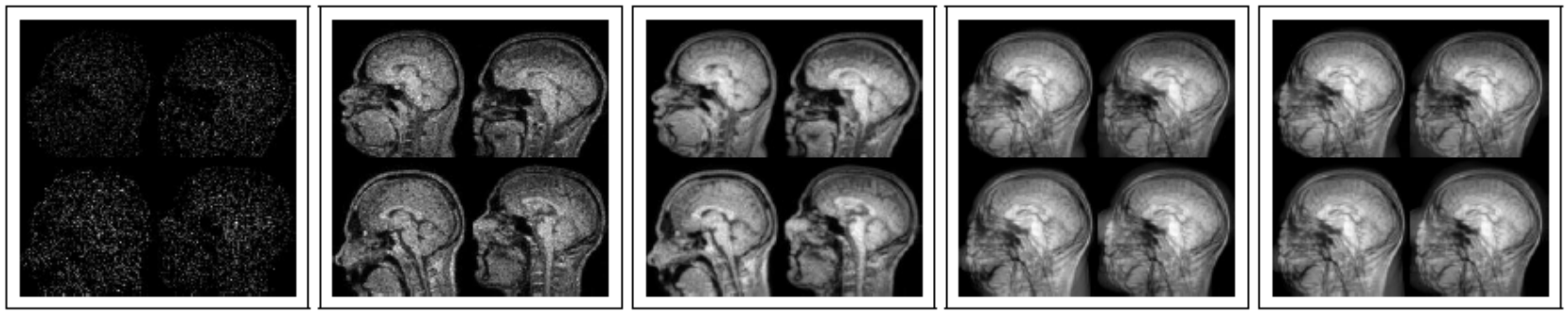


Figure: Wasserstein Barycenter for a subset of images from the IXI dataset. Each block shows the local barycenters, of 4 agents, generated by our Algorithm at different time instances. The 4 agents are connected on a cycle graph.

Optimal Transport is an interesting topic which connects many fields and has interesting applications

- Applications in image analysis
- Interplay between geometry, probability and PDEs
- Convexity, duality
- Numerical optimization
- Statistics and Machine Learning

Thank you!