Wasserstein barycenters from the computational perspective

Pavel Dvurechensky,

joint work with D. Dvinskikh (WIAS), A. Gasnikov (MIPT), S. Guminov (MIPT), A. Kroshnin (HSE), A. Nedic (ASU), N. Tupitsa (IITP RAS), C. Uribe (MIT)

Moscow Conference on Combinatorics and Applications, 02.06.2021

Outline

1. Background on Optimal Transport
2. Motivation for Wasserstein barycenters
3. Numerical methods for Wasserstein barycenters
    3.1 Iterative Bregman Projections (IBP)
    3.2 Accelerated gradient method
    3.3 Accelerated Alternating Minimization

Monge Problem

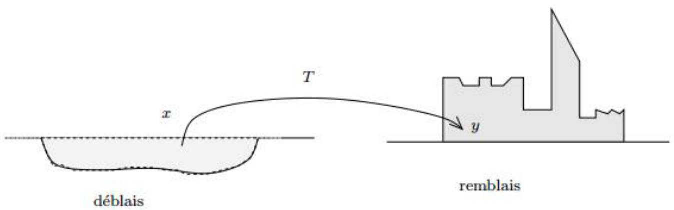
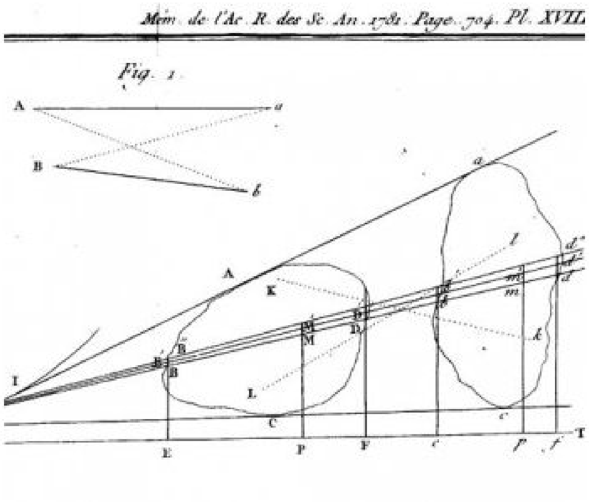


Fig. 3.1. Monge's problem of déblais and remblais



- (E, D) – metric space;
- $\mu, \nu \in \mathcal{P}_2$ – measures to be transported;
- transport map $T : E \to E$, s.t. $\forall B$, $\mu(T^{-1}(B)) = \nu(B)$.

$$\inf_{T} \int_{E} D(x, T(x))\mu(dx).$$

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

Kantorovich's problem

- (E, D) – metric space;
- $C(x, y)$ – cost function, e.g. $C(x, y) = D(x, y)$;
- $\mu, \nu \in \mathcal{P}_2$ – measures to be transported;
- $\mathcal{U}(\mu, \nu) = \left\{\pi \in \mathcal{P}(E \times E) : \int_E \pi(x, y)dy = \mu(x), \quad \int_E \pi(x, y)dx = \nu(y)\right\}$.

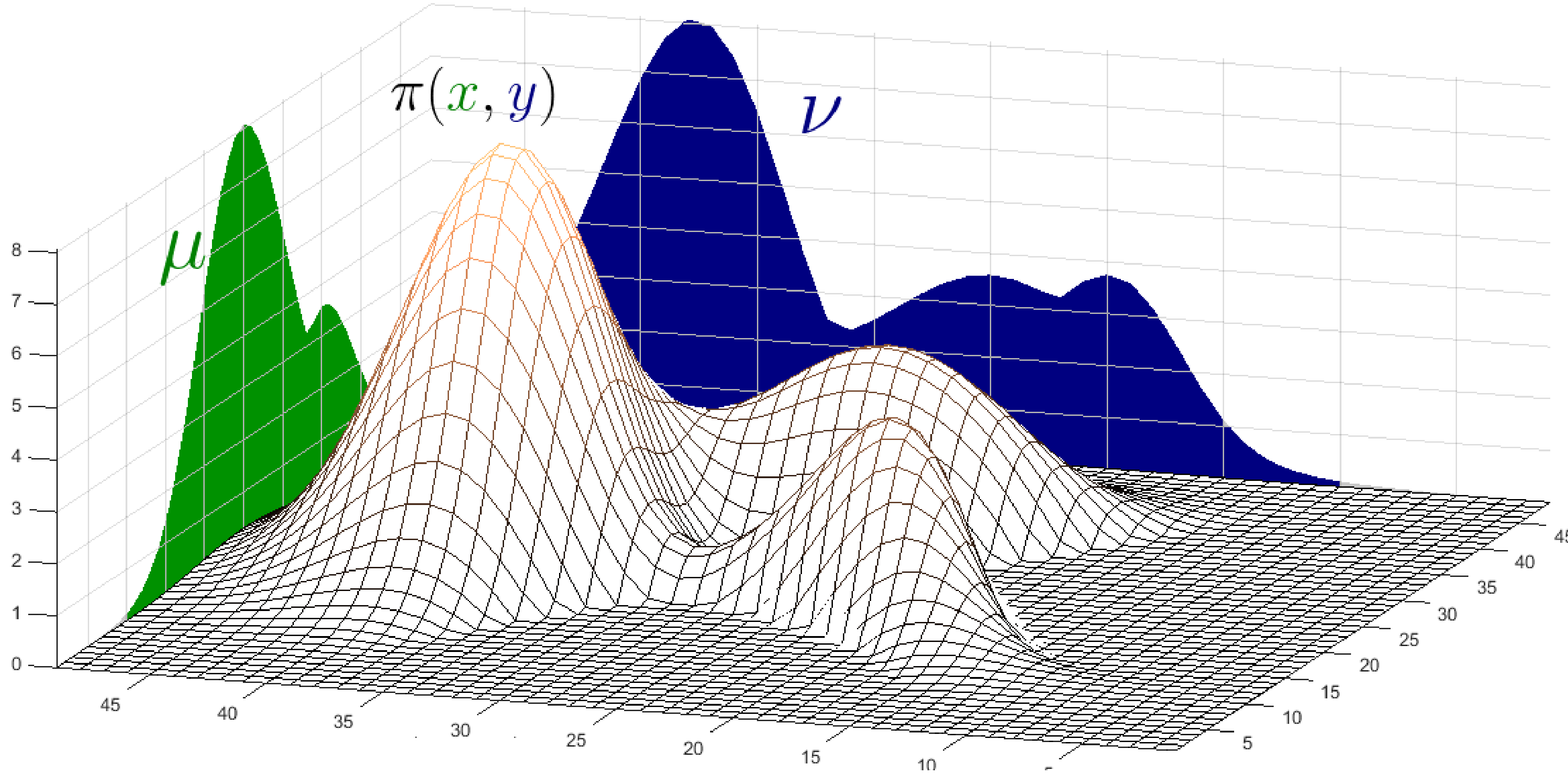
$$\inf_{\pi \in \mathcal{U}(\mu,\nu)} \int_{E\times E} C(x, y)d\pi(x, y).$$

Main feature: lifts ground metric D of a space E to the metric in the space of measures on E, e.g. p-Wasserstein distance

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

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

Kantorovich's relaxation

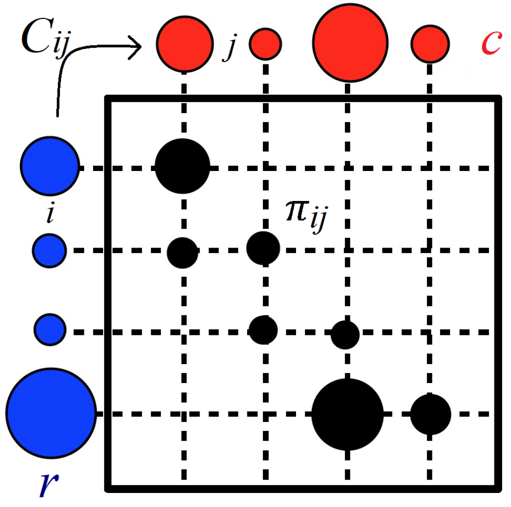


$$\inf_{\pi \in \mathcal{U}(\mu,\nu)} \int_{E\times E} C(x,y)d\pi(x,y).$$

Image: A.Suvorikova

Discrete case

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



Optimal transport problem

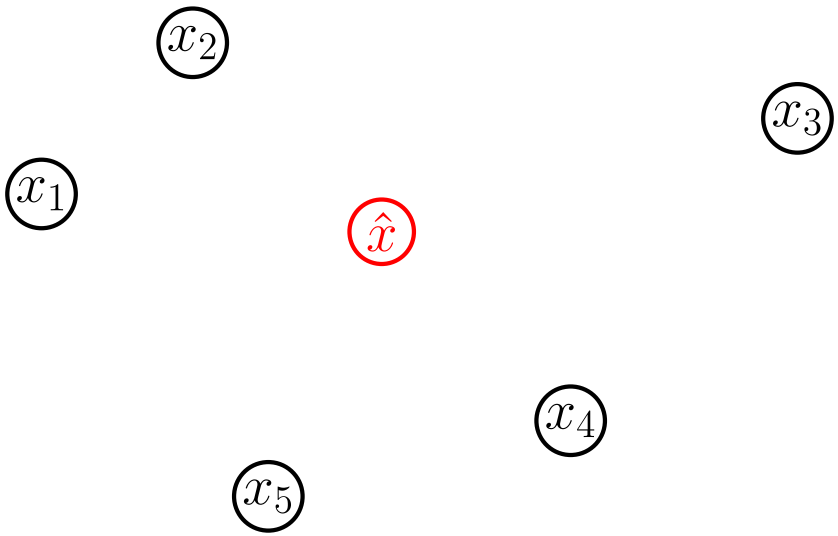
$$\min_{X \in \mathcal{U}(a,b)} \langle C, X \rangle,$$

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

Outline

1. Background on Optimal Transport
2. Motivation for Wasserstein barycenters
3. Numerical methods for Wasserstein barycenters
    - 3.1 Iterative Bregman Projections (IBP)
    - 3.2 Accelerated gradient method
    - 3.3 Accelerated Alternating Minimization

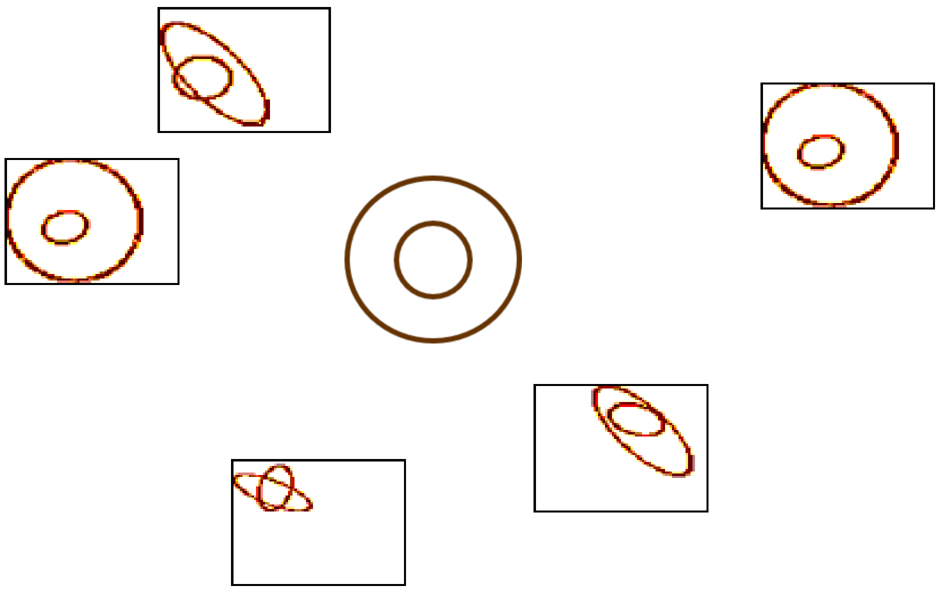
Euclidean Barycenter



$$x_i \sim \mathcal{N}(x, \sigma^2 I).$$

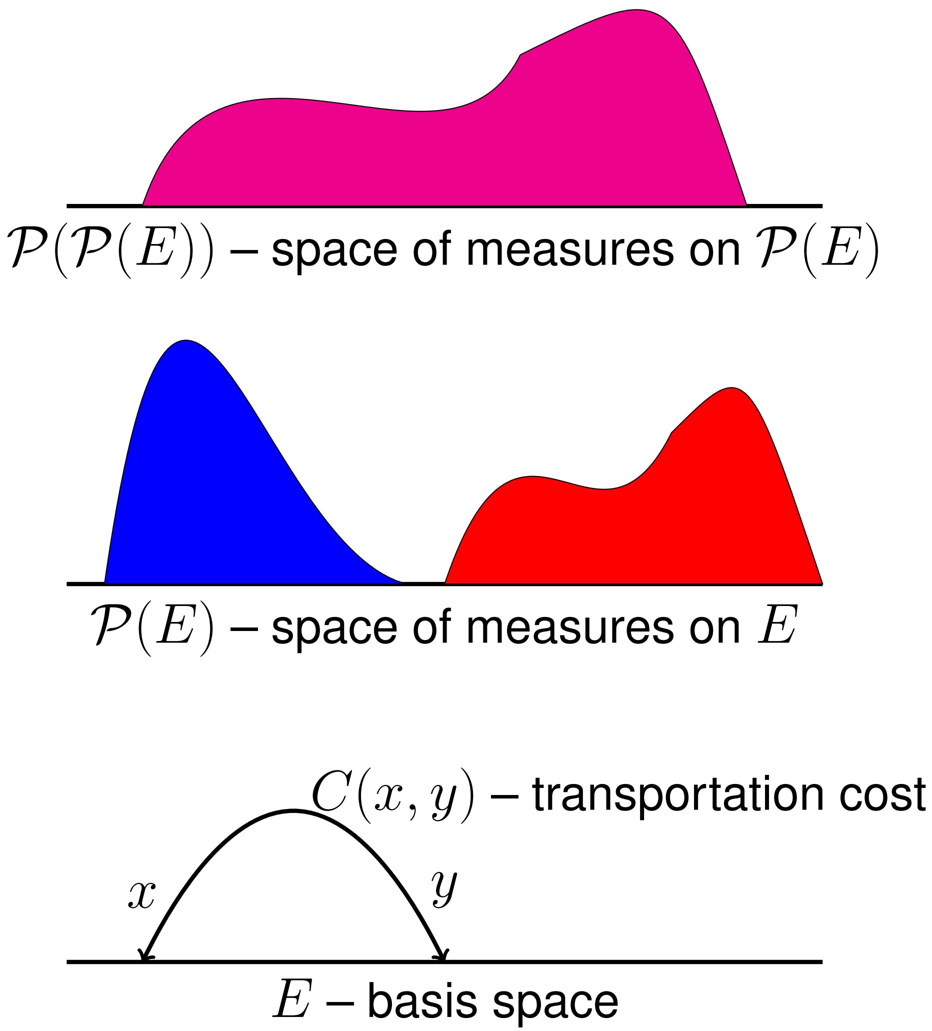
$$\hat{x} = \frac{1}{m}\sum_{i=1}^{m} x_i = \arg\min_{x} \frac{1}{m}\sum_{i=1}^{m} \|x - x_i\|_2^2.$$

Images Barycenter



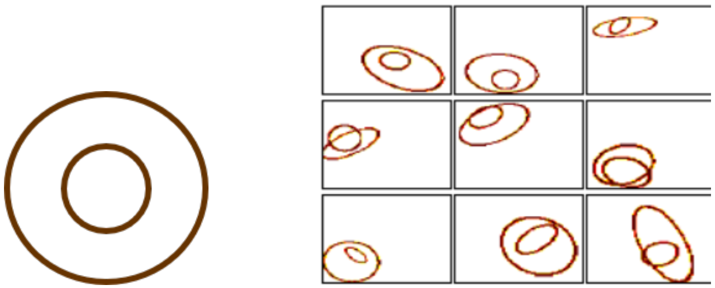
What is average image?
How to reconstruct a template?

Random picture model



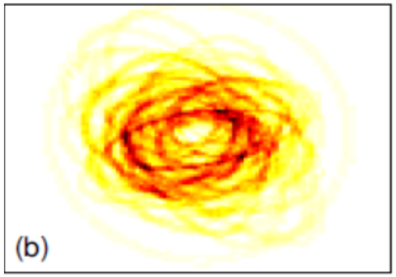
Template image reconstruction

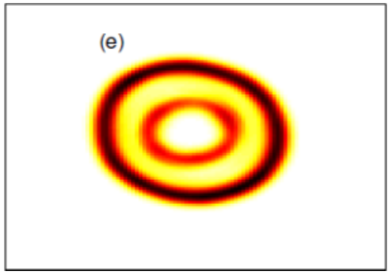
Goal: reconstruct template from its random transformations



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.

Template estimation from admissible deformations

Assume

- T – bounded random admissible transformation with finite moment $\mathbb{E}T$
- $T_i,\, i = 1, \ldots, m$ – 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

Outline

Background

Following [Cuturi & Doucet, 2014], we use entropic regularization.

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

Algorithms for barycenter

- Sinkhorn + Gradient Descent [Cuturi, Doucet, NeurIPS'13]
- 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]
- Distributed algorithms $\rightsquigarrow$ talk by Darina Dvinskikh.

Large-scale problem of dimension $mn^2 + n$.
Complexity? How to choose γ? How to scale-up computations?

Outline

Iterative Bregman Projections: Dual view

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

Dual problem:

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

$\mathtt{u} = [u_1, \dots, u_m]$, $\mathtt{v} = [v_1, \dots, v_m]$, $u_l, v_l \in \mathbb{R}^n$,
$B_l(u_l, v_l) := \operatorname{diag}(e^{u_l}) \exp(-C/\gamma) \operatorname{diag}(e^{v_l})$, $K = \exp(-C/\gamma)$.

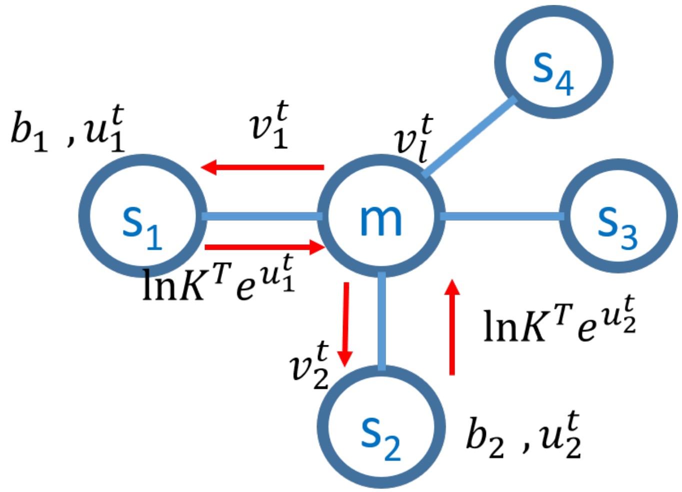
IBP is equivalent to alternating minimization for the dual problem.

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

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

IBP complexity

- 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/workers).



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

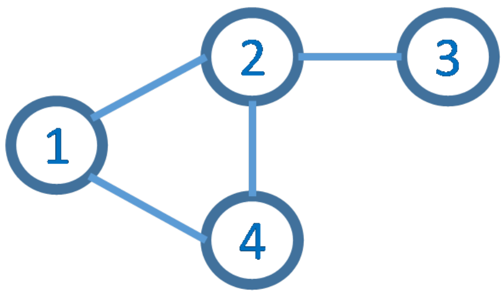
Outline

1. Background on Optimal Transport
2. Motivation for Wasserstein barycenters
3. Numerical methods for Wasserstein barycenters
    - 3.1 Iterative Bregman Projections (IBP)
    - 3.2 Accelerated gradient method
    - 3.3 Accelerated Alternating Minimization

Dual reformulation

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

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



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

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

Equivalent form of the WB problem

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

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

Dual reformulation

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

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

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 $\mathcal{W}_\gamma(\cdot, b)$

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

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

NB: $\Phi(\boldsymbol{\lambda})$ has Lipschitz-continuous gradient.

Convergence theorem

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

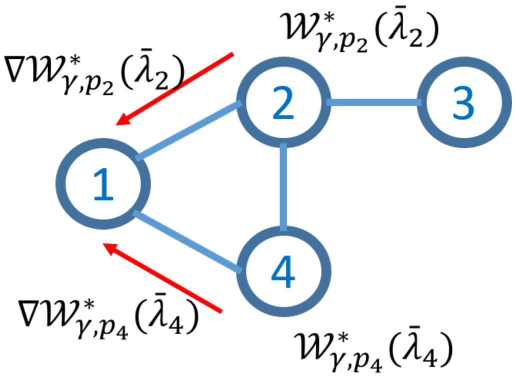
$$\hat{a}_i^{(k)} = \frac{1}{k+1} \sum_{\ell=0}^{k} a([L\boldsymbol{\lambda}^{(\ell)}]_i).$$

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

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

AGM complexity

- 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

For the extension to stochastic optimization setting see [Dvurechensky et. al., NeurIPS 2018].

Complexity of barycenter problem

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

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

$$\hat{a} \text{ s.t. } \frac{1}{m}\sum_{i=1}^{m}\mathcal{W}(\hat{a}, b_i) - \frac{1}{m}\sum_{i=1}^{m}\mathcal{W}(a^*, b_i) \le \varepsilon$$

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

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

$$\hat{a}_i,\, i = 1, \ldots, m \text{ s.t. } \frac{1}{m}\sum_{i=1}^{m}\mathcal{W}(\hat{a}_i, b_i) - \frac{1}{m}\sum_{i=1}^{m}\mathcal{W}(a^*, b_i) \le \varepsilon, \quad \|L\hat{\mathbf{a}}\|_2 \le \varepsilon.$$

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

Recently [Dvinskikh & Tiapkin, AISTATS 2021] proposed an algorithm with improved complexity $O\left(\frac{mn^2}{\varepsilon}\right)$ a.o.

Outline

1. Background on Optimal Transport
2. Motivation for Wasserstein barycenters
3. Numerical methods for Wasserstein barycenters
    - 3.1 Iterative Bregman Projections (IBP)
    - 3.2 Accelerated gradient method
    - 3.3 Accelerated Alternating Minimization

Motivation for Accelerated Alternating Minimization

Dual problem:

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

$\mathtt{u} = [u_1, \ldots, u_m]$, $\mathtt{v} = [v_1, \ldots, v_m]$, $u_l, v_l \in \mathbb{R}^n$,
$B_l(u_l, v_l) := \operatorname{diag}(e^{u_l}) \exp(-C/\gamma) \operatorname{diag}(e^{v_l})$, $K = \exp(-C/\gamma)$.

- Iteration Bregman Projection can be seen as Alternating Minimization in the dual.
- Accelerated Gradient Method also operates in the dual.
- Can we combine Alternating Minimization and Accelerated Gradient Method?

Yes.

General problem: block structure

We consider minimization problem

$$\min_{\lambda \in \mathbb{R}^N} \varphi(\lambda).$$

- The space $\mathbb{R}^N$ is divided into n disjoint subsets (blocks) I_p, $p \in \{1, \dots, n\}$.
- $S_p(\lambda) = \lambda + \operatorname{span}\{e_i : \ i \in I_p\}$, i.e. the affine subspace containing λ and all the points differing from λ only over the block p.
- λ_i – components of λ corresponding to the block i and $\nabla_i \varphi(\lambda)$ – gradient corresponding to the block i.
- Assume that for any $p \in \{1, \dots, n\}$ and any $\zeta \in \mathbb{R}^N$ the problem $\min\limits_{\lambda \in S_p(\zeta)} \varphi(\lambda)$ has a solution, and this solution is easily computable.
- $\varphi(\lambda)$ is L_φ-smooth: $\forall\ \lambda,\ \eta \in \mathbb{R}^N \quad \|\nabla \varphi(\lambda) - \nabla \varphi(\eta)\|_2 \leqslant L_\varphi \|\lambda - \eta\|_2$.

Adaptive Primal-Dual Accelerated Gradient Descent

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

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

repeat {Main iterate}

  repeat {Line search}

    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 $\tau_k = \alpha_{k+1}/\beta_{k+1}$.

    [Coupling step] $\lambda_{k+1} = \tau_k \zeta_k + (1 - \tau_k)\eta_k$.

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

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

  until

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

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

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

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

Ensure: $\hat{x}_{k+1}$, η_{k+1}.

Changes in APDAGD

Instead of gradient step $\eta_{k+1} = \lambda_{k+1} - \frac{1}{L}\nabla\varphi(\lambda_{k+1})$ we consider the Gauss-Southwell rule:
Choose $i_k = \arg\max_{i\in\{1,\dots,n\}} \|\nabla_i \varphi(\lambda_{k+1})\|_2^2$. Set $\eta_{k+1} = \arg\min_{\eta\in S_{i_k}(\lambda_{k+1})} \varphi(\eta)$.
Momentum step $\zeta_{k+1} = \zeta_k - \alpha_{k+1}\nabla\varphi(\lambda_{k+1})$, $\beta_k + \alpha_{k+1} = L\alpha_{k+1}^2$

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

$$[\eta_{k+1} = \lambda_{k+1} - \frac{1}{L}\nabla\varphi(\lambda_{k+1})] = \varphi(\lambda_{k+1}) - \frac{1}{2L}\|\nabla\varphi(\lambda_{k+1})\|_2^2$$

$$\beta_k + \alpha_{k+1} = L\alpha_{k+1}^2 \quad \to \quad \varphi(\eta_{k+1}) = \varphi(\lambda_{k+1}) - \frac{\alpha_{k+1}^2}{2(\beta_k + \alpha_{k+1})}\|\nabla\varphi(\lambda_{k+1})\|_2^2$$

Coupling step

$$\lambda_{k+1} = \tau_k\zeta_k + (1-\tau_k)\eta_k \quad \to \quad \tau_k = \arg\min_{\tau\in[0,1]} \varphi\left(\zeta_k + \tau(\eta_k - \zeta_k)\right)$$

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

Primal-dual accelerated alternating minimization

$\beta_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.

for $k \geqslant 0$ **do**

Set $\tau_k = \arg\min_{\tau \in [0,1]} \varphi\left(\eta_k + \tau(\zeta_k - \eta_k)\right)$

[Coupling step] Set $\lambda_k = \tau_k \zeta_k + (1 - \tau_k)\eta_k$

[Gauss-Southwell] Choose $i_k = \arg\max_{i \in \{1,\dots,n\}} \|\nabla_i \varphi(\lambda_k)\|_2^2$.

Set $\eta_{k+1} = \arg\min_{\eta \in S_{i_k}(\lambda_k)} \varphi(\eta)$.

Find α_{k+1}, $\beta_{k+1} = \beta_k + \alpha_{k+1}$ from

$$\varphi(\lambda_k) - \frac{\alpha_{k+1}^2}{2(\beta_k + \alpha_{k+1})} \|\nabla \varphi(\lambda_k)\|_2^2 = \varphi(\eta_{k+1})$$

[Update momentum] Set $\zeta_{k+1} = \zeta_k - \alpha_{k+1} \nabla \varphi(\lambda_k)$

[Primal update] Set $\hat{x}_{k+1} = \frac{\alpha_{k+1} x(\lambda_k) + \beta_k \hat{x}_k}{\beta_{k+1}}$.

end for

Ensure: The points $\hat{x}_{k+1}$, η_{k+1}.

Convergence theorem

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

$$f(\hat{x}_k) - f^* \leq f(\hat{x}_k) + \varphi(\eta_k) \leq \frac{4nL_\varphi R^2}{k^2} = \frac{8n\|A\|^2_{E\to H}R^2}{\gamma k^2} = O\left(\frac{n}{\gamma k^2}\right),$$

$$\|A\hat{x}_k - b\|_2 \leq \frac{8n\|A\|^2_{E\to H}R}{\gamma k^2} = O\left(\frac{n}{\gamma k^2}\right),$$

$$\|\hat{x}_k - x^*\|_E \leq \frac{4n}{k}\frac{\|A\|_{E\to H}R}{\gamma} = O\left(\frac{n}{\gamma k}\right),$$

x^*, f^* – resp. an optimal solution and the optimal value in the primal problem.
Assume that algorithm is applied for a non-convex and L_φ-smooth objective $\varphi(\cdot)$.
Then

$$\min_{i=0,\dots,k} \|\nabla\varphi(\lambda_i)\|_2^2 \leqslant \frac{2nL_\varphi(\varphi(\lambda_0) - \varphi(\lambda^*))}{k}.$$

Uniformly optimal method for smooth convex and non-convex problems, no knowledge of the parameters.

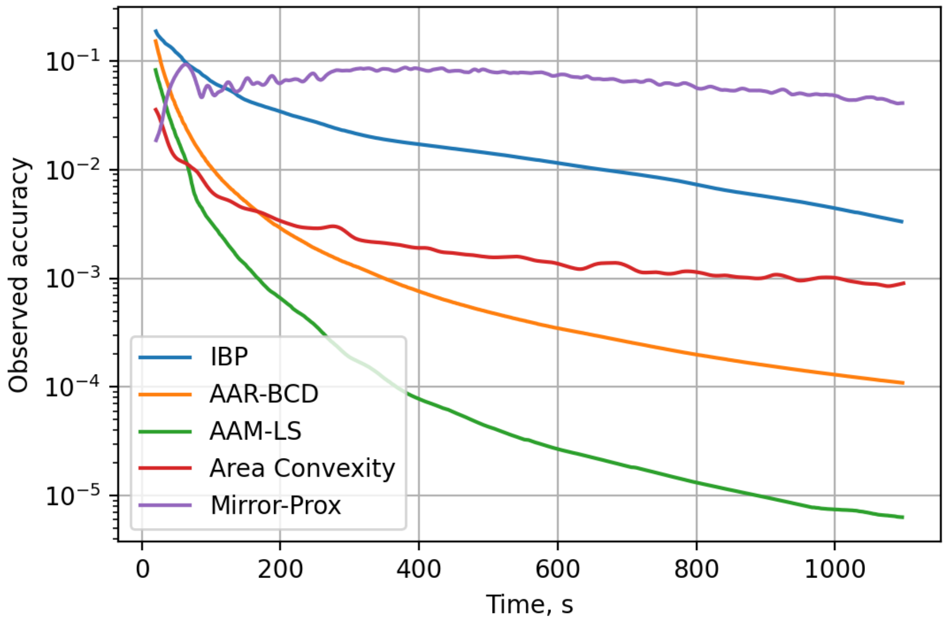
Application to Wasserstein barycenter

Modified dual problem

$$\min_{\substack{u,v \\ \sum_{l=1}^m w_l v_l = 0}} \gamma \sum_{l=1}^{m} w_l \left\{ \ln\left(\mathbf{1}^T B_l(u_l, v_l)\,\mathbf{1}\right) - \langle u_l, p_l \rangle \right\}$$

Complexity similar to Accelerated Gradient Method $O\left(\frac{mn^{2.5}}{\varepsilon}\right)$ a.o.

Good practical performance, e.g. on MNIST dataset:



Conclusions

- Wasserstein barycenter allows to “average” complex objects.
- Complexity by IBP/AGD/AAM.
- Distributed algorithms to scale-up computations.
- Accelerated algorithms and their applications to Wasserstein barycenter problem.

Publications

- D., Gasnikov, Kroshnin, Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorns algorithm. ICML 2018.
- D., Dvinskikh, Gasnikov, Uribe, Nedic Decentralize and randomize: Faster algorithm for Wasserstein barycenters. NeurIPS 2018.
- Kroshnin, Dvinskikh, D., Gasnikov, Tupitsa, Uribe On the Complexity of Approximating Wasserstein Barycenter. ICML 2019.
- S. Guminov, D., N. Tupitsa, A. Gasnikov, Accelerated Alternating Minimization, Accelerated Sinkhorn’s Algorithm and Accelerated Iterative Bregman Projections, ICML 2021.

Thank you!