Doc-Start
Doc-Start
Doc-Start
Doc-Start

Decentralized Computation of Wasserstein Barycenters

César A. Uribe

Alexey Kroshnin Darina Dvinskikh Pavel Dvurechensky

Alexander Gasnikov Nazarii Tupitsa Roman Krawtschenko

Ivan Lau Shiqian Ma Dmitry Pasechnyuk Angelia Nedić

May 19, 2025

RICE ENGINEERING

Electrical and Computer Engineering

Outline

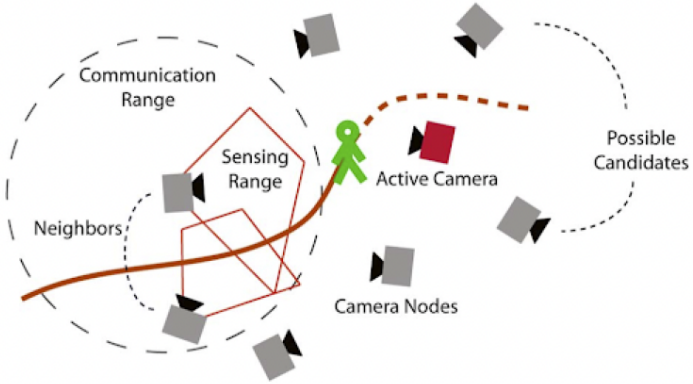
- Distributed Optimization 101
- Primal-Dual and Decentralized Computations
- Discrete and Semi-Discrete Barycenters
- Semi-Discrete Barycenters and Quantization
- Equitable OT

Distributed Optimization 101

Motivation

(a) Sensor Networks in Agriculture

(b) (Mis)information Spread



(c) Camera Networks for Security

(d) Huge-scale ML

some intersecting points

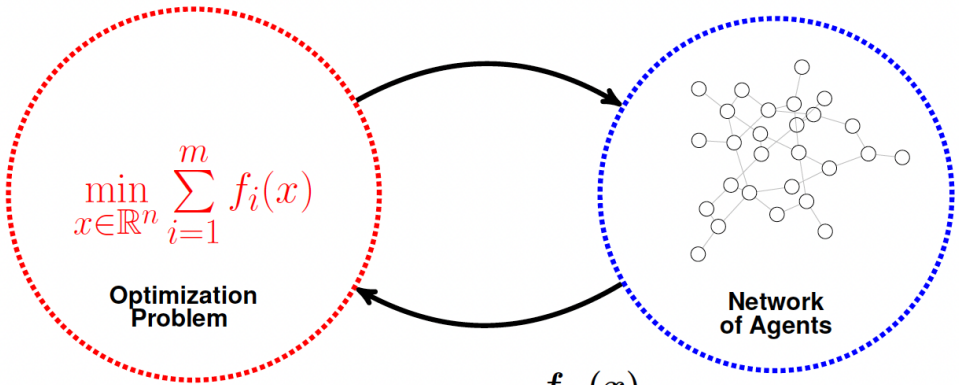
Characteristics

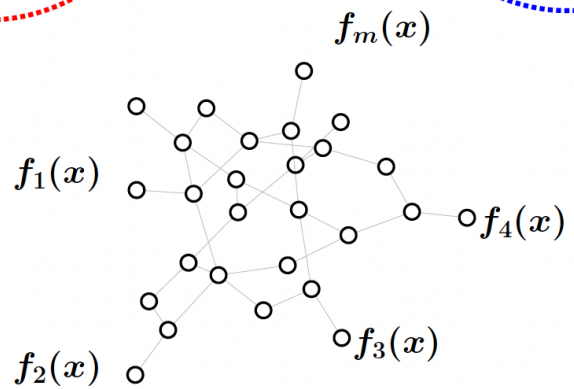
- Many components/units (we call them agents).
- Connected over networks.
- Cyber and Physical interactions.
- Distributed Storage.

Challenges

- Decentralization: distributed computations.
- Scalability: Price of decentralization.
- Optimality: Efficiency & Performance
- Robustness & Resiliency: Performance under failures

Distributed Optimization 101: Object of Study

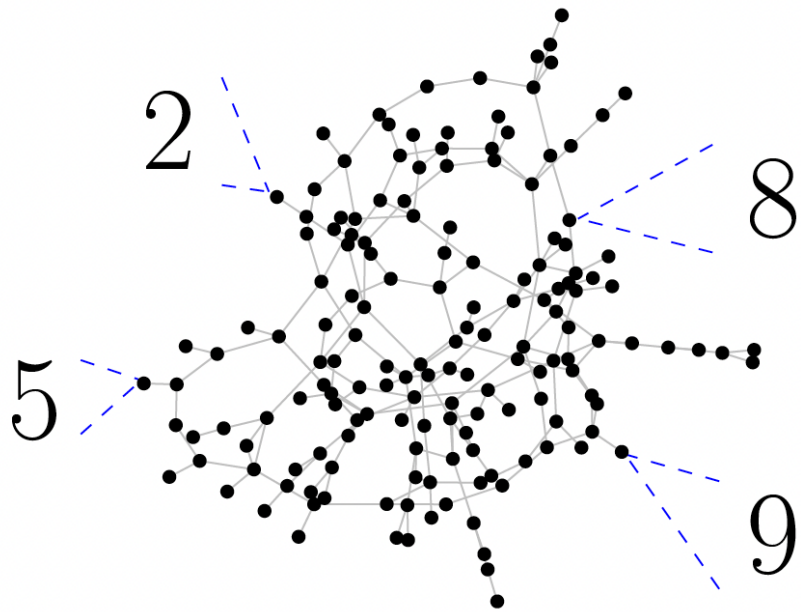




DECENTRALIZED **SCALABLE** **OPTIMAL**

An example

- There is a network of m agents, i.e. a graph $G = \{V, E\}$.
- Agent i holds an initial value $x_0^i \in \mathbb{R}$.
- Each agent needs to distributedly compute $\frac{1}{m}\sum_{i=1}^{m} x_0^i$.



Equivalently, solve $\min_{x\in\mathbb{R}} \frac{1}{2}\sum_{i=1}^{m}\|x - x_i\|_2^2$.

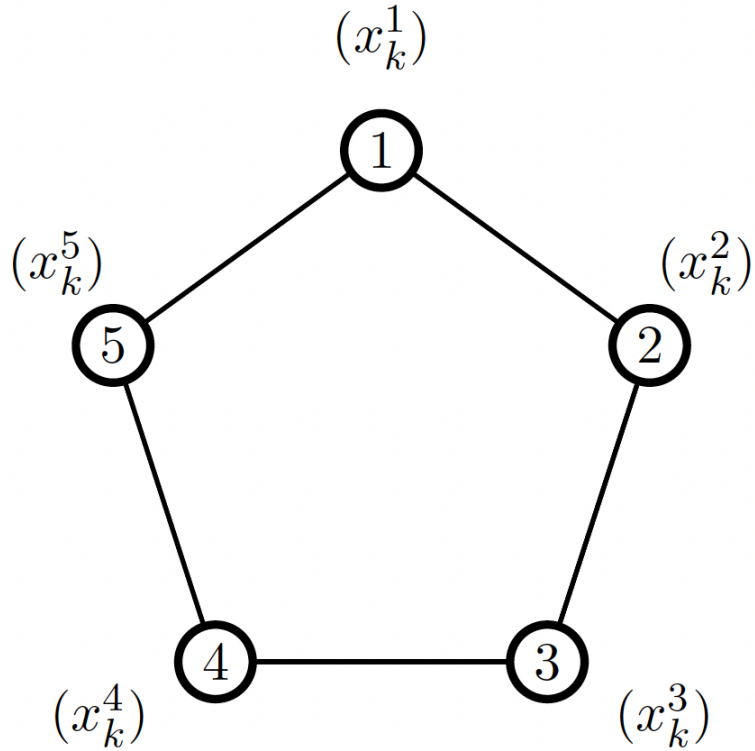
A fundamental result

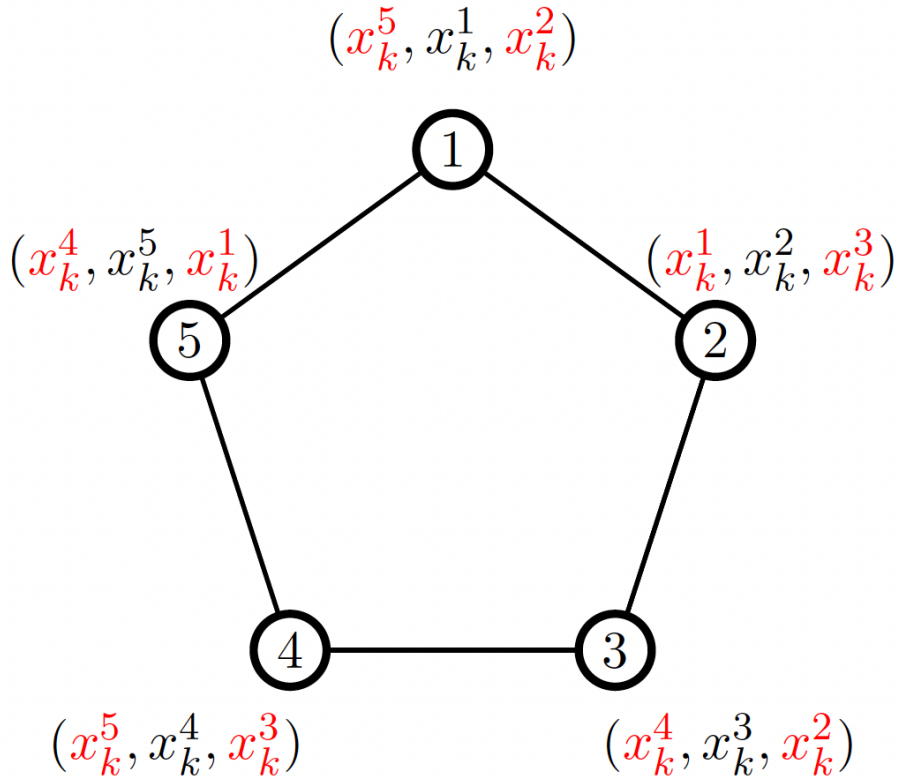
$$x_{k+1}^i = \sum_{j=1}^{m} [A]_{ij} \, x_k^j \tag{1}$$

FUNDAMENTAL RESULT:

If $\mathcal{G}$ is connected, undirected, and static, and A is *doubly stochastic* (i.e. $[A]_{ij} > 0$ iff $(j, i) \in E$), then the iterates generated by **(1)** satisfy

$$\lim_{k \to \infty} x_k^i = \frac{1}{m} \sum_{j=1}^{m} x_0^j, \qquad \forall \, i \in V.$$

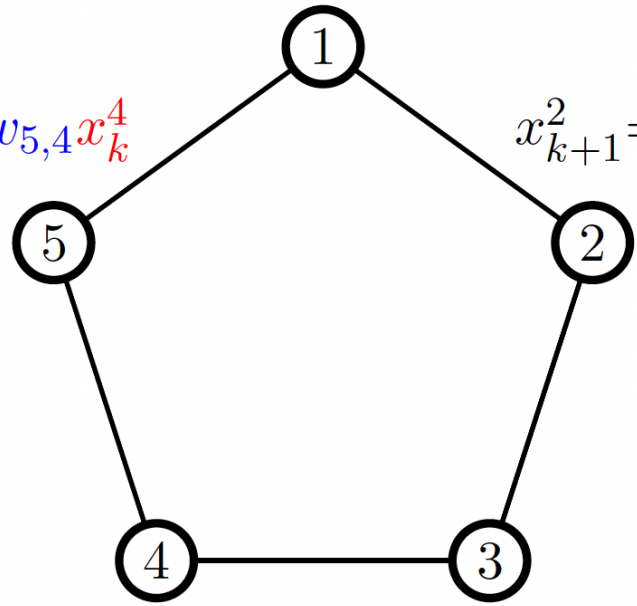

(x_k^1)
1
(x_k^5)
5
(x_k^2)
2
4
3
(x_k^4)
(x_k^3)


(x_k^5, x_k^1, x_k^2)
1
(x_k^4, x_k^5, x_k^1)
(x_k^1, x_k^2, x_k^3)
5
2
4
3
(x_k^5, x_k^4, x_k^3)
(x_k^4, x_k^3, x_k^2)

$$x^1_{k+1} = w_{1,5}x^5_k + w_{1,1}x^1_k + w_{1,2}x^2_k$$

$$x^5_{k+1} = w_{5,1}x^1_k + w_{5,5}x^5_k + w_{5,4}x^4_k$$

$$x^2_{k+1} = w_{2,1}x^1_k + w_{2,2}x^2_k + w_{2,3}x^3_k$$



$$x^4_{k+1} = w_{4,5}x^5_k + w_{4,4}x^4_k + w_{4,3}x^3_k$$

$$x^3_{k+1} = w_{3,4}x^4_k + w_{3,3}x^3_k + w_{3,2}x^2_k$$

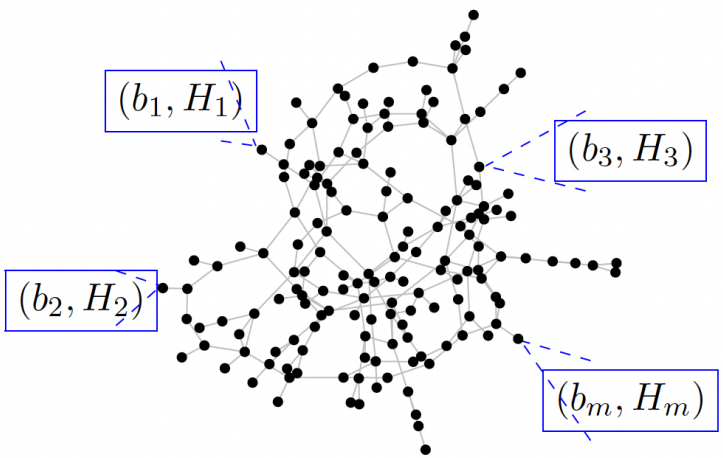
An Example: Distributed Ridge Regression

We want to estimate x assuming

$$b_i = H_i x + noise,$$

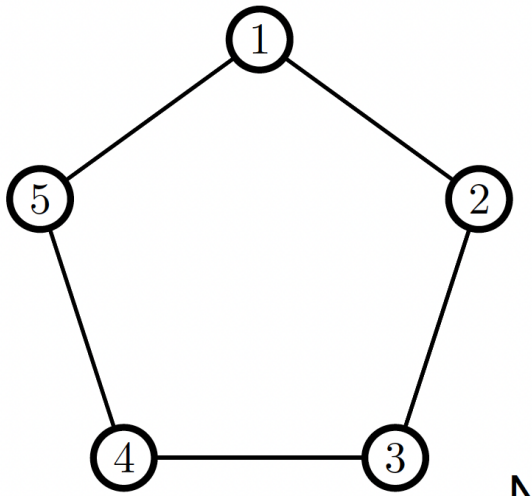
where

- $H_i \in \mathbb{R}^{d_i \times n}$: d_i data points of dimension n.
- $b_i \in \mathbb{R}^{d_i}$: d_i outputs.



$$\min_x \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m} \|b_i - H_i x\|_2^2.$$

A little bit of analysis



$$\bar{W} = \begin{bmatrix} 2 & -1 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

Note that:

- $Wx = 0$ if and only if $x_1 = \ldots = x_m$.

$$x = \begin{bmatrix} x_1 \in \mathbb{R}^n \\ x_2 \in \mathbb{R}^n \\ \vdots \\ x_m \in \mathbb{R}^n \end{bmatrix}$$

Rewrite problem (2) in an equivalent form as follows:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} f_i(x) \qquad \text{equivalent to} \qquad \min_{Wx=0} \sum_{i=1}^{m} f_i(x_i), \qquad (6)$$

$$\text{where} \quad W = \bar{W} \otimes I_n.$$

Initially, consider the general problem

$$\min_{Ax=0} f(x). \tag{7}$$

We assume that the problem has optimal solutions. Later, we will derive the specific results when

$$A = \sqrt{\bar{W}} \qquad \text{and} \qquad f(x) = \sum_{i=1}^{m} f_i(x_i).$$

Approximate Solution Definition A point $x \in \mathbb{R}^{mn}$ is said to be an $(\varepsilon, \tilde{\varepsilon})$-solution of **(7)** if

$$f(x) - f^* \ \leq \ \varepsilon \quad \text{and} \quad \|Ax\|_2 \ \leq \ \tilde{\varepsilon},$$

where f^* denotes the optimal value of (7).

The Lagrangian dual for the problem in **(7)** is given by

$$\min_{Ax=0} f(x) \;=\; \max_{y}\Big\{\; \min_{x}\{\, f(x) - \langle A^T y, x\rangle\}\Big\},$$

or equivalently

$$\min_{y} \varphi(y) \quad \text{where} \quad \varphi(y) \;\triangleq\; \max_{x}\{\, \langle A^T y, x\rangle - f(x)\},$$

$$\text{with} \quad \nabla\varphi(y) = Ax^*(A^T y) \quad \text{(Demyanov–Danskin)}$$

$x^*(A^T y) = \arg\max_x\{\, \langle A^T y, x\rangle - f(x)\}.$

We say that f is dual friendly when we can determine a solution of the preceding problem efficiently (in a closed form ideally).

- $f(x)$ is μ-strongly convex $\iff$ $\varphi(y)$ is L_φ-smooth with $L_\varphi = \lambda_{\max}(A^T A)/\mu$.
- $f(x)$ is L-smooth $\iff$ $\varphi(y)$ is μ_φ-strongly convex on range(A) with $\mu_\varphi = \lambda^+_{\min}(A^T A)/L$.

The dual problem $\min_y \varphi(y)$ may have multiple solutions of the form $y^* + \ker(A^T)$.

Informally: If $f(x)$ has condition number L/μ, then $\varphi(y)$ has condition number

$$\frac{\lambda_{\max}(A^T A)}{\lambda^+_{\min}(A^T A)} \frac{L}{\mu}.$$

Let's recall a version of Nesterov's fast gradient method for

$$\min_y \ \varphi(y)$$

$$y_{k+1} = \tilde{y}_k \ - \ \frac{1}{L_\varphi} \nabla \varphi(\tilde{y}_k),$$

$$\tilde{y}_{k+1} = y_{k+1} \ + \ \frac{\sqrt{L_\varphi} - \sqrt{\mu_\varphi}}{\sqrt{L_\varphi} + \sqrt{\mu_\varphi}} (y_{k+1} - y_k).$$

and

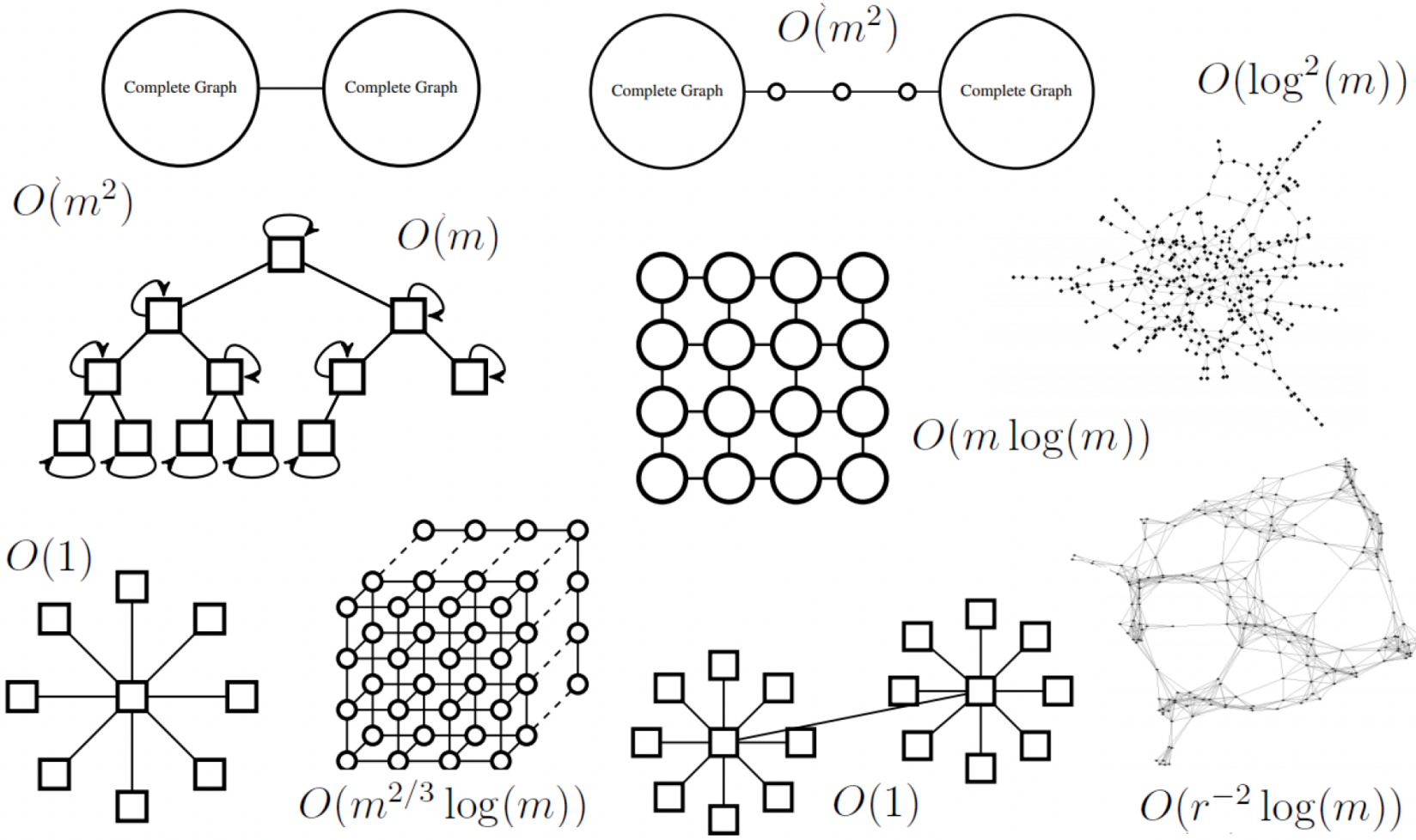
$$\varphi(y_k) - \varphi^* \ \leq \ L_\varphi \left(1 - \sqrt{\frac{\mu_\varphi}{L_\varphi}}\right)^k \left\|y_0 - y^*\right\|_2^2,$$

Set $A = \sqrt{\bar{W}},\ z_k = \sqrt{\bar{W}}\, y_k,\ \text{and}\ \tilde{z}_k = \sqrt{\bar{W}}\, \tilde{y}_k$

$$x_i^*\left(\tilde{z}_k^i\right) \ = \ \arg\max_{x_i}\left\{ \langle \tilde{z}_k^i, x_i \rangle - f_i(x_i) \right\}$$

$$z_{k+1}^i \ = \ \tilde{z}_k^i \ - \ \frac{\mu}{\lambda_{\max}(W)} \sum_{j=1}^{m} W_{ij}\, x_j^*\left(\tilde{z}_k^j\right)$$

$$\tilde{z}_{k+1}^i \ = \ z_{k+1}^i \ + \ \frac{\sqrt{\lambda_{\max}(W)/\mu} - \sqrt{\lambda_{\min}^+(W)/L}}{\sqrt{\lambda_{\max}(W)/\mu} + \sqrt{\lambda_{\min}^+(W)/L}} \left(z_{k+1}^i - z_k^i\right).$$

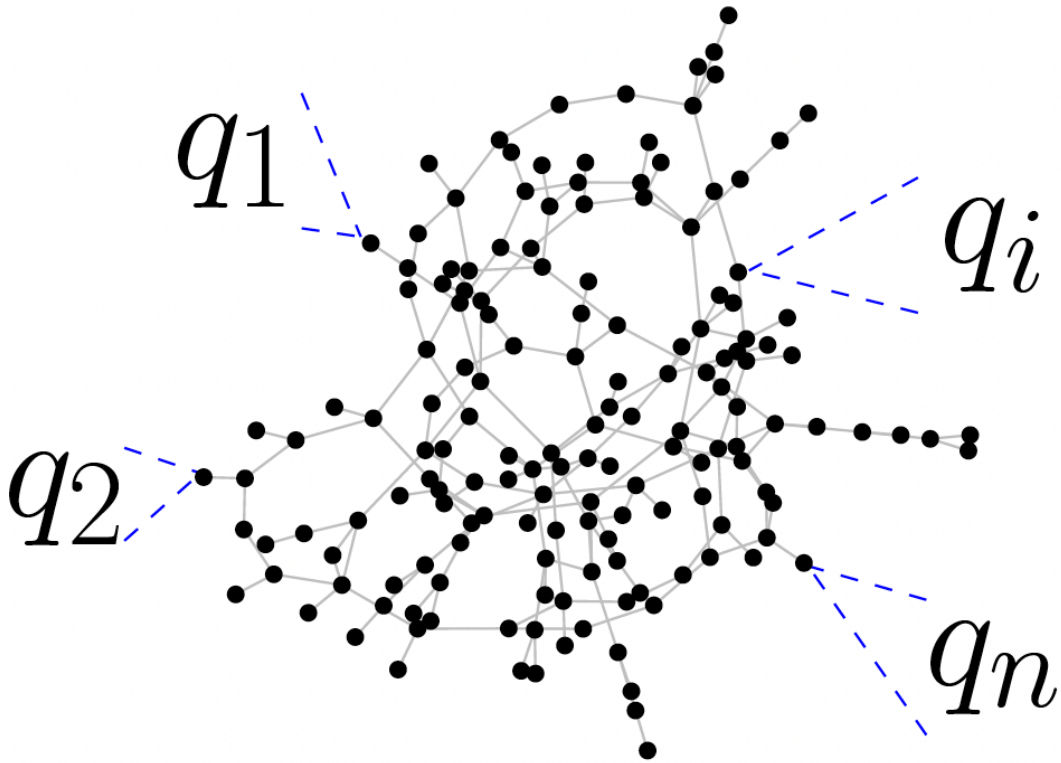

Complete Graph
Complete Graph
$O(m^2)$
Complete Graph
Complete Graph
$O(\log^2(m))$
$O(m^2)$
$O(m)$
$O(m \log(m))$
$O(1)$
$O(m^{2/3} \log(m))$
$O(1)$
$O(r^{-2} \log(m))$

Decentralized Computations of Discrete Wasserstein Barycenters

Why Decentralized Wasserstein Barycenters?

- **Distributed-data reality**
  Probability measures reside on edge devices (IoT sensors, mobile phones, robots, medical sites); raw data aggregation is infeasible or prohibited.
- **Geometry-preserving averaging**
  Wasserstein barycenters provide the correct *geometric mean* of distributions, enabling domain adaptation, sensor-fusion maps, and manifold-aware clustering.
- **Privacy & compliance**
  Local measures may contain sensitive content (e.g. medical images); decentralized protocols keep data in place, sharing only aggregated or quantized statistics.
- **Scalability & fault-tolerance**
  Centralized Sinkhorn/IBP requires $O(mn^2)$ memory; distributing both storage *and* computation removes single points of failure and scales to thousands of agents.
- **Long-term vision**
  Provably-optimal decentralized barycenters underpin federated OT learning, real-time cooperative perception, and privacy-preserving generative modelling.

What if the data are probability distributions



Equivalently, solve $\min_{q \in \mathbb{P}} \frac{1}{2} \sum_{i=1}^{m} d^2(p, q_i)$.

Entropy Regularized Discrete Optimal Transport (OT)

Given $p, q \in \Delta_n$ and cost $C \in \mathbb{R}^{n \times n}_{\geq 0}$,

$$W(p,q) = \min_{\pi \in \Pi(p,q)} \langle \pi, C \rangle, \quad \Pi(p,q) = \{\pi \geq 0 : \pi \mathbf{1} = p,\ \pi^T \mathbf{1} = q\}.$$

Cuturi (2013) solves large OT by adding entropy:

$$W_\gamma(p,q) = \min_{\pi \in \Pi(p,q)} \langle \pi, C \rangle + \gamma \mathrm{KL}(\pi \| \mathbf{1}\mathbf{1}^T).$$

Gains:

- Strong convexity $\Rightarrow$ unique optimal plan π^γ.
- Sinkhorn scaling: $O(n^2)$ per iteration with linear rate $1 - \Theta(\gamma)$.

Wasserstein Barycenter

For $\{p_\ell\}_{\ell=1}^m \subset \Delta_n$ and weights $w \in \Delta_m$,

$$q^* = \arg\min_{q \in \Delta_n} \sum_{\ell=1}^{m} w_\ell \, W_\gamma(p_\ell, q).$$

Our objective: compute q^* in a network where p_ℓ reside on distinct agents.

Dual Representation — Theorem 1

Theorem 1: (Theorem 2.4 in (Cuturi et al., 2016)

For $\gamma > 0$, the Fenchel–Legendre dual function $W^*_{\gamma,q}(y)$ is differentiable and its gradient is $\frac{1}{\gamma}$-Lipschitz in the 2-norm with

$$W^*_{\gamma,q}(y) = \gamma\big(E(q)+\langle q, \log K\alpha\rangle\big), \nabla W^*_{\gamma,q}(y) = \alpha \circ \frac{K\,q}{K\alpha} \in S_1(n),$$

where $y \in \mathbb{R}^n$, $\alpha = \exp(y/\gamma)$, and $K = \exp(-M/\gamma)$.

Centralized vs. Decentralized Formulation

Centralized

Full data at one server; complexity dominated by OT sub-problems $O(mn^2)$ each iteration.

Decentralized

$$\min_{\substack{q_1=\cdots=q_m \\ q_i\in\Delta_n}} \sum_{i=1}^{m} W_\gamma(p_i, q_i).$$

Consensus enforced by graph constraints; complexity coupled with spectral gap $\lambda^+_{\min}(W)$.

Convergence Guarantee

Theorem 2: Distributed Computation of WB

Let $\varepsilon > 0$ and $\|\nabla \mathcal{W}^*_{\gamma,\mathbf{q}}(\mathrm{y})\|_2 \leq G$ for $\mathrm{y} \in B_R(0)$ with $R = \|\mathrm{y}^*\|_2$. Then, for

$$N \geq \sqrt{\frac{16G^2}{\gamma \cdot \varepsilon} \frac{\lambda_{\max}(W)}{\lambda^+_{\min}(W)}},$$

$\mathrm{p}_N = [(p_N)_1^T, \cdots, (p_N)_m^T]^T$ and $\mathrm{y}_N = [(y_N)_1^T, \cdots, (y_N)_m^T]^T$ have the following properties:

$$\mathcal{W}_{\gamma,\mathbf{q}}(\mathrm{p}_N) + \mathcal{W}^*_{\gamma,\mathbf{q}}(\mathrm{y}_N) \leq \varepsilon \quad \text{and} \quad \|\sqrt{W}\mathrm{p}_N\|_2 \leq \varepsilon/R.$$

Lemma 8 — Radius of the Optimal Dual

Lemma 1

Let q_γ^* be the optimal solution of problem (4) with minimal 2-norm. Then there exists an optimal dual solution

$$\lambda^* = \left[\lambda_1^*, \ldots, \lambda_m^*\right]$$

for problem (29) satisfying

$$\|\lambda^*\|_2 \leq R, \qquad R^2 = \frac{2n \sum_{l=1}^m w_l^2 \, \|C_l\|_\infty^2}{\lambda_{\min}^+(W)}.$$

Here $\lambda_{\min}^+(W)$ is the minimal positive eigenvalue of the matrix W.

AGD Complexity for Barycenter

Theorem 3: Distributed Computation of WB

The Primal-Dual Accelerated Gradient Based method after

$$N = \frac{1}{\varepsilon}\sqrt{64\,\chi(\bar{W})\,mn\ln n\sum_{l=1}^{m} w_l^2\|C_l\|_\infty^2}$$

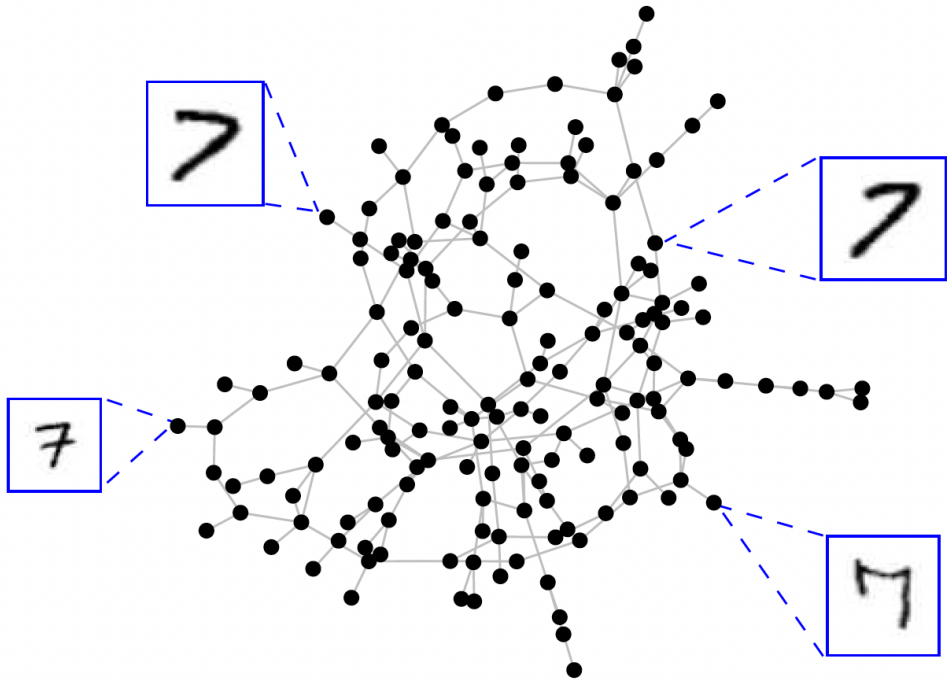
iterations generates an ε-solution of problem (2), i.e. finds a vector $q_N = [q_1^T, \ldots, q_m^T]^T$ such that

$$\sum_{l=1}^{m} w_l W(p_l, q_{N,l}) - \sum_{l=1}^{m} w_l W(p_l, q^*) \le \varepsilon, \qquad \|\sqrt{W}\, q_N\|_2 \le \varepsilon/(2R),$$

where q^* is an unregularized barycenter, and R is a bound on the solution to the dual problem. Moreover, the number of arithmetic operations is

$$O\big(N\,n\,(mn + \mathrm{nnz}(\bar{W}))/\varepsilon\big).$$

Proof by Picture



Video 1

Decentralized Computation of Semi-Discrete Wasserstein Barycenters

Semi–Discrete Regularized OT

$$W_\gamma(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \Big\{ \sum_{i=1}^{n} \int_Y c_i(y)\,\pi_i(y)\,dy + \gamma \sum_{i=1}^{n} \int_Y \pi_i(y) \log\Big(\frac{\pi_i(y)}{\xi}\Big)\,dy$$

$$\Pi(\mu,\nu) = \Big\{ \pi \in \mathcal{M}_+^1(Y) \times S_1(n) : \sum_{i=1}^{n} \pi_i(y) = q(y),\ \int_Y \pi_i(y)dy = p_i \Big\}$$

$$\min_{p\in S_1(n)} \sum_{i=1}^{m} W_{\gamma,\mu_i}(p)$$

Dual Formulation & Gradient

$$\min_{\lambda \in \mathbb{R}^{mn}} \sum_{i=1}^{m} W^*_{\gamma,\mu_i}\big([\sqrt{W}\,\lambda]_i\big)$$

$$W^*_{\gamma,\mu}(\bar{\lambda}) = \mathbb{E}_{Y\sim\mu}\,\gamma \log\Big(\frac{1}{q(Y)}\sum_{\ell=1}^{n}\exp\Big(\frac{\bar{\lambda}_\ell - c_\ell(Y)}{\gamma}\Big)\Big)$$

$$\big(\nabla W^*_{\gamma,\mu}(\bar{\lambda})\big)_l = \mathbb{E}_{Y\sim\mu}\frac{\exp\big((\bar{\lambda}_l - c_l(Y))/\gamma\big)}{\sum_{\ell=1}^{n}\exp\big((\bar{\lambda}_\ell - c_\ell(Y))/\gamma\big)}, \quad l = 1,\ldots,n$$

Smoothness & Stochastic Gradient Bounds

Lemma 1 (smoothness). For every μ, $W^*_{\gamma,\mu}$ is $\frac{1}{\gamma}$–smooth.

Lemma 2. The gradient of the dual objective satisfies

$$\|\nabla W^*_\gamma(\lambda_1) - \nabla W^*_\gamma(\lambda_2)\|_2 \leq \frac{\lambda_{\max}(W)}{\gamma} \|\lambda_1 - \lambda_2\|_2,$$

and the mini–batch estimator with size M obeys

$$\mathbb{E}\left\|\tilde{\nabla} W^*_\gamma(\lambda) - \nabla W^*_\gamma(\lambda)\right\|_2^2 \leq \frac{\lambda_{\max}(W)\, m}{M}.$$

Convergence of Algorithm 4

Theorem 4: Distributed Computation of the Semi-Discrete WB

Let $\|\lambda^*\|_2 \leq R$. After

$$N = \sqrt{\frac{32\,\lambda_{\max}(W)R^2}{\varepsilon\gamma}}$$

iterations, Algorithm 4 returns $\hat{p}_N$ such that

$$\sum_{i=1}^{m} W_{\gamma,\mu_i}\left(\mathbb{E}[\hat{p}_N]_i\right) - \sum_{i=1}^{m} W_{\gamma,\mu_i}(p_i^*) \leq \varepsilon, \qquad \left\|\sqrt{W}\,\mathbb{E}\hat{p}_N\right\|_2 \leq \varepsilon/R,$$

with total arithmetic cost

$$O\left(mn\max\left\{\sqrt{\frac{\lambda_{\max}(W)R^2}{\varepsilon\gamma}},\ \frac{\lambda_{\max}(W)mR^2}{\varepsilon^2}\right\}\right).$$

Proof by Picture 2

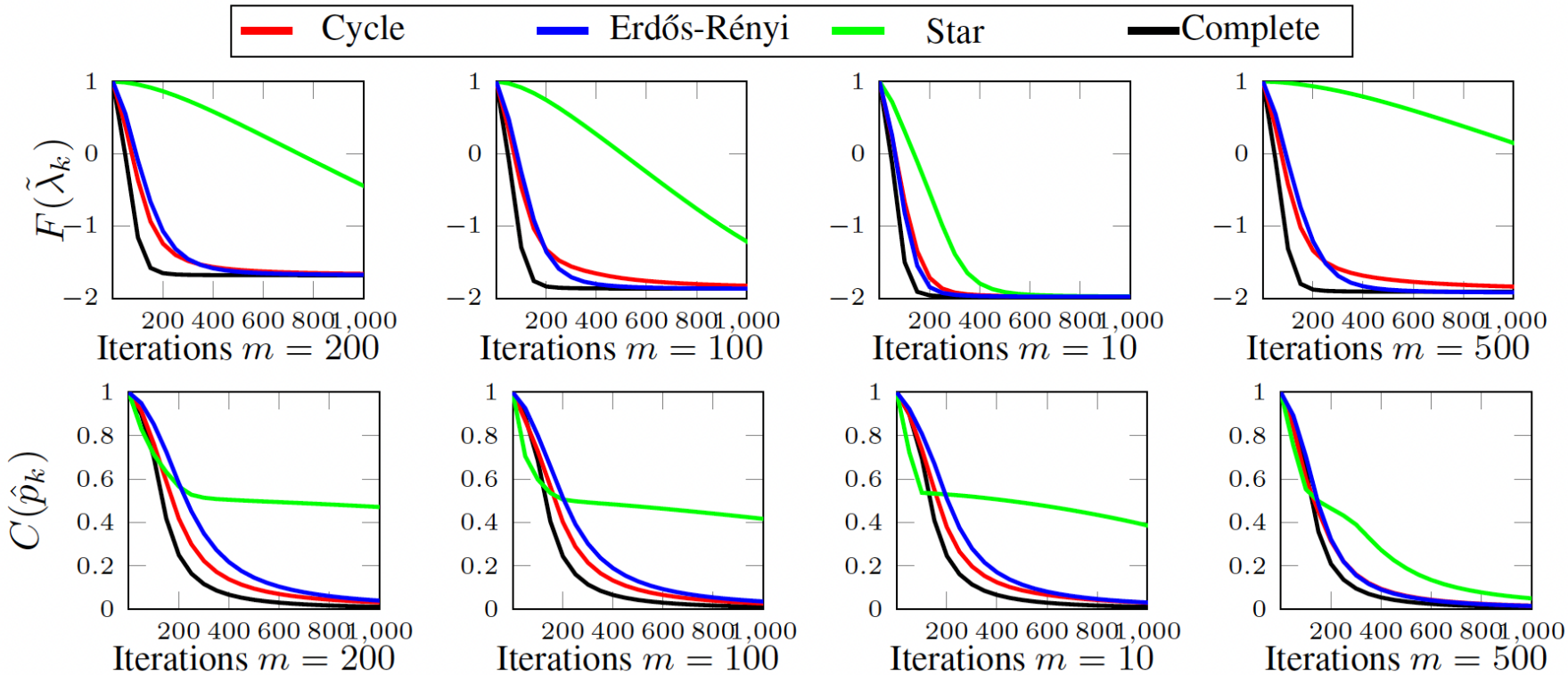
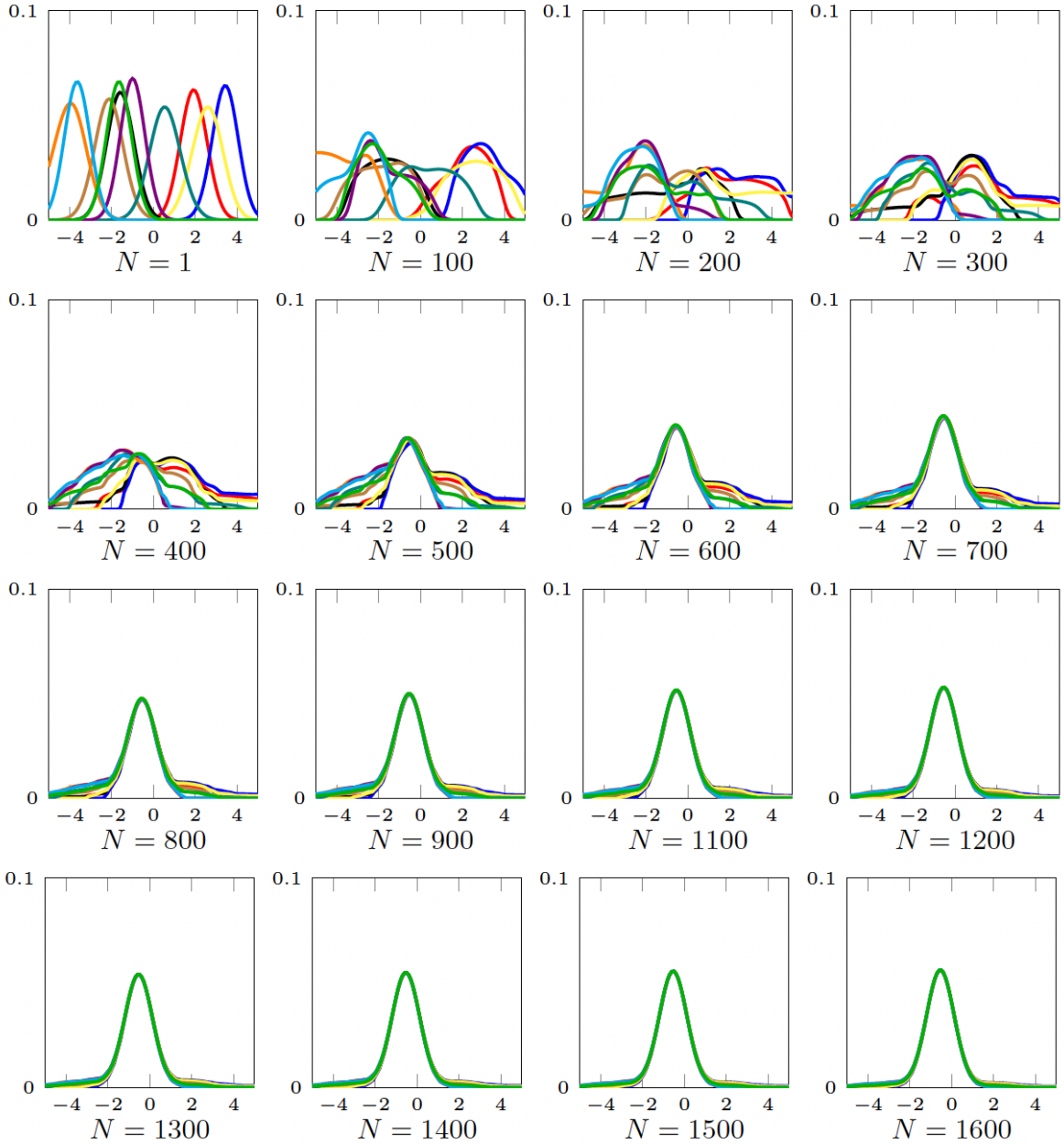


Figure 1: Dual function value and distance to consensus for 200, 100, 10, 500 agents, $M_k = 100$ and $\gamma = 0.1$.

Discrete Bounded Support Gaussian-Like



IBP: High-Accuracy Baseline

Core idea: Alternating KL projections on marginal constraints leveraging Sinkhorn kernel $K_\ell = \exp(-C_\ell/\gamma)$.

- Each outer iteration performs two Sinkhorn updates per distribution.
- Communication: vector of size n per neighbor (scalable).

Iterative Bregman Projections

$$\min_{\substack{\pi_l \mathbb{1}=p_l,\ \pi_l^{\mathsf{T}}\mathbb{1}=\pi_{l+1}^{\mathsf{T}}\mathbb{1} \\ \pi_l \in \mathbf{R}_+^{n\times n},\ l=1,\ldots,m}} \frac{1}{m}\sum_{l=1}^{m}\{\pi_l, C_l + \gamma H(\pi_l)\}$$

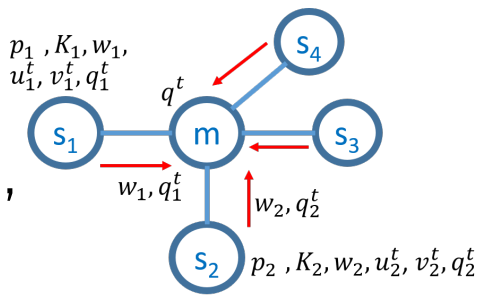
Dual problem:

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

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

IBP is equivalent to alternating minimization for the dual problem.

- $u_l^{t+1} := \ln p_l - \ln K_l e^{v_l^t}$, $\mathbf{v}^{t+1} := \mathbf{v}^t$
- $v_l^{t+1} := \frac{1}{m}\sum_{k=1}^{m} \ln K_k^{\mathsf{T}} e^{u_k^t} - \ln K_l^{\mathsf{T}} e^{u_l^t}$,
  $\mathbf{u}^{t+1} := \mathbf{u}^t$



Complexity of IBP

Theorem 5

Let $c := \max_{l=1,\dots,m} \|C_l\|_\infty$ and fix an accuracy parameter $\varepsilon' > 0$. Then Algorithm 1 (Iterative Bregman Projections) terminates after

$$N \;\le\; 4 + \frac{44\,c}{\gamma\,\varepsilon'}$$

iterations.
The couplings produced at termination, $\pi_1^{(N)}, \dots, \pi_m^{(N)}$, together with $q^{(N)} = \sum_{l=1}^m w_l\, \pi_l^{(N)} \mathbf{1}$, satisfy

$$\sum_{l=1}^m w_l \left\| \pi_l^{(N)} \mathbf{1} - q^{(N)} \right\|_1 \;\le\; \varepsilon',$$

and

$$\sum_{l=1}^m w_l \Big(\langle C_l, \pi_l^{(N)} \rangle + \gamma H\big(\pi_l^{(N)}\big) \Big) - \min_{q \in S_n(1)} \sum_{l=1}^m w_l W_\gamma(p_l, q) \;\le\; c\,\varepsilon'.$$

Barycenters and Quantization

Why Probability-Proportional-to-Size (PPS) Quantization?

- In decentralised optimisation the *communication* of dense gradient vectors dominates runtime.
- PPS converts an n-dimensional vector into an *unbiased* random sample of $M \ll n$ indices, with probabilities proportional to component magnitudes.
- Each iteration transmits only $\approx 2M \log_2 n$ bits yet preserves high-quality gradient information, matching the best known compression operators in bit-complexity while enjoying a smaller second moment when gradients lie in a simplex.
- Leveraging PPS, we derive *accelerated* primal and primal-dual methods with high-probability guarantees and near-optimal oracle and communication complexity.

Duality and Quantization

Lemma 2

Given a positive Radon probability measure $\mu \in \mathcal{M}_+^1(\mathcal{Y})$ with density $q(y)$ on a metric space $\mathcal{Y}$, the Fenchel-Legendre dual function of $\mathbf{W}_{\gamma,\mu}(p)$ can be written as

$$\mathbf{W}^*_{\gamma,\mu}(\bar{\lambda}) = \mathbf{E}_{Y\sim\mu}\left[\gamma \log\left(\frac{1}{q(Y)}\sum_{\ell=1}^{n}\exp\left(\frac{[\bar{\lambda}]_\ell - c_\ell(Y)}{\gamma}\right)\right)\right].$$

Moreover, $\mathbf{W}^*_{\gamma,\mu}(\bar{\lambda})$ has m/γ-Lipschitz gradient w.r.t. 2-norm, and its l-th coordiante, for $l = 1, \ldots, n$, is

$$[\nabla \mathbf{W}^*_{\gamma,\mu}(\bar{\lambda})]_l = \mathbf{E}_{Y\sim\mu}\left[\frac{\exp(([\bar{\lambda}]_l - c_l(Y))/\gamma)}{\sum_{\ell=1}^{n}\exp(([\lambda]_\ell - c_\ell(Y))/\gamma)}\right]. \qquad (1)$$

Sampling

Given that at iteration k each agent i can obtain $M_{i,1}^k$ independent realizations of the random variable $Y^i \sim \mu_i$:

$$\widehat{\nabla}\mathbf{W}^*_{\gamma,\mu_i}(\bar{\lambda}_i) = \frac{1}{M_{i,1}} \sum_{r=1}^{M_{i,1}} p_i(\bar{\lambda}_i, Y_r^i), \tag{2}$$

where, for all $l = 1, \cdots, n$,

$$[p_i(\bar{\lambda}_i, Y_r^i)]_l = \frac{\exp(([\bar{\lambda}_i]_l - c_l(Y_r^i))/\gamma)}{\sum_{k=1}^n \exp(([\bar{\lambda}_i]_k - c_k(Y_r^i))/\gamma)}, \tag{3}$$

IMPORTANTLY: $\widehat{\nabla}\mathbf{W}^*_{\gamma,\mu}(\bar{\lambda}) \in S_1(n)$.

Stochastic Gradient PPS Quantisation (per agent i at iter. k)

1. **Monte-Carlo OT gradient.** Draw M_1 i.i.d. samples $Y_r^i \sim \mu_i$ and form
$$\widehat{\nabla} W^*_{\gamma,\mu_i}(\boldsymbol{\lambda}_i) = \frac{1}{M_1} \sum_{r=1}^{M_1} g\big(Y_r^i; \boldsymbol{\lambda}_i\big), \quad \text{see (2.8)}.$$

2. **PPS sampling.** Let Z^i be an independent categorical r.v. with $\Pr\{Z^i = l\} = \big[\widehat{\nabla} W^*_{\gamma,\mu_i}(\boldsymbol{\lambda}_i)\big]_l$, $l = 1, \ldots, n$. Take M_2 i.i.d. draws $\{Z_r^i\}_{r=1}^{M_2}$.

3. **Quantised gradient.** Construct the histogram
$$\widetilde{\nabla} W^*_{\gamma,\mu_i}(\boldsymbol{\lambda}_i) = \frac{1}{M_2} \sum_{r=1}^{M_2} e_{Z_r^i},$$
where e_l is the l-th canonical basis vector.

This yields an *unbiased* estimator using only $M_2 \log_2 n$ bits of communication per agent at iteration k.

Unbiased estimator

Lemma 3

The full quantized stochastic gradient is unbiased, i.e., $\mathbf{E}\widetilde{\nabla}\mathbf{W}^*_\gamma(\boldsymbol{\lambda}) = \nabla\mathbf{W}^*_\gamma(\boldsymbol{\lambda})$. *Furthermore, its variance is bounded as*

$$\mathbf{E}\|\widetilde{\nabla}\mathbf{W}^*_\gamma(\boldsymbol{\lambda}) - \nabla\mathbf{W}^*_\gamma(\boldsymbol{\lambda})\|_2^2 \leq 2\lambda_{\max}(W)\sum_{i=1}^{m}\left(\frac{1}{M_{i,1}} + \frac{1}{M_{i,2}}\right).$$

Let's focus on the generic case:

$$f(x_*) = \min_{\substack{x\in\mathbb{R}^n \\ Ax=b}} f(x), \tag{4}$$

$M_1 = r$, $M_2 = M$.

Lemma 1 — Sub-Gaussianity of PPS

Lemma 4

Let $\sigma^2_{r,M} = 50\left(\frac{2(1-1/n)B^2}{\mathrm{e}M} + \frac{\sigma^2}{r}\right)$. *Then, it holds that*

$$\forall x \in \mathbb{R}^n, \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{k},\boldsymbol{l}}[\exp(\|PPS(x,\boldsymbol{\xi},\boldsymbol{k},\boldsymbol{l}) - \nabla f(x)\|^2/\sigma^2_{r,M})] \leq \mathrm{e}.$$

Theorem 1 — High-Probability Primal Error Bound

Theorem 6

For $\delta \in (0,1)$ Algorithm 1 guarantees

$$\mathbb{P}\big(f(x_T) - f(x^*) \le \varepsilon(T, \delta, \boldsymbol{\alpha}, \boldsymbol{\beta}, r, M)\big) \;>\; 1 - \delta,$$

where

$$\varepsilon = \frac{\beta_T R^2}{2A_T} + \frac{C_1' R}{A_T}\sqrt{\sum_{t=0}^{T} \alpha_t^2 \sigma_{r_t, M_t}^2} + \frac{C_2'}{A_T}\sum_{t=0}^{T} \frac{A_t\, \sigma_{r_t, M_t}^2}{\beta_t - L},$$

$C_2' = 1 + \ln\frac{4}{\delta} = \mathcal{O}(\ln\frac{1}{\delta})$ and $C_1' = \big(2J(T) + \sqrt{2} - 1\big)\big(\sqrt{2} + (\sqrt{2}+1)\sqrt{3\ln\frac{4}{\delta}}\big) + \sqrt{2} - 2 = \mathcal{O}(\mathrm{poly}(\ln T)\sqrt{\ln\frac{1}{\delta}})$.

Theorem 2 — Primal–Dual Large-Deviation Bounds

Theorem 7

For $\delta \in (0,1)$ Algorithm 2 (primal-dual) ensures

$$\mathbb{P}\big(f(x_T) - f(x^*) \le \varepsilon\big) > 1-\delta, \quad \mathbb{P}\Big(\|Ax_T - b\| \le \tfrac{\varepsilon}{R}\Big) > 1-\delta,$$

with

$$\varepsilon = \frac{\beta_T R^2}{2A_T} + \frac{C_3 R + C_4 L/\|A\|_2}{A_T}\sqrt{\sum_{t=0}^{T} \alpha_t^2 \sigma_{r_t,M_t}^2} + \frac{C_2}{A_T}\sum_{t=0}^{T} \frac{A_t\, \sigma_{r_t,M_t}^2}{\beta_t - L},$$

$$C_4 = \sqrt{2}\big(1 + \sqrt{3\ln(5/\delta)}\big), \quad C_3 = C_1 + 2\sqrt{2}J(T)\big(1 + \sqrt{3\ln(5/\delta)}\big).$$

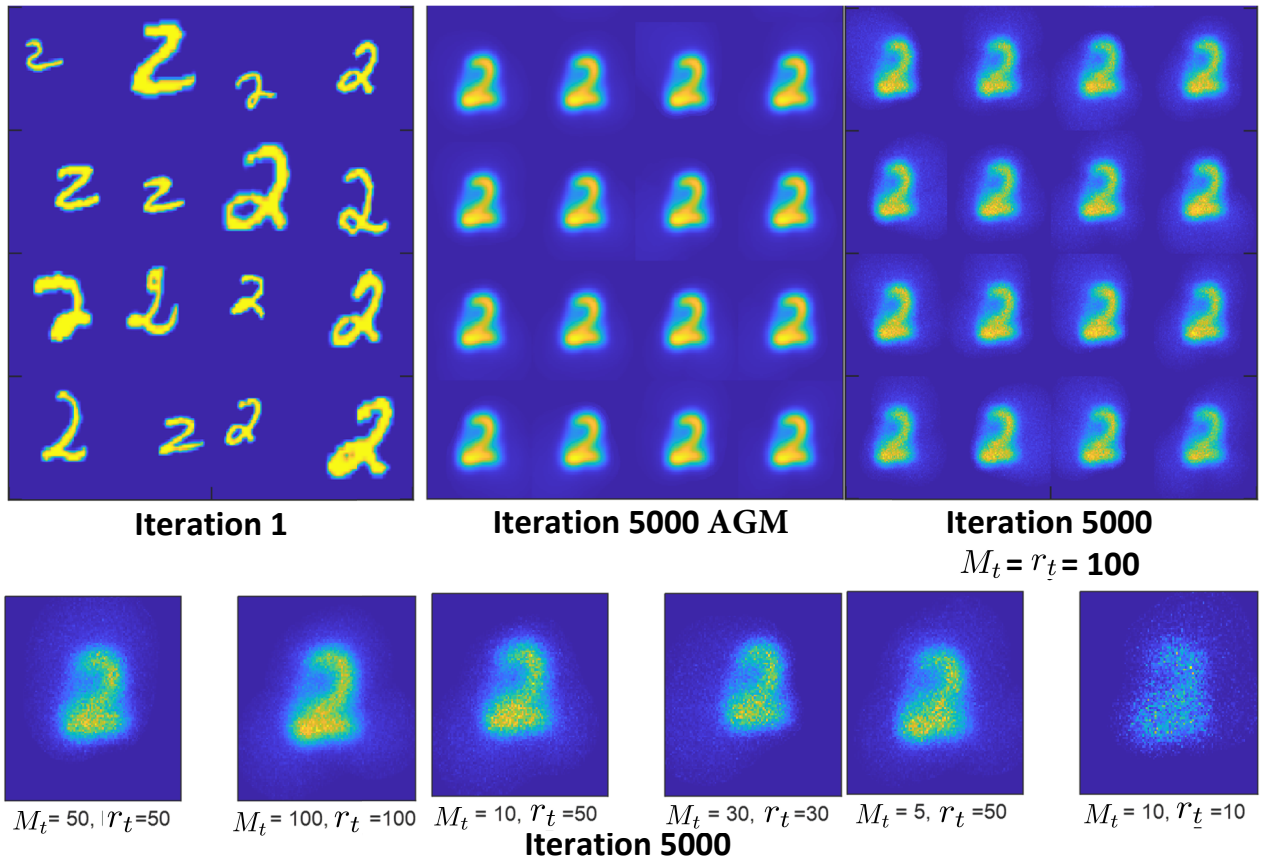
Communication Complexity

Theorem 8

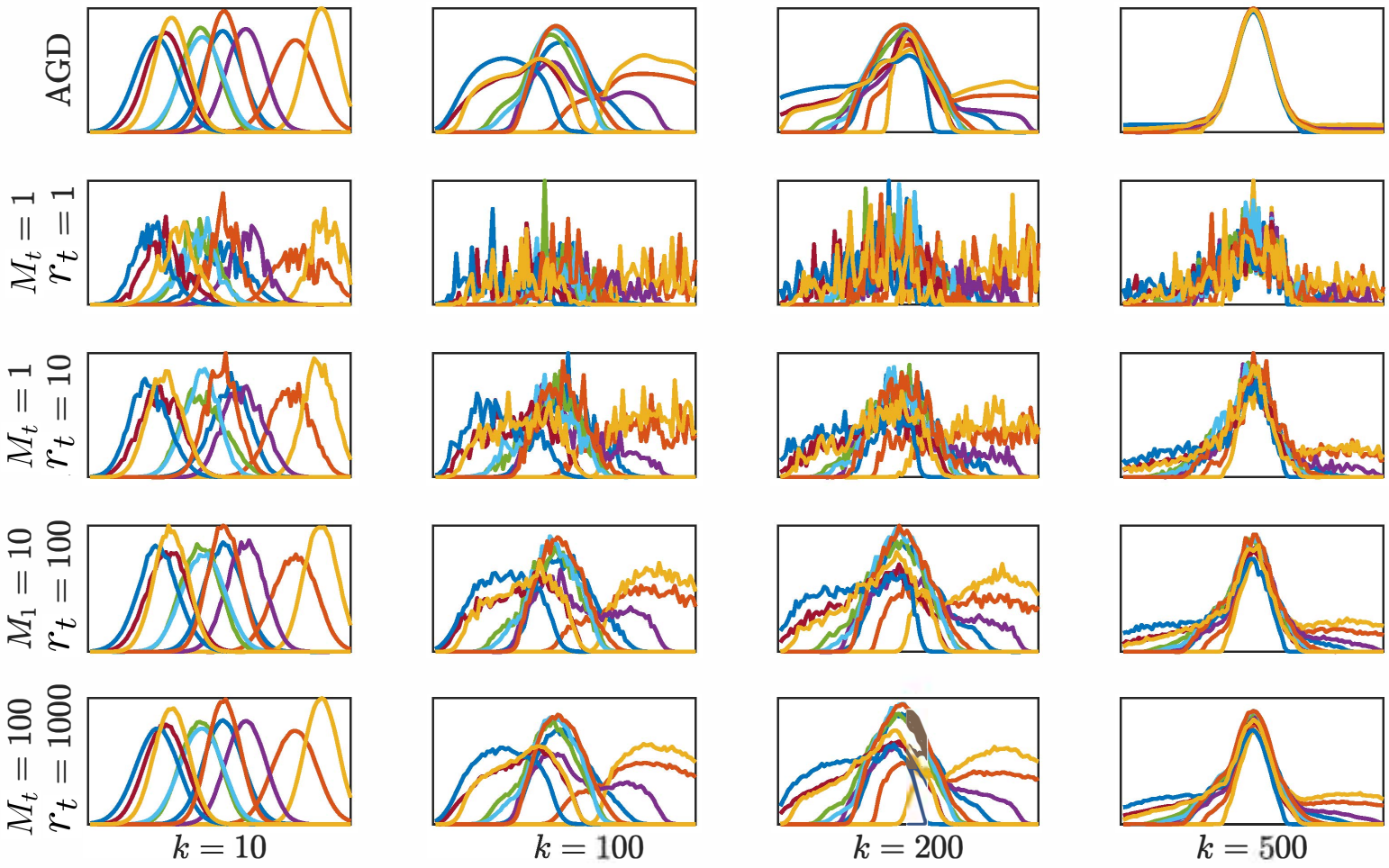
Given $\varepsilon > 0$, $0 < \delta < 1$, $\boldsymbol{r}$ and $\boldsymbol{M}$ selected appropriately, one can ensure $P(f(\boldsymbol{x}) - f(\boldsymbol{x}_*) \leq \varepsilon) > 1 - \delta$ and $P\left(\|\sqrt{\boldsymbol{W}}\boldsymbol{x}\| \leq \frac{\varepsilon\sqrt{2m}}{B_*}\right) > 1 - \delta$, if any node sends

$$\widetilde{O}\left(B^2 d \log n \max\left\{\frac{1}{\sigma_i^2}\sqrt{\frac{DB_*^2}{\gamma\varepsilon md}}, \frac{DB_*^2}{\varepsilon^2}, \frac{m^2}{\gamma^2\varepsilon^2}\right\}\right) \text{ bits.}$$

Proof by Picture



Proof by Picture



Decentralized and Equitable OT

Decentralized Optimal Transport

There is an undirected and connected network of n, where each agent has access to a column of the matrix C.

$$\min_{x_i \in \mathbf{R}_+^n} \sum_{i=1}^{n} c_i^\mathsf{T} x_i \text{ s.t. } \sum_{i=1}^{n} x_i = p \text{ and } x_i^\mathsf{T} \mathbb{1}_n = q_i \tag{5}$$

- Problem setup involves multiple agents, each with access to part of the cost information.
- Agents collaborate to design a transportation plan that minimizes the total transportation cost.
- Communication is limited to immediate neighbors in a network.

Decentralized and Equitable Optimal Transport (DE-OT)

Each agent k has its own private cost function C^k.

- Combines D-OT and EOT, focusing on decentralized networks and equitable cost sharing.
- Problem formulation:

$$\begin{aligned} &\min_{X^k \in \mathbf{R}_+^{n\times n}} \sum_{k=1}^{N} \langle C^k, X^k \rangle \\ &\text{s.t. } \langle C^k, X^k \rangle = \langle C^l, X^l \rangle \text{ for all } k, l \in [N], \\ &\left(\sum_{k=1}^{N} X^k\right) \mathbf{1}_n = p \text{ and } \left(\sum_{k=1}^{N} X^k\right)^{\mathsf{T}} \mathbf{1}_n = q. \end{aligned} \tag{6}$$

- Ensures fairness among agents.

Reformulating D-OT as a DCCO problem

Proposition 5

The DOT problem (5) *is equivalent to*

$$\min_{x_i} \sum_{i=1}^{n} c_i^{\mathsf{T}} x_i + \iota_{\geq \mathbf{0}}(x_i) \; s.t. \; \sum_{i=1}^{n} M_i x_i = \begin{bmatrix} p \\ q \end{bmatrix}, \tag{7}$$

where the indicator function $\iota_{\geq \mathbf{0}}(y)$ *is defined by*

$$\iota_{\geq \mathbf{0}}(y) \begin{cases} 0 & \text{if } y \geq 0 \\ \infty & \text{otherwise,} \end{cases} \tag{8}$$

and the matrix $M_i \in \mathbf{R}^{2n \times n}$ *is the* i^{th} *column-block of*

$$[M_1, \ldots, M_n] = M \begin{bmatrix} I_n & I_n & \cdots & I_n \\ \mathbb{1}_n^{\mathsf{T}} & \mathbf{0}_n^{\mathsf{T}} & \cdots & \mathbf{0}_n^{\mathsf{T}} \\ \mathbf{0}_n^{\mathsf{T}} & \mathbb{1}_n^{\mathsf{T}} & \cdots & \mathbf{0}_n^{\mathsf{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_n^{\mathsf{T}} & \mathbf{0}_n^{\mathsf{T}} & \cdots & \mathbb{1}_n^{\mathsf{T}} \end{bmatrix}. \tag{9}$$

Reformulating DE-OT as a DCCO problem

Proposition 6

Problem (6) *is equivalent to*

$$\min_{x_i} \sum_{i=1}^{N} c_i^{\mathsf{T}} x_i + \iota_{\geq \mathbf{0}}(x_i) \; s.t. \; \sum_{i=1}^{N} \begin{bmatrix} M \\ E_i \end{bmatrix} x_i = \begin{bmatrix} p \\ q \\ \mathbf{0}_{N-1} \end{bmatrix}, \tag{10}$$

where the indicator function $\iota_{\geq \mathbf{0}}$*, and matrices* E_i *are defined as follows:*

$$E_i \left(\tilde{L}_{*,i} \right) c_i^{\mathsf{T}} = \tilde{L} \begin{bmatrix} \mathbf{0}_{(i-1)\times n^2} \\ c_i^{\mathsf{T}} \\ \mathbf{0}_{(N-i)\times n^2} \end{bmatrix} \in \mathbf{R}^{(N-1)\times n^2}. \tag{11}$$

Common formulation

$$\begin{aligned} \min_{\boldsymbol{x}=(x_i)\in\mathbf{R}^{Nd}} \; & f(\boldsymbol{x}) \triangleq \mathbf{c}^{\mathsf{T}}\boldsymbol{x} + \iota_{\geq \mathbf{0}}(\boldsymbol{x}) = \sum_{i=1}^{N} c_i^{\mathsf{T}} x_i + \iota_{\geq \mathbf{0}}(x_i) \\ \text{s.t.} \quad & A\boldsymbol{x} = \sum_{i=1}^{N} A_i x_i = b. \end{aligned} \tag{12}$$

Table 1: Algorithms for Problem (12).

Paper	Convergence rate	Singe-loop
Falsone et al., 2020	Asymptotic	No
Chang et al. 2014,2016	$O(1/k)$, ergodic	No
Su et al. 2021	$O(1/k)$, non-ergodic	No
Alghunaim et al. 2019 and Li et al. 2023	Asymptotic	Yes

Euclidean Regularization

$$\begin{aligned} \min_{\substack{\boldsymbol{x}=(x_i)\in\mathbf{R}^{Nd},\\ \boldsymbol{y}=(y_i)\geq 0}} \quad & f(\boldsymbol{x}) \triangleq \sum_{i=1}^{N} c_i^{\mathsf{T}} x_i + \frac{\eta}{2}\|x_i\|^2 \\ \text{s.t.} \quad & A\boldsymbol{x} = \sum_{i=1}^{N} A_i x_i = b \text{ and } \boldsymbol{y} - \boldsymbol{x} = \mathbf{0}. \end{aligned} \tag{13}$$

Algorithm: PDC-ADMM

Algorithm 1 PDC-ADMM for Decentralized Optimal Transport

Penalty parameters $\rho, \beta_i, \tau_i > 0$ satisfying conditions Optimal solution x^* Initialize $k = 0$ and $x^0, y^0, z^0, \lambda^0, p^0$

while stopping criterion not met **do** Exchange λ_i^k with neighbors N_i Update y_i^{k+1}:

$$y_i^{k+1} := \left(1 - \frac{1}{\beta_i \tau_i}\right) y_i^k + \frac{1}{\beta_i \tau_i}\left(x_i^k - \tau_i z_i^k\right)$$

Update x_i^{k+1}:

$$x_i^{k+1} := x_i^k - \frac{1}{\beta_i}\left(c_i + \eta x_i^k + \frac{1}{\tau_i}\left(x_i^k - y_i^{k+1} - \tau_i z_i^k\right) + \frac{A_i^T}{2\rho|N_i|}\left(A_i x_i^{k+1} - \frac{1}{N} b - p_i^k + \rho \sum_{j \in N_i}\left(\lambda_j^k + \lambda_i^k\right)\right)\right)$$

Update λ_i^{k+1}:

$$\lambda_i^{k+1} := \frac{1}{2|N_i|}\left(\sum_{j \in N_i}\left(\lambda_j^k + \lambda_i^k\right) + \frac{1}{\rho} p_i^{k+1} + \frac{1}{\rho}\left(A_i x_i^{k+1} - \frac{1}{N} b\right)\right)$$

Update z_i^{k+1}:

$$z_i^{k+1} := z_i^k + \frac{1}{\tau_i}\left(y_i^{k+1} - x_i^{k+1}\right)$$

Update p_i^{k+1}:

$$p_i^{k+1} := p_i^k + \rho \sum_{j \in N_i}\left(\lambda_i^{k+1} - \lambda_j^{k+1}\right)$$

Increment k Output x_i^k: $x^* := \frac{1}{k}\sum_{t=1}^k x_i^t$

Theorem

Theorem 9:

Let the graph be connected and undirected. Then the sequence $(\boldsymbol{x}^k, \boldsymbol{y}^k)$ generated by PDC-ADMM with parameters $\rho, \beta_i, \tau_i > 0$ satisfying

$$\beta_i \tau_i \geq 1 \text{ and } \left(\beta_i - \frac{\beta_i}{\beta_i \tau_i - 1} - 1\right) I_d - \frac{1}{2\rho|\mathcal{N}_i|} A_i^\intercal A_i \succ \mathbf{0}, \quad (14)$$

converges to the optimal solution $(\boldsymbol{x}^*, \boldsymbol{y}^* = \boldsymbol{x}^*)$ of Problem (13). Furthermore,

$$|f(\bar{\boldsymbol{x}}^k) - f(\boldsymbol{x}^*)| + \left\|A\bar{\boldsymbol{x}}^k - b\right\| + \|\bar{\boldsymbol{y}}^k - \bar{\boldsymbol{x}}^k\| = O(1/k), \quad (15)$$

where $(\bar{\boldsymbol{x}}^k, \bar{\boldsymbol{y}}^k) = \frac{1}{k}\sum_{t=1}^k (\boldsymbol{x}^t, \boldsymbol{y}^t)$.

Numerical Results

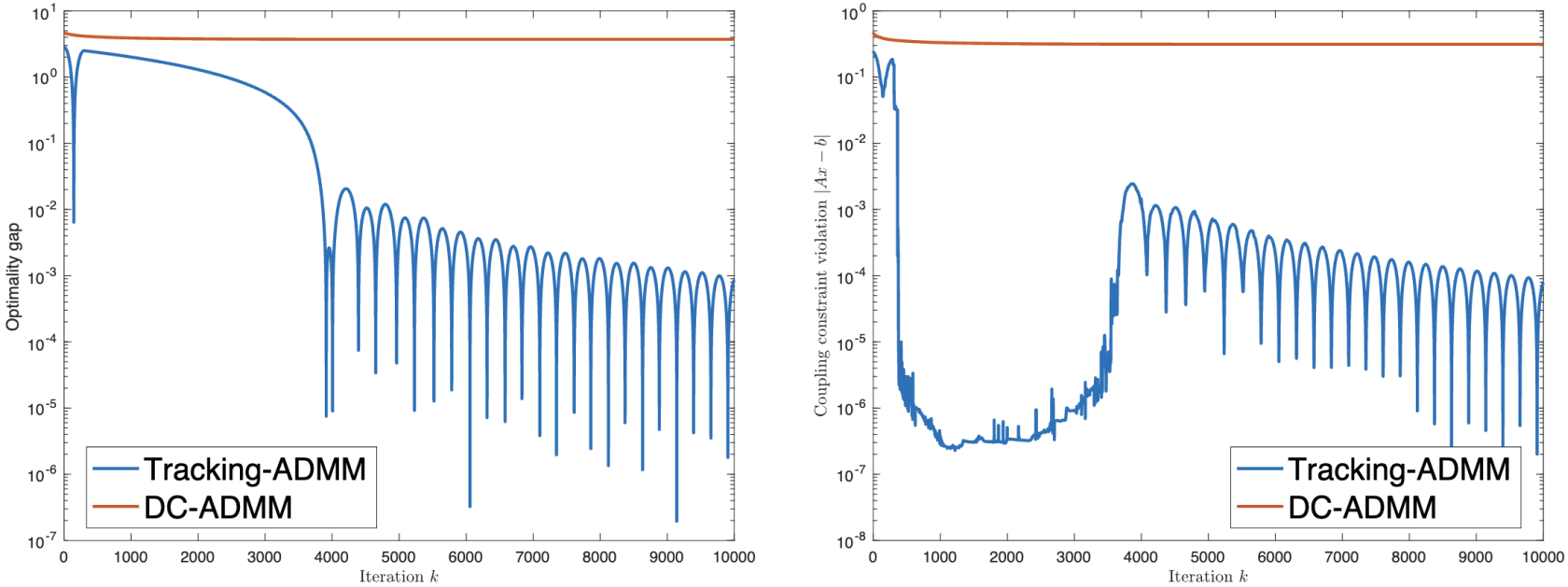


Figure 1: Optimality gap $f(\boldsymbol{x}^k) - f^*$ and feasibility violation $\|A^k - b\|$ for D-OT and DE-OT under Tracking-ADMM and DC-ADMM.

Decentralized Computation of Wasserstein Barycenters

César A. Uribe

Alexey Kroshnin Darina Dvinskikh Pavel Dvurechensky

Alexander Gasnikov Nazarii Tupitsa Roman Krawtschenko

Ivan Lau Shiqian Ma Dmitry Pasechnyuk Angelia Nedić

May 19, 2025

RICE ENGINEERING

Electrical and Computer Engineering

References I

[1] A. Kroshnin, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, N. Tupitsa, C. Uribe. On the complexity of approximating Wasserstein barycenter. 2020.

[2] P. Dvurechensky, D. Dvinskikh, A. Gasnikov, C. Uribe, A. Nedić. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In *NeurIPS*, 2018.

[3] C. Uribe, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, A. Nedić. Distributed computation of Wasserstein barycenters over networks. In *IEEE CDC*, 2018.

[4] R. Krawtschenko, C. Uribe, A. Gasnikov, P. Dvurechensky. Distributed optimization with quantization for computing Wasserstein barycenters. 2020.

[5] D. Pasechnyuk, P. Dvurechensky, C. Uribe, A. Gasnikov. Decentralised convex optimisation with probability-proportional-to-size quantization. 2025.

[6] I. Lau, S. Ma, C. Uribe. Decentralized and equitable optimal transport. 2024.