# First order methods and rates for approximate optimal transport and Wasserstein barycenter problems

Antonin Chambolle

CEREMADE, CNRS et
Université Paris-Dauphine (PSL), France

(Joint with Juan-Pablo Contreras, U. Adolfo Ibañez, Chile)

Inverse Problem on Large Scales,
RICAM, Linz, 28/11–2/12 2022

# Outline

- first order algorithms for approximate OT and WB?
- some properties of OT solutions and approximate solutions;
- Euclidean and nonlinear saddle-point algorithms;
- basic complexity bounds;
- improvements: acceleration, linesearch;
- extensions, examples.

# (Discrete) Optimal transportation problem (OT)

*Data:* distributions $(\mu_i)_{i=1,\dots,n}$ $(\nu_j)_{j=1,\dots n}$ (to simplify), with $\mu_i \ge 0$, $\nu_j \ge 0$, $\sum_i \mu_i = \sum_j \nu_j = 1$;
a cost matrix $(C_{i,j})_{i,j}$, with (wlog) $C_{i,j} \ge 0$.

*Problem:* minimal cost assignment (or transportation) from $\mu$ to $\nu$ (a minimal cost flow problem).

$$\min_{X \ge 0} C : X =: \sum_{i,j} C_{i,j} X_{i,j} \ : \ \sum_j X_{i,j} = (X\mathbf{1}_n)_i = \mu_i \, , \ \sum_i X_{i,j} = (X^T\mathbf{1}_n)_i = \nu_j \tag{OT}$$

*(in particular $\sum_{i,j} X_{i,j} = 1$).*

We denote $\Delta_n$ the unit simplex in $\mathbb{R}^n$, $\Delta_{n\times n}$ the unit simplex in $\mathbb{R}^{n\times n}$,

$$\Delta_{\mu,\nu} := \left\{ (x_{i,j}) \in \mathbb{R}_+^{n\times n} : \sum_j x_{i,j} = \mu_j, \sum_i x_{i,j} = \nu_j \right\} \subset \Delta_{n\times n}.$$

# (Discrete) Wasserstein barycenter problem (WB)

An extension is the discrete transportation barycenter problem: given $(\mu^l)$, $l = 1, \ldots, m$ in $\Delta_1$, we look for the "barycenter" $\nu$ of the measures, given the cost matrices $C^l$, and the scalar weights $w^l \geq 0$ with $\sum_{l=1}^m w^l = 1$, solving:

$$\min_{\nu \in \Delta_n} \min_{X^l \in \Delta_{\mu_l, \nu}} \sum_{l=1}^{m} w^l C^l : X^l . \qquad (WB)$$

Here, $\nu$ is the *common* second marginal of the transportation plans $(X^l)_l$. For $C^l = C$ given by $C_{i,j} = |x_i - x_j|^2$, $(x_i)_{i=1}^n$ a sampling of some domain in $\mathbb{R}^d$, $\nu$ will be an approximation of the (2-)Wasserstein barycenter of the $(\mu^l)_l$ with weights $(w^l)_l$.

# Our goal

- We want to study non-linear continuous optimization algorithms for *approximate* (OT) or (WB);
- Why? linear programming works very well (network simplex implemented in python-OT);

- Theoretical complexity scales a bit better ($\sim n^{5/2}$ rather than $n^3$ or $n^4$);
- Efficient LP for (WB)?
- Straightforward extension to nonlinear problems such as:

$$\min_{X : X\mathbf{1}_n = \mu} C : X + \psi(X^T \mathbf{1}_n)$$

for $\psi$ a convex function.

# Classical trick for approximate OT: entropic regularization

- Replace $X \geq 0$ by the entropic barrier $\gamma \sum_{i,j} X_{i,j} \ln X_{i,j} = \gamma X : (\ln X)$, $\gamma > 0$; [Cuturi 2013]
- Allows for explicit solution for one fixed marginal ($X\mathbf{1}_n = \mu$ **or** $X^T\mathbf{1}_n = \nu$);
- Alternating maximization for the dual / alternating "Bregman" projection in the primal on each marginal leads to the Sinkhorn algorithm [Sinkhorn, S-Knopp, 64–67];
- Very efficient for large $\gamma$ ($\rightarrow$ large error), hard to implement and slow for small $\gamma$ (involves $exp(-C/\gamma)$).

# Approximate OT: rates

Many recent works have addressed the complexity of solving the OT up to some error: given $\varepsilon > 0$, one looks for $X$ admissible with $C : X \leq C : X^* + \varepsilon$. First order / randomized / alternating minimization approaches. Here $\|C\| = \max_{i,j} |C_{i,j}|$.

- Sinkhorn: $O(n^2\|C\|^2/\varepsilon^2)$ (up to log factors) [Dvurechensky Gasnikov Kroshnin 18]. Randomized "Randkhorn" is $O(n^{7/3}(\|C\|/\varepsilon)^{4/3})$ [Lin-Ho-Chen-Cuturi-Jordan 2020];
- Accelerated first order methods: $O(n^{5/2}\|C\|/\varepsilon)$ (a bit worse wr $n$, better wr $\varepsilon$) [DGK18], [Lin Ho Jordan 2019];
- [Sherman 2017] "Area convexity": non-linear (Bregman type) descent with a non-convex but "area convex" Bregman function: $O(n^2\|C\|/\varepsilon)$ in theory, very slow in practice;
- [Blanchet-Kent-Jambulapati-Sidford 2020] : $O(n^2\|C\|/\varepsilon)$ using linear programming techniques (for "packing") / interior point type (Newton/matrix scaling) (Implementation?) + This is optimal.

# Approximate OT: rates

**Our contribution:** we show that standard saddle-point (that is, Prox method of [Nemirovsky 2004] or non-linear primal-dual [C-Pock 2016]) yield the nearly optimal rate $O(n^{5/2}\|C\|/\varepsilon)$, and that heuristic improvements (line-search, [Malitsky-Pock 2018]) yield competitive methods wr the state-of-the art.

- Would need to be compared with implementation of [Blanchet et al. 2020];
- Not competitive with Network Simplex for middle-sized OT problems.
- Yet quite better than LP based methods for barycenter problems. Generalizes easily to nonlinear.

# Some basic facts about OT

1. **Duality:**

$$\min_{X\in\Delta_{\mu,\nu}} C:X = \min_{X\geq 0}\max_{f,g} C:X + f\cdot(\mu - X\mathbf{1}_n) + g\cdot(\nu - X^T\mathbf{1}_n)$$
$$= \max_{f,g}\min_{X\geq 0} f\cdot\mu + g\cdot\nu + X:(C - f\otimes\mathbf{1}_n - \mathbf{1}_n\otimes g)$$
$$= \max_{f,g}\{f\cdot\mu + g\cdot\nu \ : \ f\otimes\mathbf{1}_n + \mathbf{1}_n\otimes g \leq C\}\,.$$

The Lagrangian:

$$\mathcal{L}(X,f,g) := C:X + f\cdot(\mu - X\mathbf{1}_n) + g\cdot(\nu - X^T\mathbf{1}_n)$$

(*cf* Monge / Kantorovich / Rubinstein in the continuous setting.)

# Some basic facts about OT

**2. Bounds:** Here we assume (wlog): $C_{i,j} \geq 0$, $\min_i C_{i,j} = \min_j C_{i,j} = 0$.
Why? because $(C_{i,j} + a)_{i,j}$, $a \in \mathbb{R}$, $(C_{i,j} + a_i)_{i,j}$, $a \in \mathbb{R}^n$, $(C_{i,j} + b_j)_{i,j}$, $b \in \mathbb{R}^n$ yield the same solutions. (Indeed:
$(C + a \otimes \mathbf{1}_n) : X = C : X + a \cdot (X\mathbf{1}_n) = C : X + a \cdot \mu$, etc.)
We also assume $\mu_i, \nu_j > 0$ (else we can remove the corresponding coordinate).

Basic remark: $(X, f, g)$ solution (saddle-point of $\mathcal{L}$) $\rightarrow (X, (f_i + a)_i, (g_j - a)_j)$ solution. As a consequence:

**Lemma**: There is a saddle-point with $|f_i|, |g_j| \leq \|C\|/2$.
(Again $\|C\| = \max_{i,j} C_{i,j}$. This is sharp.)

# Some basic facts about OT

**2. Bounds:** Here we assume (wlog): $C_{i,j} \geq 0$, $\min_i C_{i,j} = \min_j C_{i,j} = 0$.
Why? because $(C_{i,j} + a)_{i,j}$, $a \in \mathbb{R}$, $(C_{i,j} + a_i)_{i,j}$, $a \in \mathbb{R}^n$, $(C_{i,j} + b_j)_{i,j}$, $b \in \mathbb{R}^n$ yield the same solutions. (Indeed:
$(C + a \otimes \mathbf{1}_n) : X = C : X + a \cdot (X\mathbf{1}_n) = C : X + a \cdot \mu$, etc.)
We also assume $\mu_i, \nu_j > 0$ (else we can remove the corresponding coordinate).

Basic remark: $(X, f, g)$ solution (saddle-point of $\mathcal{L}$) $\to (X, (f_i + a)_i, (g_j - a)_j)$ solution. As a consequence:

**Lemma**: There is a saddle-point with $|f_i|, |g_j| \leq \|C\|/2$.
(Again $\|C\| = \max_{i,j} C_{i,j}$. This is sharp.)

*Proof:* Relies on complementary conditions. Assume wlog $f_i \geq 0$, $\min_i f_i = 0$ ($f_i \leftarrow f_i - \min_{i'} f_{i'}$)
Complementary shows: $X_{i,j} > 0 \Rightarrow f_i + g_j = C_{i,j}$.
Then $f_i + g_j \leq C_{i,j} \Rightarrow g_j \leq \min_i C_{i,j} - f_i \leq 0$ (as $\min_j C_{i,j} = 0$). Using then that $\min_i f_i = 0$ and that for all $i$ ($j$), $\exists j$ ($i$) with $f_i + g_j = C_{i,j}$ (since $\sum_i X_{i,j} > 0$, $\sum_j X_{i,j} > 0$), we easily deduce that there is $i_0, j_0$ with $f_{i_0} = g_{i_0} = C_{i_0,j_0} = 0$ and then:

$$0 \leq f_i \leq \|C\|, \quad -\|C\| \leq g_j \leq 0.$$

Then $(f_i - \|C\|/2, g_j + \|C\|/2)$ satisfies the thesis of the Lemma.

# A consequence

The problem is equivalent to

$$\min_{X\geq 0} \max_{|f_i|,|g_j|\leq\lambda} \mathcal{L}(X,f,g)$$
$$= \min_{X\geq 0} C : X + \lambda|\mu - X\mathbf{1}_n|_1 + \lambda|\nu - X^T\mathbf{1}_n|_1$$

as soon as $\lambda \geq \|C\|/2$. We solve the saddle-point with a primal-dual method.

# Primal-dual algorithm

Recall: $\mathcal{L}(X, f, g) = C : X + f \cdot (\mu - X\mathbf{1}_n) + g \cdot (\nu - X^T\mathbf{1}_n).$

$$\begin{cases} f^{k+1} = \arg\max_{|f|\leq\lambda} -\frac{1}{2\sigma}\|f - f^k\|^2 + f \cdot (\mu - X^k\mathbf{1}_n) = \Pi_{[-\lambda,\lambda]}(f^k + \sigma(\mu - X^k\mathbf{1}_n)) \\ g^{k+1} = \arg\max_{|g|\leq\lambda} -\frac{1}{2\sigma}\|g - g^k\|^2 + g \cdot (\nu - (X^k)^T\mathbf{1}_n) \\ \tilde{f}^{k+1} = 2f^{k+1} - f^k, \quad \tilde{g}^{k+1} = 2g^{k+1} - g^k, \\ X^{k+1} = \arg\min_{X\geq 0} \frac{1}{\tau} D_X(X, X^k) + X : (C - \tilde{f}^{k+1} \otimes \mathbf{1}_n - \mathbf{1}_n \otimes \tilde{g}^{k+1}). \end{cases}$$

with $D_X(X, X^k)$ a "Bregman distance[1]", such as $\|X - X^k\|^2/2$ (in this case $X^{k+1} = (X^k - \tau(C - \tilde{f}^{k+1} \otimes \mathbf{1}_n - \mathbf{1}_n \otimes \tilde{g}^{k+1}))^+$ is also easy to compute).

[1] $D_X(X, X^k) := \psi(X) - \psi(X^k) - \nabla\psi(X^k) \cdot (X - X^k)$ for $\psi$ some convex function with domain $\mathbb{R}_+^{n\times n}$ or $\Delta_{n\times n}$

# Primal-dual algorithm: basic estimates

Letting $\bar{X}^k = (1/k)\sum_{i=1}^k X^i$, etc, we have the following [C-Pock, 2016]: for all $X, f, g$,

$$\mathcal{L}(\bar{X}^k, f, g) - \mathcal{L}(X, \bar{f}^k, \bar{g}^k) \leq \frac{2}{k}\left(\frac{1}{\tau}D_X(X, X^0) + \frac{\|f - f^0\|^2 + \|g - g^0\|^2}{2\sigma}\right)$$

And introducing the primal-dual gap (primal - dual values)

$$\mathcal{G}(\bar{X}, \bar{f}, \bar{g}) := \max_{|f|\leq\lambda, |g|\leq\lambda, X\in\Delta_{n\times n}} \mathcal{L}(\bar{X}, f, g) - \mathcal{L}(X, \bar{f}, \bar{g})$$

one gets (choosing $f^0 = g^0 = 0$):

$$\mathcal{G}(\bar{X}^k, \bar{f}^k, \bar{g}^k) \leq \frac{2}{k}\left(\frac{1}{\tau}\max_X D_X(X, X^0) + \frac{n\lambda^2}{\sigma}\right).$$

# Global rate?

A crucial point: this rate holds under restrictive assumptions on $\tau, \sigma$. Namely:

$$\tau\sigma L^2 \leq 1 \text{ where } L := \max_{\|X\|_{\mathcal{X}} \leq 1} \max_{\|(f,g)\|_{\mathcal{Y}} \leq 1} X : (f \otimes \mathbf{1}_n + \mathbf{1}_n \otimes g).$$

Here, the choices of the norms in $\mathcal{X} \ni X, \mathcal{Y} \ni (f, g)$ are important. For $\mathcal{Y}$, we use $\| \cdot \|_2$ the Euclidean norm.
For $\mathcal{X}$, we need the Bregman function $\psi$ from which $D_X$ is obtained:

$$D_X(X, X') := \psi(X) - \psi(X') - \nabla\psi(X') \cdot (X - X')$$

to be 1-convex: $D_X(X, X') \geq \|X - X'\|_{\mathcal{X}}^2 / 2$.

# Global rate?

For $\psi(X) = \|X\|_2^2/2$ (Euclidean), one has

$$L = \max_{\sum_{i,j} X_{i,j}^2 \leq 1} \max_{\sum_i f_i^2 + g_i^2 \leq 1} \sum_{i,j} X_{i,j}(f_i + g_j) = \max_{\sum_i f_i^2 + g_i^2 \leq 1} \sqrt{\sum_{i,j} (f_i + g_j)^2} = \sqrt{2n}$$

Hence one can choose $\tau = 1/(2n\sigma)$ and one gets a rate:

$$\frac{2}{k}\left(\frac{1}{\tau} + \frac{n\lambda^2}{\sigma}\right) = \frac{2}{k}\left(2n\sigma + \frac{n\lambda^2}{\sigma}\right) \overset{\min_\sigma}{\rightarrow} \frac{4\sqrt{2}n\lambda}{k}$$

Hence one needs $\sim \lambda n/\varepsilon$ iterations (and $\lambda n^3/\varepsilon$ computations) to reach a precision $\varepsilon$ (using the optimal steps). Same as Network simplex, but no sparsity, and very slow in practice.

# Improvement by non-linear optimization

To improve the rate we use $\psi(X) = X \cdot \ln X = \sum_{i,j} X_{i,j} \ln X_{i,j}$ if $X \in \Delta_{n\times n}$, and $+\infty$ else, and non-linear proximal updates:
$D_\psi(X, X') = \sum_{i,j} X_{i,j} \ln(X_{i,j}/X'_{i,j})$ is the KL divergence. Then $\psi$ is $1$-strongly convex on the simplex, wr the $\ell_1$ norm (*cf* Pinsker's inequality).
Hence, the right norm for $X$ is $\ell_1$ and

$$L = \max_{\sum_{i,j}|X_{i,j}|\leq 1} \max_{\sum_i f_i^2+g_i^2\leq 1} \sum_{i,j} X_{i,j}(f_i + g_j) = \max_{\sum_i f_i^2+g_i^2\leq 1} \max_{i,j} |f_i + g_j| = \sqrt{2}$$

$\rightarrow$ improvement by a factor $\sqrt{n}$ (choosing again the optimal $\tau, \sigma$), but we lose a factor $\log n$ ("diameter" of the unit simplex in the KL divergence).

# Improvement by non-linear optimization

- The estimate on the gap has to be turned into an estimate for an approximate feasible point. This is obtained by a rounding procedure (Altschuller, Niles-Weed, Rigollet 2017) (for which we slightly improved the constant);
- Same complexity as the most recent approaches based on first order methods (except "area convexity" / [Blanchet et al]): $n^{5/2}\|C\|/\varepsilon$ $(\times \ln n)$;
- Nonlinear updates are easily performed exactly (similar to Sinkhorn-type update);
- Sinkhorn-type update: one can enforce $X\mathbf{1}_n = \mu$ (or $X^T\mathbf{1}_n = \nu$) at each iteration and drop the corresponding dual variable (simpler, and slightly faster);
- $\varepsilon$ needs not be fixed in advance (may use other stopping criterion);
- Not as fast as best methods such as [Dvurechensky et al, 18].
- Generalizes to WB problem which has the same structure.

# Further improvements? Acceleration, line-search

**Acceleration:** One can smooth the problem (as for Sinkhorn), as also proposed by [Dvurechensky et al, 18], by adding $\gamma X \cdot \ln X = \gamma\psi(X)$ ($\to$ $\gamma$-convex in $\ell_1$):

$$\mathcal{L}_\gamma(X, f, g) = \mathcal{L}(X, f, g) + \gamma X \cdot \ln X\,.$$

Dvurechensky et al. propose then to compute the dual (which has then Lipschitz gradient in $(\ell_1, \ell_\infty)$ and use an accelerated gradient scheme inspired by Nesterov's/Tseng's accelerated methods.

On the other hand, the primal objective becomes "relatively strongly convex" wr to $\psi(X) = \gamma X \ln X$ [Lu, Freund, Nesterov 18], that is, $\mathcal{L}_\gamma(\cdot, f, g) - \gamma\psi$ is convex (for all $(f, g)$), and one can revert to an accelerated method as shown in [C-Pock 16].
The rate of convergence is now $O(1/(\gamma k^2))$ (with essentially the same constants), however the global complexity is unchanged, as one needs to choose $\gamma \sim \varepsilon$ (and then $k \sim 1/\varepsilon$) to maintain an error of order $\varepsilon$.

# Improvements? Acceleration, line-search

**Linesearch:** [Malitsky and Pock 2018] introduce a primal-dual algorithm with linesearch in the Euclidean case. It was observed in [Jiang-Vandenberghe 2022] that it could be extended to the case where one variable has a non-linear prox function, as in our case.
We extend this result to the (relatively) strongly convex case, improving in fact both settings from [Malitsky-Pock] and [Jiang-VdB].
The theoretical rate is the same as before, and the complexity is not changed.
But the empirical convergence is improved.

(Numerics)

# Wasserstein barycenter

Similarly to OT, we can solve the barycenter problem with the saddle-point formulation:

$$\min_{X^l \in \Delta_{n\times n}, l=1,\ldots,m} \max_{|f^l|,|g^l|\le\lambda} \sum_{l=1}^{m} w_l \left( C^l : X^l + f^l \cdot (\mu^l - X^l \mathbf{1}_n) + g^l \cdot ((X^m - X^l)^T \mathbf{1}_n) \right) \left[ + \gamma \sum_{l=1}^{m} w_l X^l \cdot \ln X^l \right].$$

$\rightarrow$ one can adapt the same algorithms. One can also remove the variables $f^l$ and solve the $X$ problems directly with the constraint $X^l \mathbf{1}_n = \mu^l$.

# Remark: scaled entropy kernel

We propose to replace the entropy $\psi$ by, for $\delta > 0$ small:

$$\psi_\delta(X) = \frac{1}{(1-\delta)^2}\psi\left(\frac{\delta}{n^2}\mathbf{1}_n \otimes \mathbf{1}_n + (1-\delta)X\right)$$

for $X \geq 0$, and $+\infty$ else, which is still $1$-convex on the unit simplex (for the $\ell_1$ norm).
The idea is that $\nabla\psi_\delta(X)$ is now finite when some $X_{i,j} = 0$, so this does not rule out sparse solutions.

# Remark: scaled entropy kernel

We propose to replace the entropy $\psi$ by, for $\delta > 0$ small:

$$\psi_\delta(X) = \frac{1}{(1-\delta)^2}\psi\left(\frac{\delta}{n^2}\mathbf{1}_n \otimes \mathbf{1}_n + (1-\delta)X\right)$$

for $X \geq 0$, and $+\infty$ else, which is still $1$-convex on the unit simplex (for the $\ell_1$ norm).
The idea is that $\nabla\psi_\delta(X)$ is now finite when some $X_{i,j} = 0$, so this does not rule out sparse solutions.

**Remark 1:** this is stupid since this kernel does not act as a barrier any longer for the constraint $X \geq 0$;

# Remark: scaled entropy kernel

We propose to replace the entropy $\psi$ by, for $\delta > 0$ small:

$$\psi_\delta(X) = \frac{1}{(1-\delta)^2}\psi\left(\frac{\delta}{n^2}\mathbf{1}_n \otimes \mathbf{1}_n + (1-\delta)X\right)$$

for $X \geq 0$, and $+\infty$ else, which is still $1$-convex on the unit simplex (for the $\ell_1$ norm).
The idea is that $\nabla\psi_\delta(X)$ is now finite when some $X_{i,j} = 0$, so this does not rule out sparse solutions.

**Remark 1:** this is stupid since this kernel does not act as a barrier any longer for the constraint $X \geq 0$;

**Remark 2:** this is not totally stupid as one still may solve the corresponding "prox" efficiently.

# Remark: scaled entropy kernel

Letting $X^\delta := \frac{\delta}{n^2}\mathbf{1}_n \otimes \mathbf{1}_n + (1-\delta)X$ the corresponding prox is solved by computing:

$$\min_{X^\delta \geq \delta/n^2} Y : X^\delta + \frac{1}{\tau(1-\delta)} D_X(X^\delta, \bar{X}^\delta).$$

Optimality conditions are:

$$Y_{i,j} + \frac{1}{\tau(1-\delta)}(\log X^\delta_{i,j} - \log \bar{X}^\delta_{i,j}) + \alpha_{i,j} = \beta$$

with $\alpha_{i,j} > 0$ only when $X^\delta_{i,j} = \delta/n^2$ and $\beta$ the Lagrange multiplier for the constraint $\sum X^\delta_{i,j} = 1 \to X_{i,j} = \bar{X}^\delta_{i,j} \exp(-\tau(1-\delta)Y_{i,j})e^{-\beta}$ or $\delta/n^2$.

# Remark: scaled entropy kernel

One shows (from optimality) that there exists $s > 0$ such that

$$X_{i,j}^{\delta} = \max\left\{\frac{1}{s}\bar{X}_{i,j}^{\delta}\exp(-\tau(1-\delta)\,Y_{i,j}),\, \frac{\delta}{n^2}\right\}.$$

# Remark: scaled entropy kernel

One shows (from optimality) that there exists $s > 0$ such that

$$sX_{i,j}^{\delta} = \max\left\{\bar{X}_{i,j}^{\delta} \exp(-\tau(1-\delta)Y_{i,j}),\, s\frac{\delta}{n^2}\right\}.$$

# Remark: scaled entropy kernel

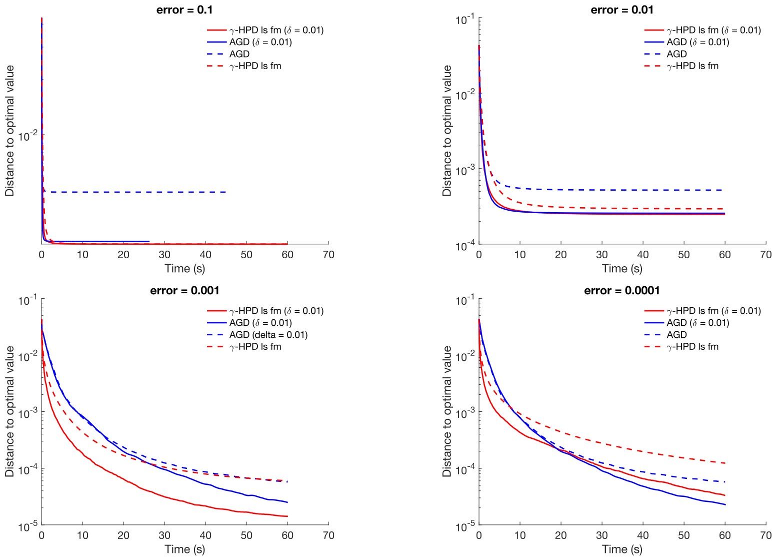
One shows (from optimality) that there exists $s > 0$ such that

$$s = \sum_{i,j} \max \left\{ \bar{X}^{\delta}_{i,j} \exp(-\tau(1-\delta) Y_{i,j}), s \frac{\delta}{n^2} \right\}.$$

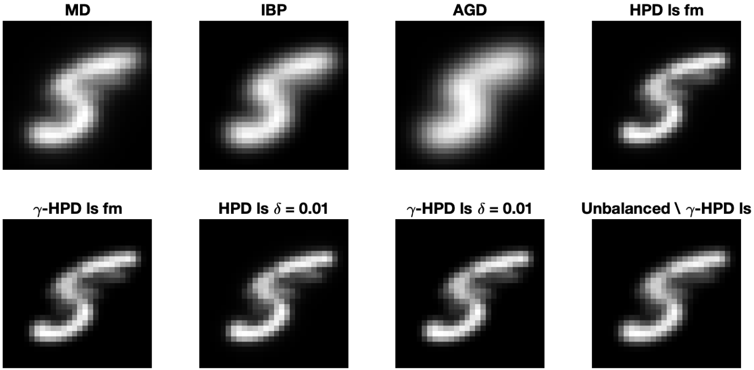
Letting $Z_{i,j} := \bar{X}^{\delta}_{i,j} \exp(-\tau(1-\delta) Y_{i,j})$, one needs to solve $s = T(s)$ where $T(s) = \sum_{i,j} \max\{Z_{i,j}, s\delta/n^2\}$ is $\delta$-Lipschitz: very contractive if $\delta$ is small.

Alternatively, we can use Newton's method to solve $s - Ts = 0$.

# Some Results:



# Barycenter problems




(Barycenters computed via various algorithms)

# To do?

- nonlinear problems? (Wasserstein flows?)
- faster matrix/vector products for $W_2^2$ (convolutions)?
- Exploit sparsity ($\mathrm{spt} X^* \leq 2n - 1$)
  (*cf* Network simplex, or sparse interior point method [Zanetti-Gondzio 2022])

Thank you for your attention.