Weierstrass Institute for

Applied Analysis and Stochastics

Computational Optimal Transport: Accelerated Gradient Descent vs Sinkhorn

Pavel Dvurechensky (joint work with Alexander Gasnikov and Alexey Kroshnin)

Grenoble Optimization Days 2018

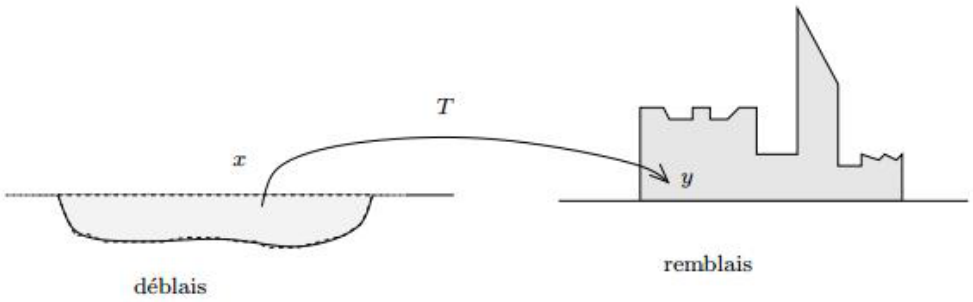
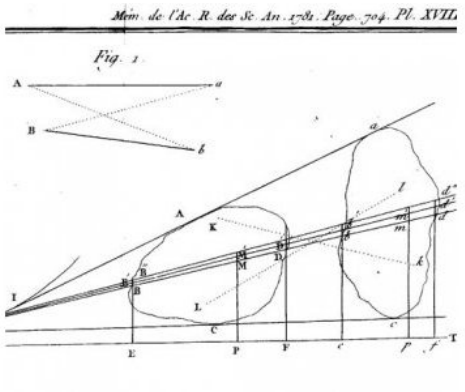


Fig. 3.1. Monge's problem of déblais and remblais



- (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.

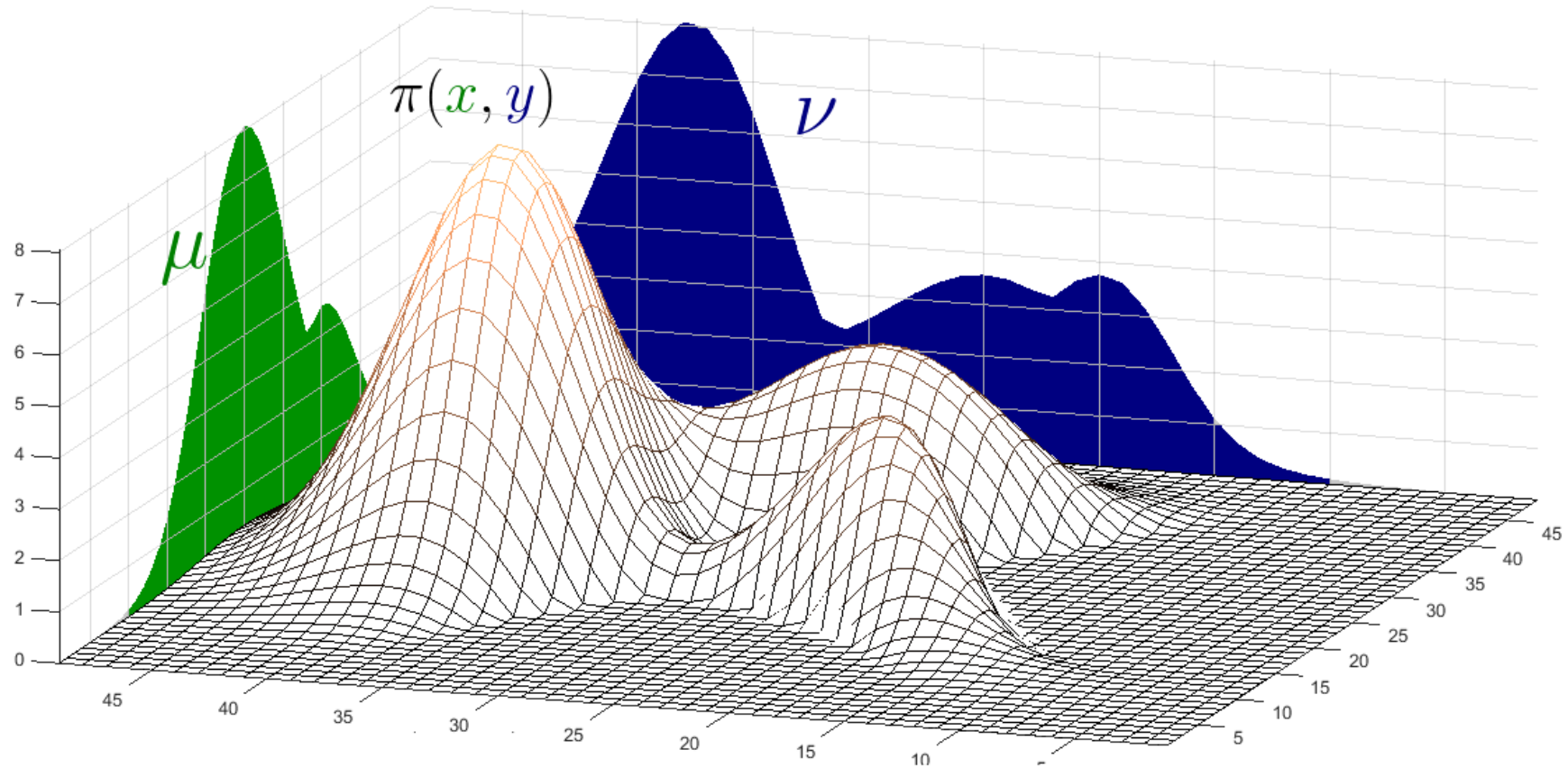
- (E, D) – metric space;
- $C(x, y)$ – cost function, e.g. $C(x, y) = D(x, y)$;
- $\mu, \nu \in \mathcal{P}_2$ – measures to be transported;
- $\mathcal{U}(\mu, \nu)$ – set of all probability measures on $E \times E$ with marginals μ, ν.

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

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

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

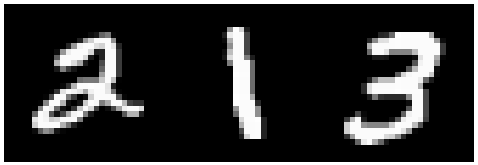
$$W_2^2(\mu, \nu) = \inf_{\pi \in \mathcal{U}(\mu,\nu)} \int_{E\times E} \|x - y\|_2^2 d\pi(x, y).$$

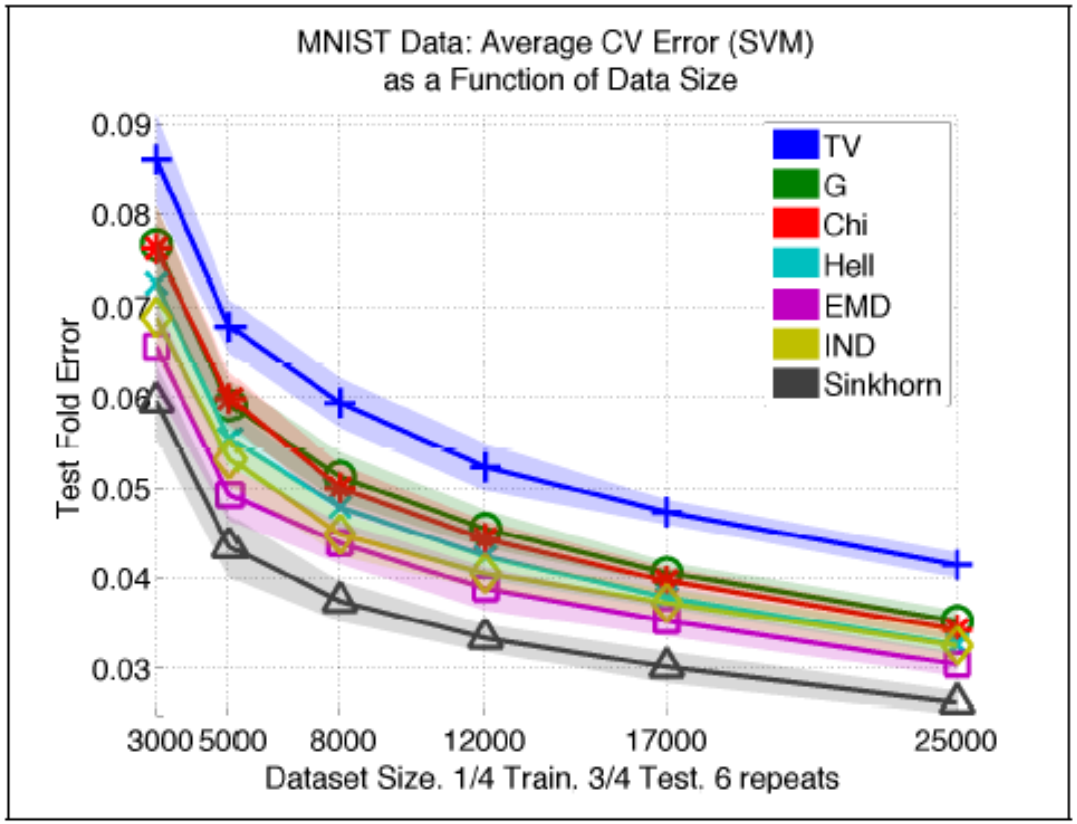


$$\inf_{\pi \in \mathcal{U}(\mu,\nu)} \int_{E \times E} C(x, y) d\pi(x, y).$$

Image: A.Suvorikova

Euclidean distance between pixels defines a distance between images as the minimum amount of work to transport one image to another.





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

- $\xi_i \in \mathbb{R}^d$, $i = 1, ..., n$ – support of μ, $\eta_i \in \mathbb{R}^d$, $i = 1, ..., n$ – support of ν;
- $\mu = \sum_{i=1}^n r_i \delta(\xi_i)$, $\nu = \sum_{j=1}^n c_j \delta(\eta_j)$;
- $C_{ij} = C(\xi_i, \eta_j)$, $i, j = 1, ..., n$ – ground cost matrix;
- $X_{ij} = \pi(\xi_i, \eta_j)$, $i, j = 1, ..., n$ – transportation plan;

Optimal transport problem

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

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

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

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

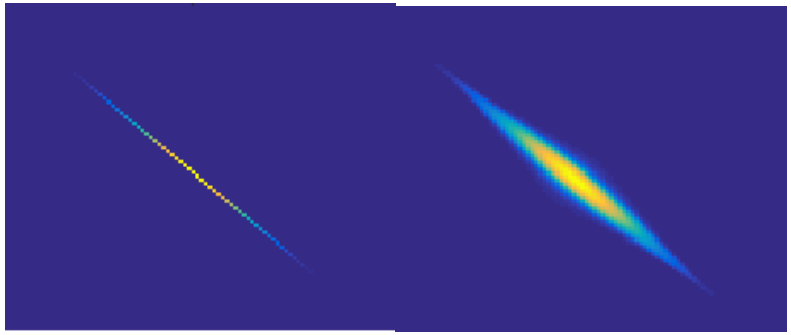
- Linear programming problem with complexity $O(n^3 \ln n)$ arithmetic operations [Pele & Werman, 2009].
- State-of-the-art approach [Cuturi, 2013]. Solve by Sinkhorn's algorithm an *entropy-regularized optimal transport* problem

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

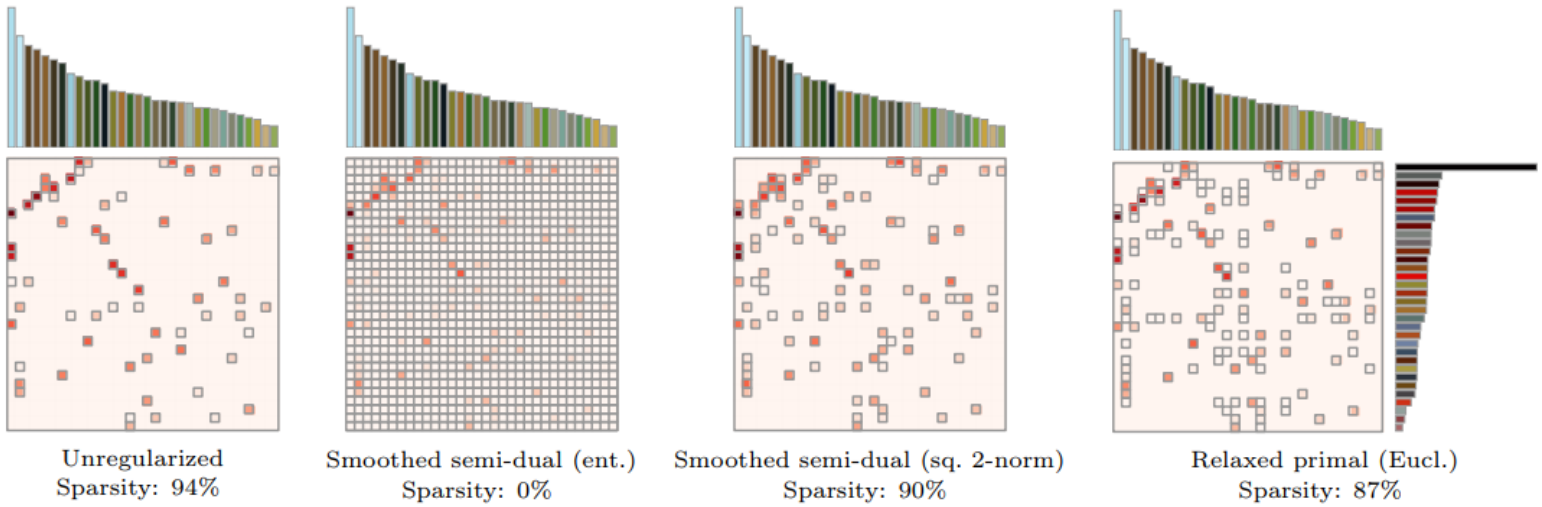
- Complexity [Altschuler et.al.,2017]

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

- Blurring in the transportation plan.



- Dense transportation plan.



Lower image: Blondel et al., 2017

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

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

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

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

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**
    $u_{k+1} = u_k + \ln(r/(B(u_k, v_k)\mathbf{1}))$, $v_{k+1} = v_k$
  else
    $v_{k+1} = v_k + \ln(c/(B(u_k, v_k)^T \mathbf{1}))$, $u_{k+1} = u_k$
  end if
  $k = k + 1$
until $\|B(u_k, v_k)\mathbf{1} - r\|_1 + \|B(u_k, v_k)^T \mathbf{1} - c\|_1 \leq \varepsilon'$

Bounds for the iterates and optimal solution

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

Sinkhorn's convergence rate

Sinkhorn's algorithm requires no more than

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

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

Input: Accuracy ε.

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

2: Define $(\tilde{r}, \tilde{c}) = \left(1 - \frac{\varepsilon'}{8}\right)\left((r, c) + \frac{\varepsilon'}{n(8-\varepsilon')}(\mathbf{1}, \mathbf{1})\right)$.
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 $\mathcal{U}(r, c)$ by Algorithm 2 in [Altschuler et.al.,2017].

Complexity of OT by Sinkhorn

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

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

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

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

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

Desired features:

- accelerated convergence rates $O(1/k^2)$ separately for $f(x_k) - f^*$ and $\|Ax_k - b\|$;
- line-search;
- entropy friendliness.

Algorithm	Rates	LS	Entr.
Beck & Teboulle, 2014	×	√	√
Chambolle & Pock, 2011	×	×	×
Malitsky & Pock, 2016	×	√	×
Tran-Dinh & Cevher, 2014	√	×	√
Yurtsever et al., 2015	√	×[1]	√
Patrascu et al., 2015	√	×	√
Gasnikov et al., 2016	√	×	√
Li et al., 2016	√	×	√
Lan et al., 2011	×	×	√
Ouyang et al., 2015	×	√	×
Xu, 2016	√	×	×
Our algorithm	√	√	√

[1] Their algorithm uses Lipschitz constant in the stopping criterion and, hence, is not completely adaptive.

Primal Problem

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

Dual problem

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

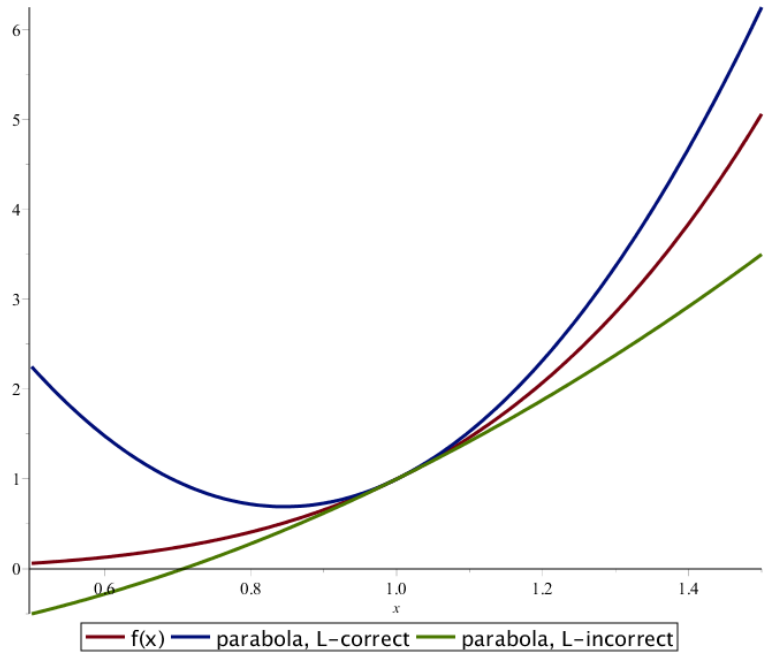
$$\nabla \varphi(\lambda) = b - Ax(\lambda), \quad x(\lambda) := \arg\max_{x \in Q} \left(-f(x) - \langle A^T \lambda, x \rangle \right).$$

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

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

We assume that the dual problem has a solution λ^* s.t. $\|\lambda^*\|_2 \leq R < +\infty$.

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



Input: 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: $\zeta_{k+1} = \zeta_k - \alpha_{k+1} \nabla \varphi(\lambda_{k+1})$.

7: $\eta_{k+1} = \tau_k \zeta_{k+1} + (1 - \tau_k)\eta_k$.

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: $\hat{x}_{k+1} = \tau_k x(\lambda_{k+1}) + (1 - \tau_k)\hat{x}_k$.

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

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

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

APDAGD Convergence Rate

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. Moreover, the stopping criterion in step 11 is correctly defined.

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

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

- Entropy regularization: $f(X) = \langle C, X \rangle + \gamma \langle X, \ln X \rangle$ is strongly convex w.r.t. $\| \cdot \|_1$.
- Squared Euclidean norm: $f(X) = \langle C, X \rangle + \gamma \|X\|_2^2$ is strongly convex w.r.t. the Euclidean norm.
- Other strongly convex regularizes are also possible.

Input: Accuracy ε.

1: Set $\gamma = \frac{\varepsilon}{3 \ln n}$.
2: **for** $k = 1, 2, \ldots$ **do**
3: Make step of APDAGD and calculate $\widehat{X}_k$ and η_k.
4: Find $\widehat{X}$ as the projection of $\widehat{X}_k$ on $\mathcal{U}(r, c)$ by Algorithm 2 in [Altschuler et.al.,2017].
5: **If** $\langle C, \widehat{X} - \widehat{X}_k \rangle \leq \frac{\varepsilon}{6}$ and $f(\hat{x}_k) + \varphi(\eta_k) \leq \frac{\varepsilon}{6}$, **Then** Return $\widehat{X}$. **Else** $k = k + 1$ and continue.
6: **end for**

Complexity theorem

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

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

Goal: find $\widehat{X} \in \mathcal{U}(r,c)$ s.t. $\langle C, \widehat{X} \rangle \leq \min_{X \in U(r,c)} \langle C, X \rangle + \varepsilon$

- State of the art bound by Sinkhorn's algorithm [Altschuler et.al.,2017]

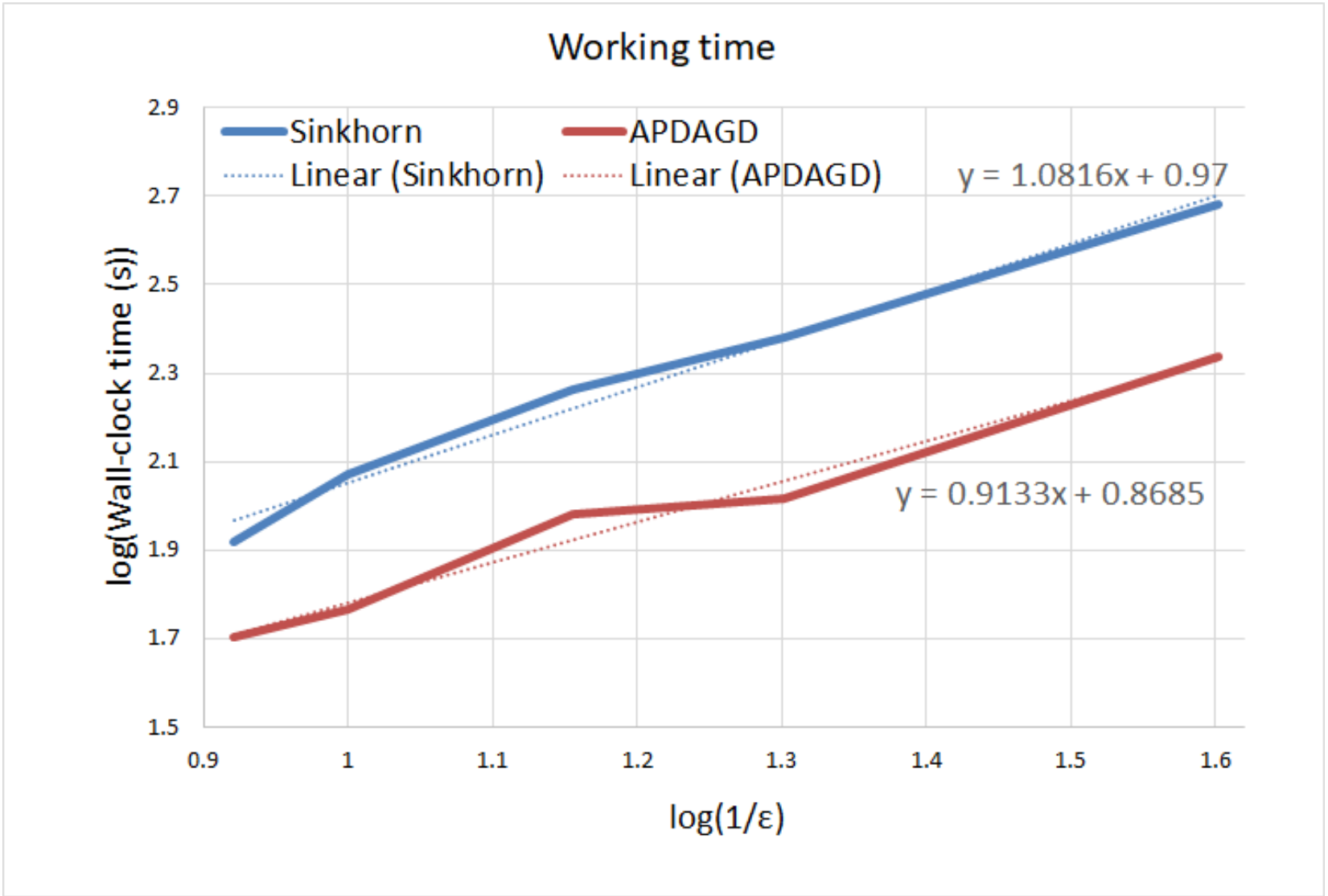
$$O\left(\frac{n^2 \|C\|_\infty^3 \ln n}{\varepsilon^3}\right).$$

- Our improved bound by Sinkhorn's algorithm

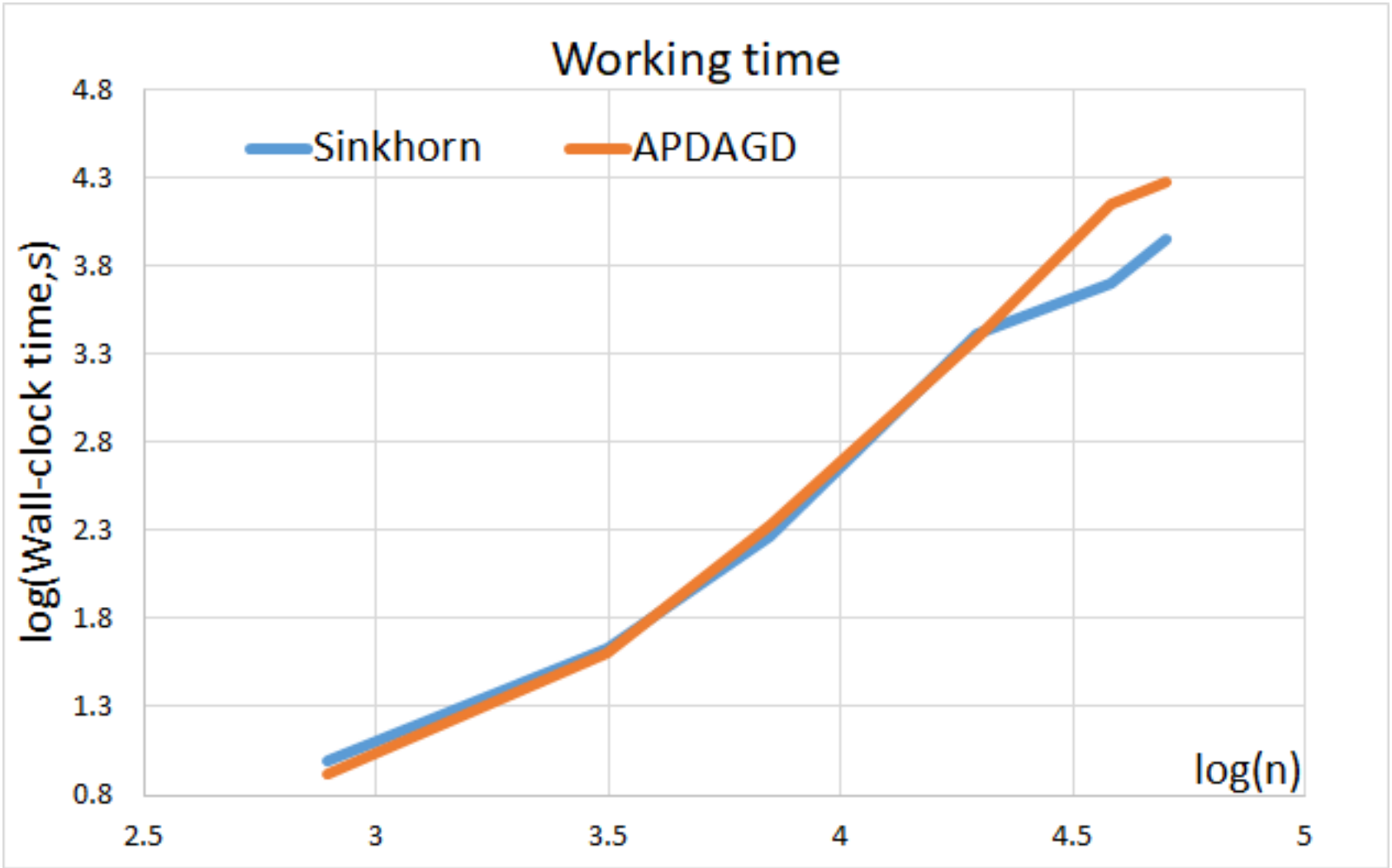
$$O\left(\frac{n^2 \|C\|_\infty^2 \ln n}{\varepsilon^2}\right).$$

- Our bound by APDAGD

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



MNIST dataset, average in 10 randomly chosen images, $\varepsilon \in [0.025, 0.12]$, $n = 784$.



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

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

We consider (regularized) optimal transport problem and

- improve complexity bounds for its solution by Sinkhorn's algorithm,
- develop an adaptive primal-dual accelerated gradient descent, which can be applied for OT problem with different regularizers
- obtain complexity bounds for OT using entropic regularization and our new APDAGD.

References:

- P. Dvurechensky, A. Gasnikov, A.Kroshnin Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm arXiv:1802.04367, to appear in ICML 2018.
- P. Dvurechensky, A. Gasnikov, S. Omelchenko, A. Tiurin Adaptive Similar Triangles Method: a Stable Alternative to Sinkhorn's Algorithm for Regularized Optimal Transport arXiv:1706.07622.

Thank you!

Sinkhorn kernel matrix $\exp(-C/\gamma)$ is easy to apply, e.g. the measures are supported on regular grids and C is given by squared Euclidean distance.

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

$$= \max_{y,z \in \mathbb{R}^n} -\langle y, r \rangle - \langle z, c \rangle - \gamma \sum_{i,j=1}^{n} \exp\left(-\frac{1}{\gamma}(y^i + z^j + C^{ij}) - 1 \right).$$

- $f(x) = \langle C, X \rangle + \gamma \langle X, \ln X \rangle$
- $Q = \mathbb{R}_+^{n^2}$, $b^T = (r^T, c^T)$
- $A : \mathbb{R}^{n^2} \to \mathbb{R}^{2n}$ defined by the identity $(A \operatorname{vec}(X))^T = ((X\mathbf{1})^T, (X^T\mathbf{1})^T)$
- $\lambda = (y, z)$

$$\nabla \varphi(\lambda) = b - Ax(\lambda) = \begin{pmatrix} r - e^{-1} \cdot \operatorname{diag}\left(e^{y/\gamma}\right) K e^{z/\gamma} \\ c - e^{-1} \cdot \operatorname{diag}\left(e^{z/\gamma}\right) K e^{y/\gamma} \end{pmatrix}$$