Weierstrass Institute for
Applied Analysis and Stochastics

Leibniz Association

Accelerated Alternating Minimization Methods with Applications to Optimal Transport

Pavel Dvurechensky

Based on joint works with D. Dvinskikh (WIAS), A. Gasnikov (MIPT), S. Guminov (MIPT), A. Kroshnin (HSE), A. Nedic (ASU), Yu. Nesterov (CORE UCL), S. Omelchenko (MIPT), N. Tupitsa (IITP RAS), C. Uribe (Rice)

One World Optimization Seminar, 14.02.2022

Content

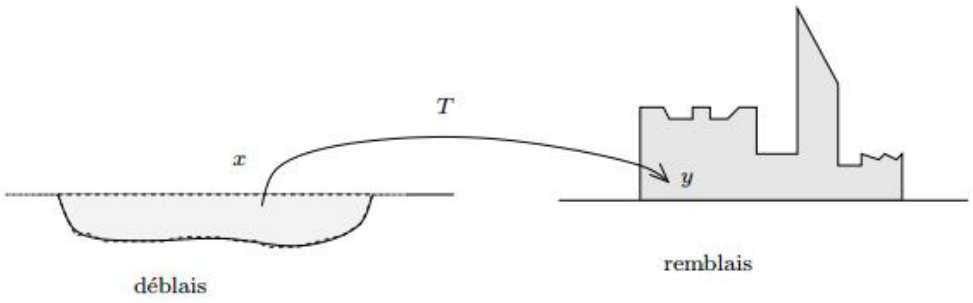
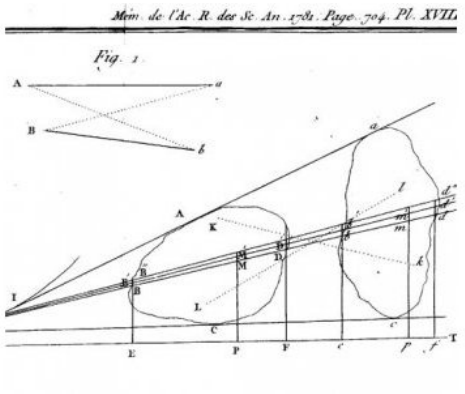


Fig. 3.1. Monge's problem of déblais and remblais



- (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. $\forall B$, $\mu(T^{-1}(B)) = \nu(B) \iff \nu = T_{\#}\mu$.

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

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

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:

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

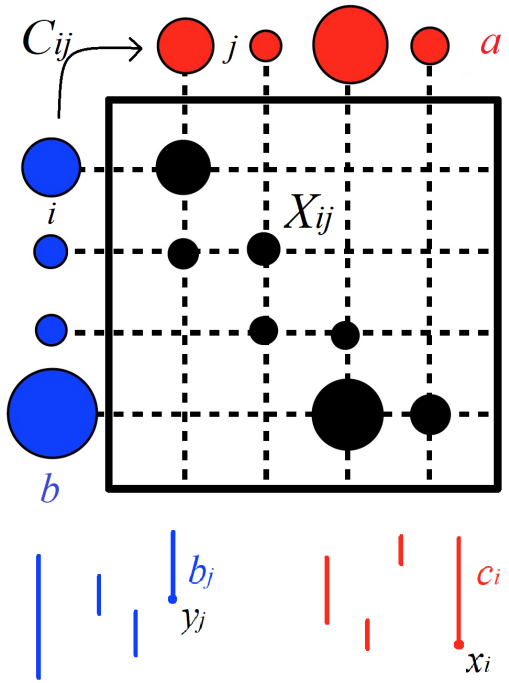
$$\inf_{\pi \in \mathcal{U}(\mu,\nu)} \int_{E \times E} C(x, y) d\pi(x, y).$$

Main feature: lifts ground metric 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.

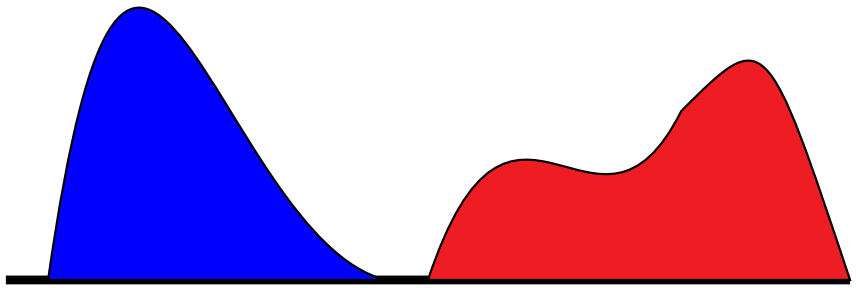
- $x_i \in \mathbb{R}^d$, $i = 1, ..., n$ – support of μ;
- $y_i \in \mathbb{R}^d$, $i = 1, ..., n$ – support of ν;
- $\mu = \sum_{i=1}^n a_i \delta(x_i), \quad a \in S_n(1)$;
- $\nu = \sum_{j=1}^n b_j \delta(y_j), \quad b \in S_n(1)$;
- $C_{ij} = C(x_i, y_j), \quad i, j = 1, ..., n$ – ground cost matrix;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



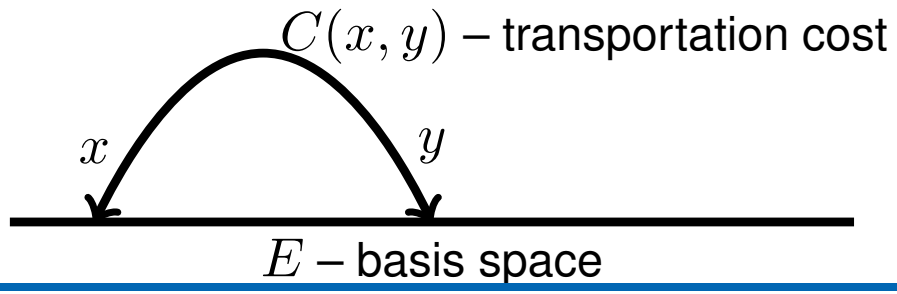
Optimal Transport (OT) Problem

$$\min_{X \in \mathcal{U}(a,b)} \langle C, X \rangle = \sum_{i,j=1}^{n} C_{ij} X_{ij},$$

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



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



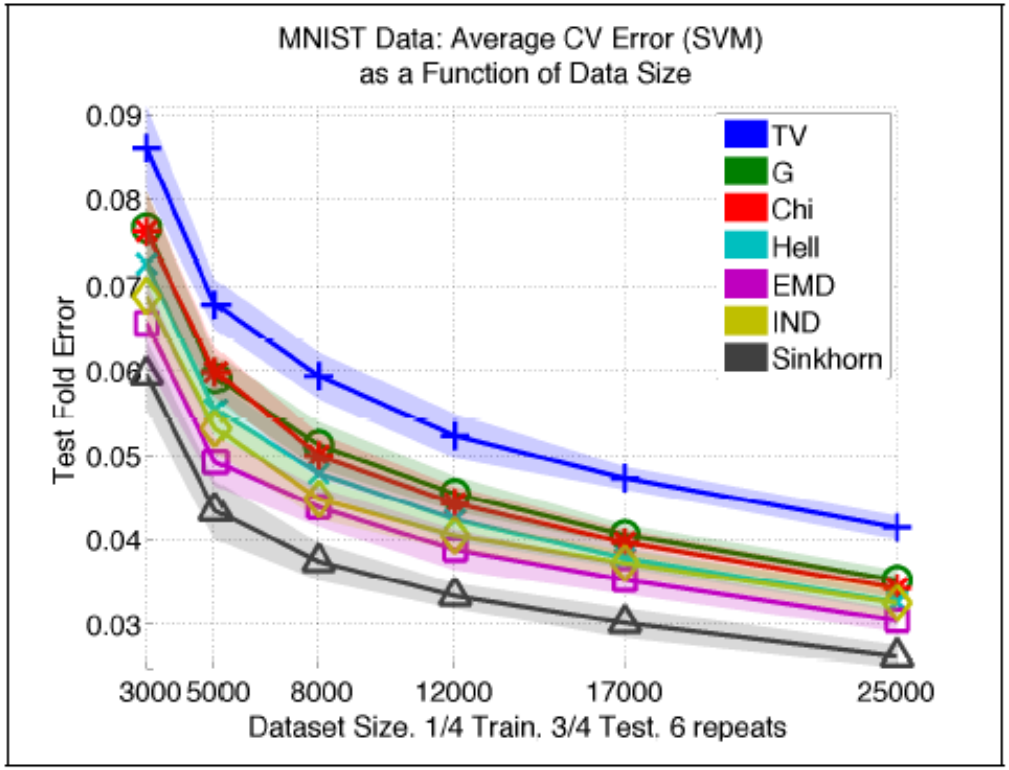
Goal: classify images from MNIST dataset 

Basis space – pixel grid

Cost – Squared Euclidean distance

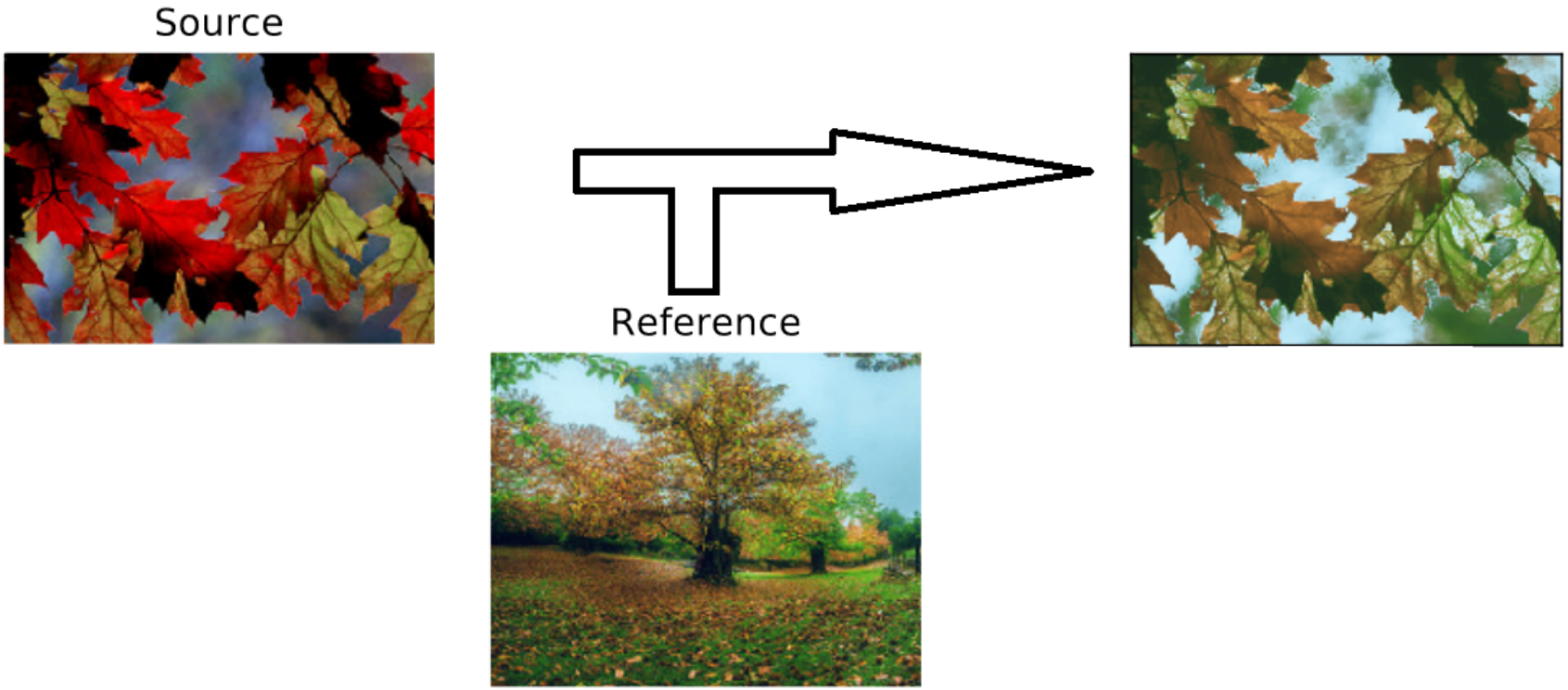
Measures – histograms of pixel intensities

Run standard SVM based on distance between images



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

Goal: transfer color from one image to another

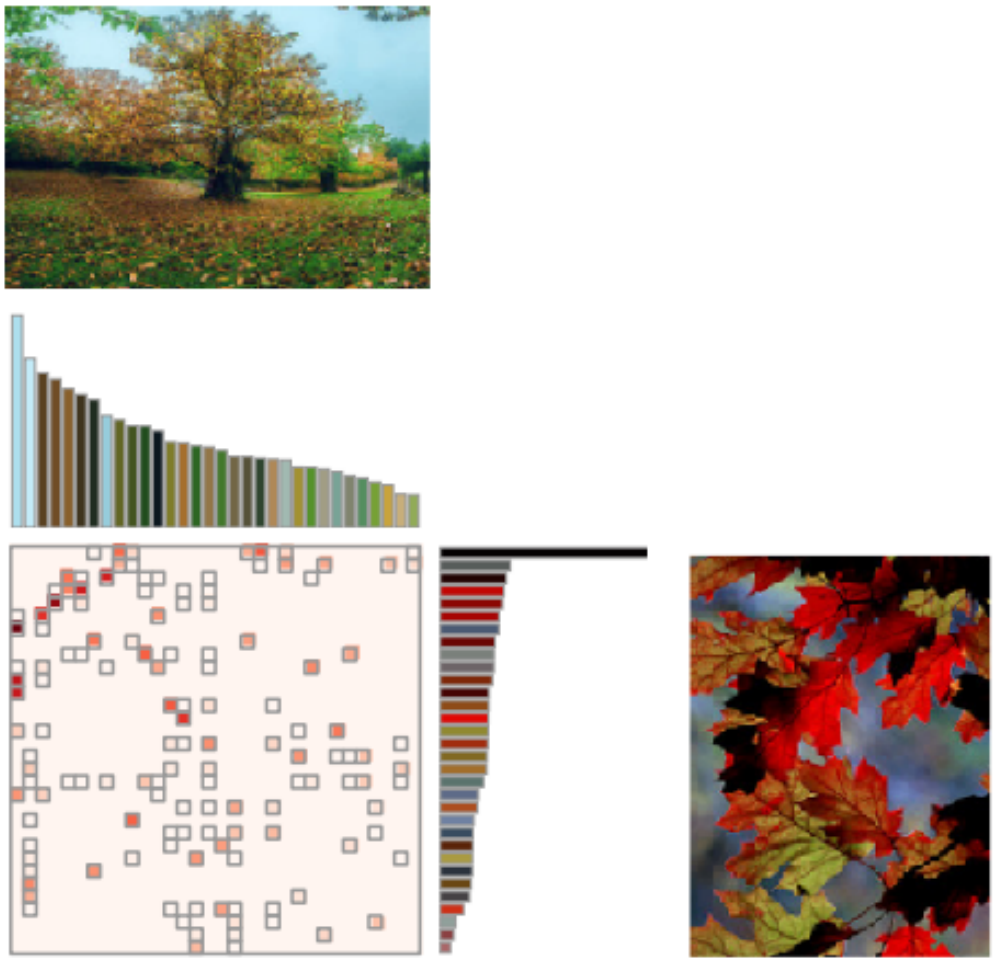


M. Blondel, V. Seguy, A. Rolet, Smooth and Sparse Optimal Transport, AISTATS 2018.

Basis space – RGB color space

Cost – Squared Euclidean distance

Measures – histograms given by clustering



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

1 Motivation: Optimal Transport (OT)

2 **Numerical methods for OT distances**

- Sinkhorn's algorithm
- Accelerated gradient method

3 Accelerated alternating minimization

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

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

- Linear programming problem with complexity $O(n^3 \ln n)$ arithmetic operations [Pele & Werman, 2009]. For $10^3 \times 10^3$ image, $n = 10^6$.
- Widespread approach [Cuturi, 2013]. Solve by Sinkhorn's algorithm an entropy-regularized optimal transport problem

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

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

O. Pele, M. Werman, Fast and robust earth mover's distances, ICCV 2009.
M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, NIPS 2013.

Primal problem

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

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

Dual problem

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

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:

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

NB: Adaptive algorithm: no need to know any smoothness parameters.

Denote $K := e^{-C/\gamma}$, $u = \xi/\gamma$, $v = \eta/\gamma$, $B(u, v) := \operatorname{diag}(e^u) K \operatorname{diag}(e^v)$.

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^*.

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

Sinkhorn's algorithm requires no more than

$$k \leq 2 + \frac{4R}{\varepsilon'} = O\left(\frac{1}{\gamma\varepsilon'}\right)$$

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

[Altschuler, et. al., 2017]: Project $B(u_k, v_k)$ on $\mathcal{U}$ to obtain the desired ε-solution.

J. Altschuler, J. Weed, P. Rigollet, Near-linear time approximation algorithms for optimal transport..., NIPS 2017.
D., A. Gasnikov, A. Kroshnin, Computational Optimal Transport: Complexity by Accelerated Gradient..., ICML 2018.

- Entropy-specific.
- Complexity $\frac{1}{\gamma\varepsilon}$ or rate $\frac{1}{\gamma k}$.
- Adaptivity.
- May be unstable for small γ.

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

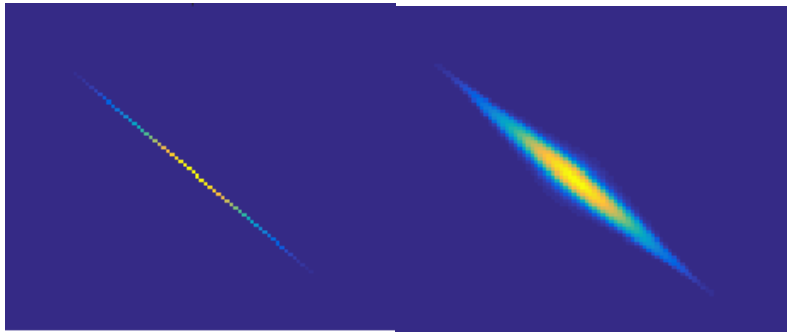
$$O\left(\frac{n^2\|C\|_\infty^2 \ln n}{\varepsilon^2}\right) \quad \textit{arithmetic operations}.$$

Previous bound $O\left(\frac{n^2\|C\|_\infty^3 \ln n}{\varepsilon^3}\right)$ by [Altschuler, Weed, Rigollet, 2017].

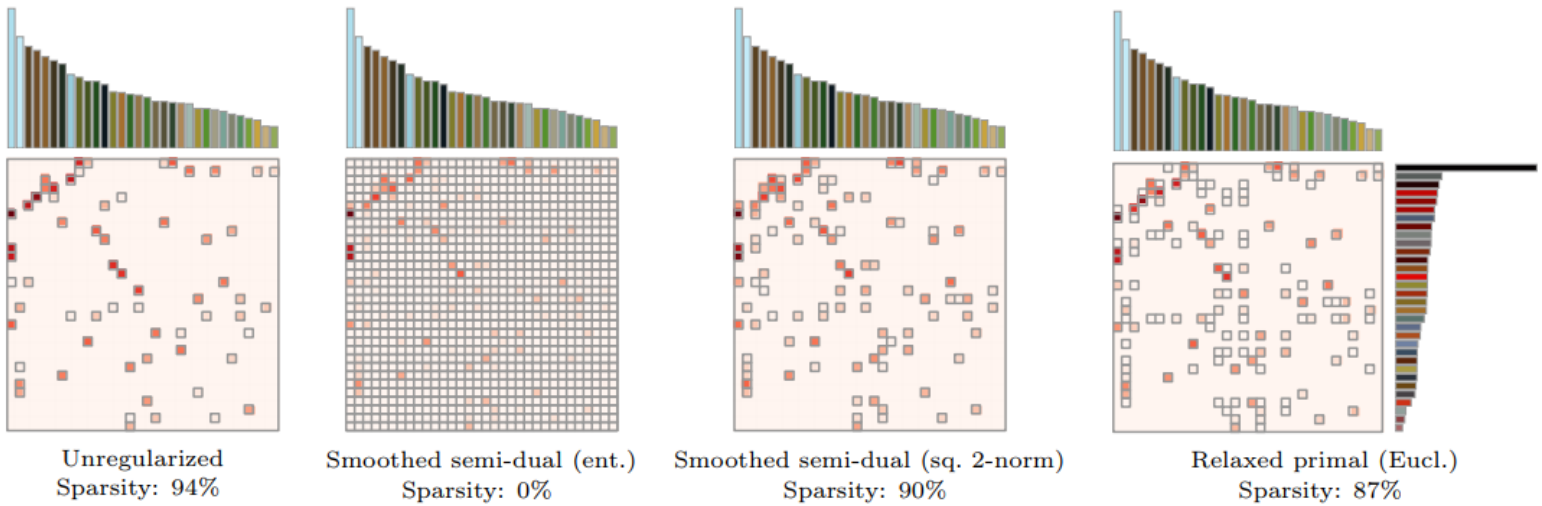
Can we propose something else?

D., A. Gasnikov, A. Kroshnin, Computational Optimal Transport: Complexity by Accelerated Gradient..., ICML 2018.
J. Altschuler, J. Weed, P. Rigollet, Near-linear time approximation algorithms for optimal transport..., NIPS 2017.

- Blurring in the transportation plan.



- Dense transportation plan.



Lower image: **M. Blondel, V. Seguy, A. Rolet, Smooth and Sparse Optimal Transport, AISTATS 2018.**

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

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

- $E = \mathbb{R}^{n^2}, H = \mathbb{R}^{2n}, \|\cdot\|_E = \|\cdot\|_1, Q = S_{n^2}(1)$;
- $f(x) = \langle C, X \rangle + \gamma \langle X, \ln X \rangle$; (we can use another regularizer, e.g., $\|X\|_2^2$)
- $\{x : Ax = b\} = \{X : X\mathbf{1} = a,\ X^T\mathbf{1} = b\}$.

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

Dual problem:

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

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

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

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

- Beck, A., Teboulle, M., A fast dual proximal gradient algorithm for convex minimization..., 2014.
- Chambolle, A., Pock, T. A first-order primal-dual algorithm for convex problems..., 2011.
- Malitsky, Y., Pock, T. A first-order primal-dual algorithm with linesearch, 2016.
- Tran-Dinh, Q., Cevher, V. Constrained convex minimization via model-based excessive gap, 2014.
- Yurtsever, A., Tran-Dinh, Q., Cevher, V. A universal primal-dual convex optimization framework, 2015.
- Patrascu, A., Necoara, I., Findeisen, R. Rate of convergence analysis of a dual fast gradient..., 2015.
- Gasnikov, A., Gasnikova, E., Nesterov, Y., Chernov, A. Efficient numerical methods for entropy..., 2016.
- Chernov, A., Dvurechensky, P., Gasnikov, A. Fast primal-dual gradient method..., 2016.
- Li, J., Wu, Z., Wu, C., Long, Q., Wang, X. An inexact dual fast gradient-projection method, 2016.
- Lan, G., Lu, Z., Monteiro, R. D. C. Primal-dual first-order methods with $O(1/\varepsilon)$ iteration..., 2011.
- Ouyang, Y., Chen, Y., Lan, G., Eduardo Pasiliao, J. An accelerated linearized alternating direction..., 2015.
- Xu, Y. Accelerated first-order primal-dual proximal methods for linearly constrained..., 2016.
- Tran-Dinh, Q., Fercoq, O., Cevher, V. A Smooth Primal-Dual Optimization Framework..., 2015.
- Alacaoglu, A., Tran-Dinh, Q., Fercoq, O., Cevher, V. Smooth Primal-Dual Coordinate Descent... , 2017.
- * Tran-Dinh, Q., Alacaoglu, A., Fercoq, O., Cevher, V. An Adaptive Primal-Dual Framework..., 2018.
- ...

Desired features:

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

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

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

    [$\sim$ 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

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

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

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

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

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

    [$\sim$ 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

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

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

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

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

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

    [$\sim$ 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

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

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

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

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

Let f in the primal problem be γ-strongly convex and the dual solution λ^* satisfy $\|\lambda^*\|_2 \leq R$. Then, for $k \geq 1$, the points $\hat{x}_k$, η_k in APDAGD satisfy

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

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

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

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

Complexity $O\left(\frac{1}{\sqrt{\gamma\varepsilon}}\right)$. (cf. $O\left(\frac{1}{\gamma\varepsilon}\right)$ for the Sinkhorn's algorithm.)

D., A. Gasnikov, A. Kroshnin, Computational Optimal Transport: Complexity by Accelerated Gradient..., ICML 2018.

- General regularizers.
- Complexity $\frac{1}{\sqrt{\gamma\varepsilon}}$ or rate $\frac{1}{\gamma k^2}$.
- Adaptivity.
- Extra dimension-dependent factor in the complexity for the OT problem.

Complexity of OT by APDAGD [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

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

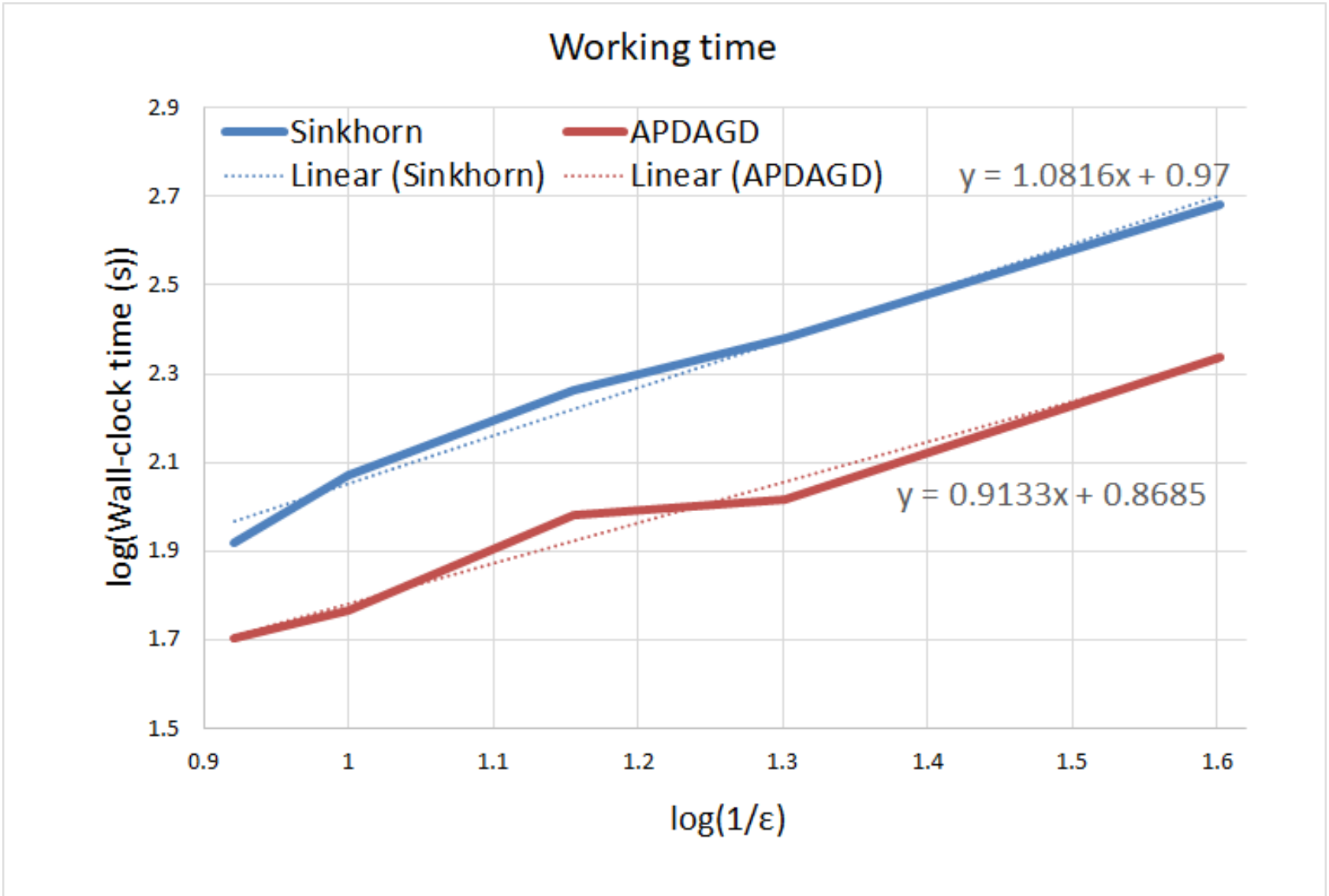
From the Sinkhorn's anslysis (slide 16), one obtains [Lin, Ho, Jordan, 2019], [Guminov, et. al., 2021] that $R \leq \|C\|_\infty\sqrt{n}$, and the bound $O\left(\frac{n^{5/2}\|C\|_\infty\sqrt{\ln n}}{\varepsilon}\right)$.

D., A. Gasnikov, A. Kroshnin, Computational Optimal Transport: Complexity by Accelerated Gradient..., ICML 2018.
T. Lin, N. Ho, M. Jordan, On Efficient Optimal Transport: An Analysis of Greedy and Accelerated..., ICML 2019.
S. Guminov, D., N. Tupitsa, A. Gasnikov, On a Combination of Alternating Minimization and Nesterov..., ICML 2021.

Algorithm	Complexity
Sinkhorn/Greenkhorn [Altschuler, Weed, Rigollet, 2017]	$n^2\|C\|_\infty^3/\varepsilon^3$
Sinkhorn [D., Gasnikov, Kroshnin, 2018]	$n^2\|C\|_\infty^2/\varepsilon^2$
Greenkhorn [Lin, Ho, Jordan, 2019a]	$n^2\|C\|_\infty^2/\varepsilon^2$
Randkhorn [Lin, Ho, Jordan, 2019b]	$n^{7/3}\|C\|_\infty^{4/3}/\varepsilon$
APDA(G/M)D [D., Gasnikov, Kroshnin, 2018], [Lin, Ho, Jordan, 2019a]	$n^{5/2}\|C\|_\infty/\varepsilon$
Mirror-Prox [Jambulapati, A. Sidford, K. Tian, 2019]	$n^2\|C\|_\infty/\varepsilon$
Accelerated Sinkhorn [Guminov, D., Tupitsa, Gasnikov, 2021]	$n^{5/2}\|C\|_\infty/\varepsilon$

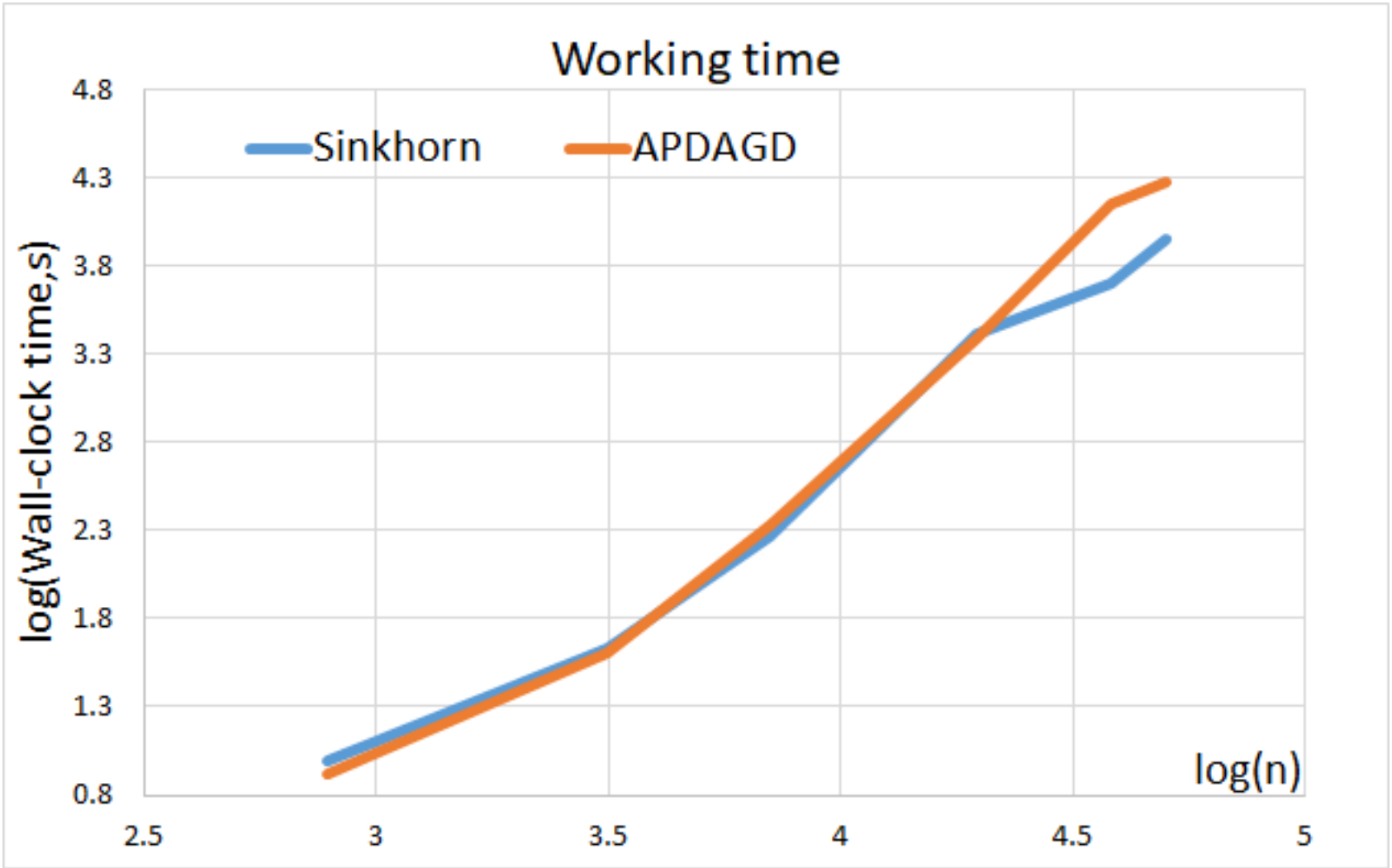
J. Altschuler, J. Weed, P. Rigollet, Near-linear time approximation algorithms for optimal transport..., NeurIPS 2017.
D., A. Gasnikov, A. Kroshnin, Computational Optimal Transport: Complexity by Accelerated Gradient..., ICML 2018.
T. Lin, N. Ho, M. Jordan, On Efficient Optimal Transport: An Analysis of Greedy and Accelerated..., ICML 2019a.
T. Lin, N. Ho, M. Jordan, On the efficiency of the Sinkhorn and Greenkhorn algorithms..., 2019b.
A. Jambulapati, A. Sidford, K. Tian, A direct $\tilde{O}(1/\varepsilon)$ iteration parallel algorithm for OT, NeurIPS 2019.
S. Guminov, D., N. Tupitsa, A. Gasnikov, On a Combination of Alternating Minimization and Nesterov's Momentum, ICML 2021.

Comparison with Sinkhorn's algorithm



MNIST dataset, average in 10 randomly chosen images.

Comparison with Sinkhorn's algorithm 2



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

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

- Adaptive idea 1: Alternating minimization in the dual a.k.a. Sinkhorn's algorithm
  - Complexity
$$O\left(\frac{1}{\gamma\varepsilon}\right).$$
  - Very fast convergence for large γ, unstable for small γ.
  - Empirically faster than in theory.
- Adaptive idea 2: adaptive to Lipschitz constant AGD in the dual
  - Complexity
$$O\left(\frac{1}{\sqrt{\gamma\varepsilon}}\right).$$
  - Stable for small γ.
  - Still based on Lipschitz constant of the gradient.

Can we combine alternating minimization and AGD?

We consider the (dual) minimization problem

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

- The space $\mathbb{R}^N$ is divided into n disjoint subsets (blocks) I_i, $i \in \{1, \dots, n\}$.
- $S_i(\lambda) = \lambda + \operatorname{span}\{e_j : \ j \in I_i\}$, i.e. the affine subspace containing λ and all the points differing from λ only over the block i.
- 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 $i \in \{1, \dots, n\}$ and any $\zeta \in \mathbb{R}^N$ the problem $\min\limits_{\lambda \in S_i(\zeta)} \varphi(\lambda)$ has a solution, and this solution is easily computable.
- $\varphi(\lambda)$ is L_φ-smooth: $\|\nabla \varphi(\lambda) - \nabla \varphi(\eta)\|_2 \leqslant L_\varphi \|\lambda - \eta\|_2, \ \ \forall\, \lambda,\ \eta \in \mathbb{R}^N$.

Primal problem:

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

Dual problem:

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

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

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

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

Related works

- Beck, A. and Tetruashvili, L. On the convergence of block coordinate descent type methods, 2013.
- Beck, A. On the convergence of alternating minimization for convex programming with appl..., 2015.
- Saha, A. and Tewari, A. On the nonasymptotic convergence of cyclic coordinate descent methods, 2013.
- Sun, R. and Hong, M. Improved iteration complexity bounds of cyclic block coordinate descent..., 2015.
- Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems, 2012.
- Lee, Y. T. and Sidford, A. Efficient accelerated coordinate descent methods and faster algorithms..., 2013.
- Shalev-Shwartz, S. and Zhang, T. Accelerated proximal stochastic dual coordinate ascent..., 2014.
- Fercoq, O. and Richtárik, P. Accelerated, parallel, and proximal coordinate descent, 2015.
- Lin, Q., Lu, Z., and Xiao, L. An accelerated proximal coordinate gradient method, 2014.
- Allen-Zhu, Z., Qu, Z., Richtarik, P., and Yuan, Y. Even faster accelerated coordinate descent..., 2016.
- Nesterov, Y. and Stich, S. U. Efficiency of the accelerated coordinate descent method..., 2015.
- Lu, H., Freund, R., and Mirrokni, V. Accelerating greedy coordinate descent methods, 2018.
- Tran-Dinh, Q., Fercoq, O., Cevher, V. A Smooth Primal-Dual Optimization Framework..., 2015.
- Alacaoglu, A., Tran-Dinh, Q., Fercoq, O., Cevher, V. Smooth primal-dual coordinate descent..., 2017.
- Diakonikolas, J. and Orecchia, L. Alternating randomized block coordinate descent, 2018.
- ...

Desired features:

- parameter-free/adaptive;
- primal-dual algorithm;
- accelerated convergence rates $O(1/k^2)$ separately for $f(x_k) - f^*$ and $\|Ax_k - b\|$;
- guarantees for non-convex minimization;
- arbitrary number of blocks.

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

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

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

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

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

- Instead of gradient step $\eta_{k+1} = \lambda_{k+1} - \frac{1}{L}\nabla\varphi(\lambda_{k+1})$ we consider the Gauss-Southwell rule + block minimization:

Choose $i_k = \arg\max_{i\in\{1,\dots,n\}} \|\nabla_i\varphi(\lambda_{k+1})\|_2^2$. Set $\eta_{k+1} = \arg\min_{\eta\in S_{i_k}(\lambda_{k+1})} \varphi(\eta)$.

- Instead of gradient step $\eta_{k+1} = \lambda_{k+1} - \frac{1}{L}\nabla\varphi(\lambda_{k+1})$ we consider the Gauss-Southwell rule + block minimization:

  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: $\zeta_{k+1} = \zeta_k - \alpha_{k+1}\nabla\varphi(\lambda_{k+1})$, $\beta_k + \alpha_{k+1} = L\alpha_{k+1}^2$

- Instead of gradient step $\eta_{k+1} = \lambda_{k+1} - \frac{1}{L}\nabla\varphi(\lambda_{k+1})$ we consider the Gauss-Southwell rule + block minimization:

Choose $i_k = \arg\max\limits_{i\in\{1,\dots,n\}} \|\nabla_i\varphi(\lambda_{k+1})\|_2^2$. Set $\eta_{k+1} = \arg\min\limits_{\eta\in S_{i_k}(\lambda_{k+1})} \varphi(\eta)$.

- Momentum step: $\zeta_{k+1} = \zeta_k - \alpha_{k+1}\nabla\varphi(\lambda_{k+1})$, $\beta_k + \alpha_{k+1} = L\alpha_{k+1}^2$

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

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

- Instead of gradient step $\eta_{k+1} = \lambda_{k+1} - \frac{1}{L}\nabla\varphi(\lambda_{k+1})$ we consider the Gauss-Southwell rule + block minimization:

Choose $i_k = \arg\max\limits_{i\in\{1,\dots,n\}} \|\nabla_i\varphi(\lambda_{k+1})\|_2^2$. Set $\eta_{k+1} = \arg\min\limits_{\eta\in S_{i_k}(\lambda_{k+1})} \varphi(\eta)$.

- Momentum step: $\zeta_{k+1} = \zeta_k - \alpha_{k+1}\nabla\varphi(\lambda_{k+1})$, $\beta_k + \alpha_{k+1} = L\alpha_{k+1}^2$

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

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

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

- Instead of gradient step $\eta_{k+1} = \lambda_{k+1} - \frac{1}{L}\nabla\varphi(\lambda_{k+1})$ we consider the Gauss-Southwell rule + block minimization:

Choose $i_k = \arg\max\limits_{i\in\{1,\dots,n\}} \|\nabla_i\varphi(\lambda_{k+1})\|_2^2$. Set $\eta_{k+1} = \arg\min\limits_{\eta\in S_{i_k}(\lambda_{k+1})} \varphi(\eta)$.

- Momentum step: $\zeta_{k+1} = \zeta_k - \alpha_{k+1}\nabla\varphi(\lambda_{k+1})$, $\beta_k + \alpha_{k+1} = L\alpha_{k+1}^2$

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

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

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

- Coupling step:

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

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

Y. Nesterov, A. Gasnikov, S. Guminov, and D., Primal-dual accelerated gradient methods with small-dimen..., 2020.

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

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

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

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

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

6: [Block minimization] Set $\eta_{k+1} = \arg\min_{\eta \in S_{i_k}(\lambda_k)} \varphi(\eta).$

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

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

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

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

10: **end for**

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

S. Guminov, D., N. Tupitsa, A. Gasnikov, On a Combination of Alternating Minimization and Nesterov's Momentum, ICML 2021.

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

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

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

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

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

6: [Block minimization] Set $\eta_{k+1} = \arg\min\limits_{\eta \in S_{i_k}(\lambda_k)} \varphi(\eta)$.

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

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

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

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

10: **end for**

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

S. Guminov, D., N. Tupitsa, A. Gasnikov, On a Combination of Alternating Minimization and Nesterov's Momentum, ICML 2021.

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

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

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

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

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

6: [Block minimization] Set $\eta_{k+1} = \arg\min\limits_{\eta \in S_{i_k}(\lambda_k)} \varphi(\eta)$.

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

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

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

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

10: **end for**

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

S. Guminov, D., N. Tupitsa, A. Gasnikov, On a Combination of Alternating Minimization and Nesterov's Momentum, ICML 2021.

When applied to a convex and L_φ-smooth objective $\varphi(\cdot)$:

$$\varphi(\eta_k) - \varphi(\lambda^*) \leq \frac{2nL_\varphi \|\lambda_0 - \lambda^*\|_2^2}{k^2} = O\left(\frac{n}{k^2}\right),$$

When applied to a non-convex and L_φ-smooth objective $\varphi(\cdot)$:

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

Uniformly optimal in terms of k method for smooth convex and non-convex problems, no knowledge of the convexity and parameters like L_φ.

In the primal-dual setting (slide 34), if f is γ-strongly convex and $\|\lambda^*\|_2 \leq R$:

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

$$\|A\hat{x}_k - b\|_2 \leq \frac{8n\|A\|^2 R}{\gamma k^2} = O\left(\frac{n}{\gamma k^2}\right), \quad \|\hat{x}_k - x^*\|_E \leq \frac{4n}{k}\frac{\|A\|R}{\gamma} = O\left(\frac{n}{\gamma k}\right),$$

x^*, f^* – resp. an optimal solution and the optimal value in the primal problem.

S. Guminov, D., N. Tupitsa, A. Gasnikov, On a Combination of Alternating Minimization and Nesterov's Momentum, ICML 2021.

When φ is μ-strongly convex, we can find α_{k+1} from

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

where $\rho_0 = 1$ and recursively $\rho_{k+1} = \rho_k + \mu\alpha_{k+1}$.

Then, we obtain the rate

$$\varphi(\eta_k) - \varphi(\eta^*) \leq nL_\varphi\|\eta_0 - \eta^*\|^2 \min\left\{\frac{4}{k^2}, \left(1 - \sqrt{\frac{\mu}{nL_\varphi}}\right)^{k-1}\right\}.$$

Under the Polyak-Lojasiewicz (PL) condition, i.e., $\|\nabla\varphi(\eta)\|_2^2 \geq 2\sigma(\varphi(\eta) - \varphi(\eta^*))$, we can run this algorithm with $\mu = 0$ and obtain linear (non-accelerated) convergence

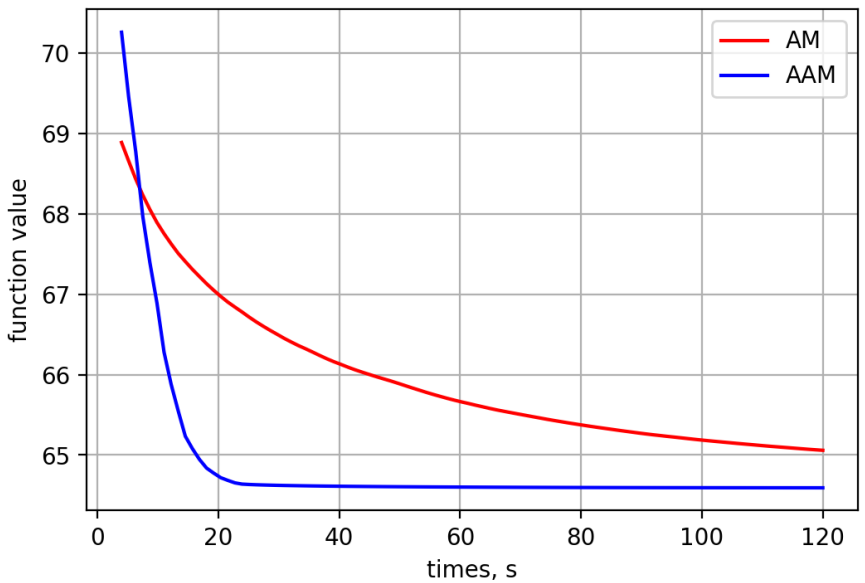
$$\varphi(\eta_k) - \varphi(\eta^*) \leq \prod_{i=0}^{k-1}\left(1 - \frac{\sigma}{\hat{L}_i}\right)(\varphi(\eta_0) - \varphi(\eta^*)) \leq \left(1 - \frac{\sigma}{nL_\varphi}\right)^k (\varphi(\eta_0) - \varphi(\eta^*)),$$

where $\hat{L}_i = \frac{\beta_{k+1}}{\alpha_{k+1}^2} \leq nL_\varphi$.

N. Tupitsa, P. Dvurechensky, A. Gasnikov, and S. Guminov. Alternating minimization methods for strongly convex optimization, 2021.

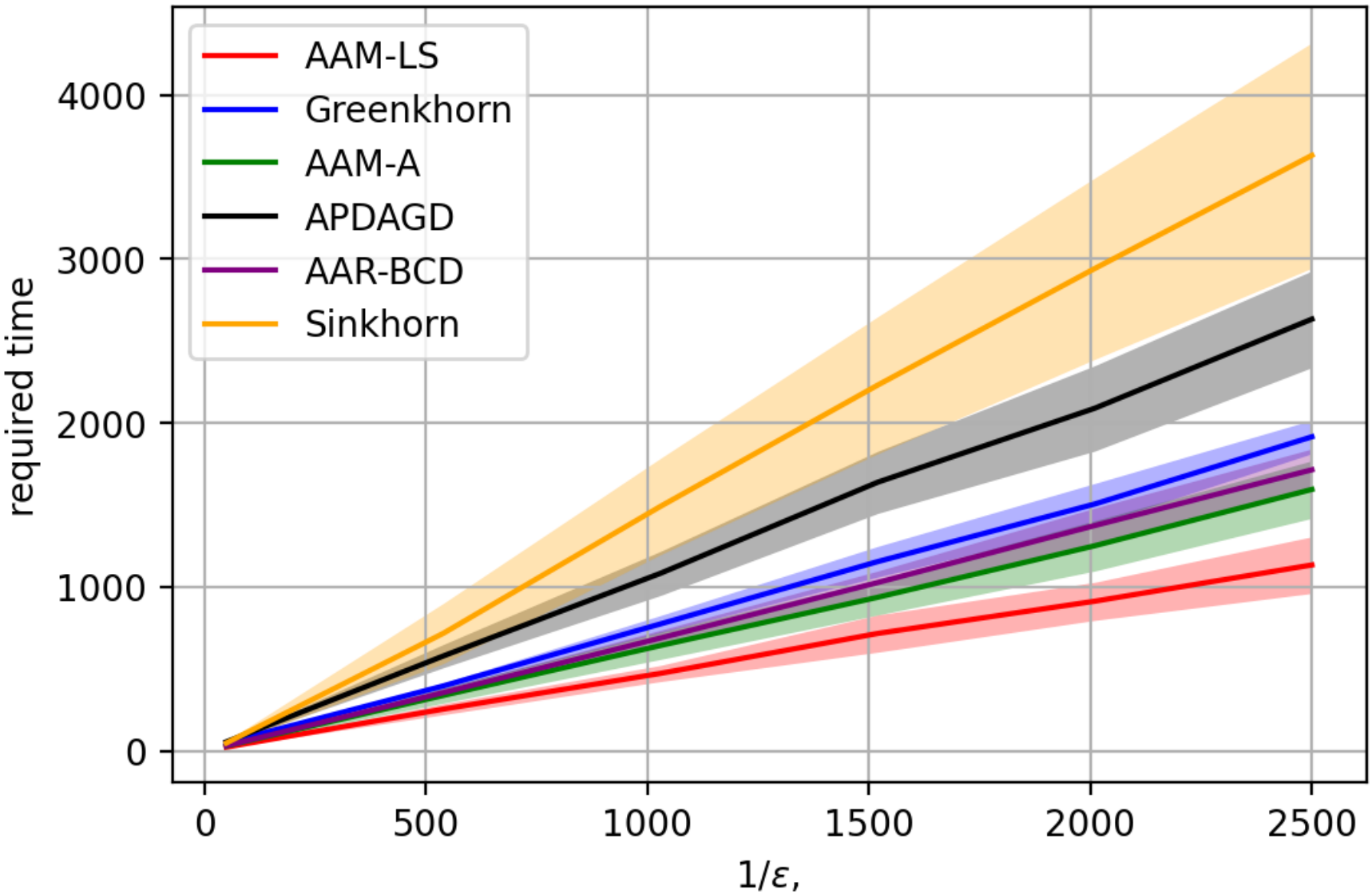
The unknown ratings $\hat{r}_{ui}$ associated with the user u and the item i are sought as a product $x_u^\top y_i$, where the vectors x_u and y_i are the optimized variables. We assume that we are given r_{ui} – observed preference rates associated with some users and items.

$$\min_{x,y} \; F(x,y) = \sum_{\text{observed } u,i} c_{ui} \left(r_{ui} - x_u^\top y_i\right)^2 + \lambda \sum_u \|x_u\|_2^2 + \lambda \sum_i \|y_i\|_2^2.$$



Algorithm	Complexity
Sinkhorn/Greenkhorn [Altschuler, Weed, Rigollet, 2017]	$n^2\|C\|_\infty^3/\varepsilon^3$
Sinkhorn [D., Gasnikov, Kroshnin, 2018]	$n^2\|C\|_\infty^2/\varepsilon^2$
Greenkhorn [Lin, Ho, Jordan, 2019a]	$n^2\|C\|_\infty^2/\varepsilon^2$
Randkhorn [Lin, Ho, Jordan, 2019b]	$n^{7/3}\|C\|_\infty^{4/3}/\varepsilon$
APDA(G/M)D [D., Gasnikov, Kroshnin, 2018], [Lin, Ho, Jordan, 2019a]	$n^{5/2}\|C\|_\infty/\varepsilon$
Mirror-Prox [Jambulapati, A. Sidford, K. Tian, 2019]	$n^2\|C\|_\infty/\varepsilon$
Accelerated Sinkhorn [Guminov, D., Tupitsa, Gasnikov, 2021]	$n^{5/2}\|C\|_\infty/\varepsilon$

J. Altschuler, J. Weed, P. Rigollet, Near-linear time approximation algorithms for optimal transport..., NeurIPS 2017.
D., A. Gasnikov, A. Kroshnin, Computational Optimal Transport: Complexity by Accelerated Gradient..., ICML 2018.
T. Lin, N. Ho, M. Jordan, On Efficient Optimal Transport: An Analysis of Greedy and Accelerated..., ICML 2019a.
T. Lin, N. Ho, M. Jordan, On the efficiency of the Sinkhorn and Greenkhorn algorithms..., 2019b.
A. Jambulapati, A. Sidford, K. Tian, A direct $\tilde{O}(1/\varepsilon)$ iteration parallel algorithm for OT, NeurIPS 2019.
S. Guminov, D., N. Tupitsa, A. Gasnikov, On a Combination of Alternating Minimization and Nesterov's Momentum, ICML 2021.


AAM-LS
Greenkhorn
AAM-A
APDAGD
AAR-BCD
Sinkhorn
required time
0
1000
2000
3000
4000
0
500
1000
1500
2000
2500
1/ε,

Motivated by Optimal Transport we considered

- Sinkhorn's algorithm as an Alternating Minimization algorithm,
- Adaptive Primal-Dual Accelerated Gradient Descent (APDAGD),
- Accelerated Alternating Minimization (AAM) algorithm.

Obtained results

- Improved complexity bounds for OT by Sinkhorn's algorithm.
- Convergence rate analysis of APDAGD with complexity bounds for OT.
- AAM that is universal for convex and non-convex optimization and adaptive to smoothness, with optimal in k convergence rates. Linearly convergent extensions for strongly convex functions and under PL condition.
- Primal-dual AAM algorithm with complexity bounds for OT.

P. Dvurechensky, A. Gasnikov, A. Kroshnin, Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm, ICML 2018

Yu. Nesterov, A. Gasnikov, S. Guminov, P. Dvurechensky, Primal–dual accelerated gradient methods with small-dimensional relaxation oracle, Optimization methods and software, 2020

S. Guminov, P. Dvurechensky, N. Tupitsa, A. Gasnikov, On a Combination of Alternating Minimization and Nesterov's Momentum, ICML 2021

N. Tupitsa, P. Dvurechensky, A. Gasnikov, and S. Guminov. Alternating minimization methods for strongly convex optimization, 2021.

Similar ideas turned out to be productive for

- Fixed-support Wasserstein barycenter problem. Given a set of m measures $b_i \in S_n(1), i = 1, ..., m$, their Wasserstein barycenter is a minimizer $\hat{a}$ of

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

Our results include stochastic and distributed accelerated primal-dual methods.

P. Dvurechensky, D. Dvinskikh, A. Gasnikov, C. A. Uribe, and A. Nedic, Decentralize and randomize: Faster algorithm for Wasserstein barycenters, NeurIPS 2018

A. Kroshnin, N. Tupitsa, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and C.A. Uribe, On the complexity of approximating Wasserstein barycenters, ICML 2019

S. Guminov, P. Dvurechensky, N. Tupitsa, A. Gasnikov, On a Combination of Alternating Minimization and Nesterov's Momentum, ICML 2021

- Multimarginal optimal transport.

$$\min_{X \in \mathbb{R}_+^{n \times \cdots \times n},\quad X[-i][\mathbf{1}]^{n-1} = a_i, i = 1, \ldots, m} \langle C, X \rangle.$$

N. Tupitsa, P. Dvurechensky, A. Gasnikov, and C. A. Uribe, Multimarginal optimal transport by accelerated alternating minimization, Conference on Decision and Control (CDC), 2020

Thank you!