Weierstrass Institute for
Applied Analysis and Stochastics

Leibniz Association

Primal-dual accelerated gradient methods with alternating minimization

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 (MIT)

Optimization without borders, July 12-17, 2021

First-order methods generate dual information that can be accumulated to

- accelerate the algorithm;
- construct the model of the primal objective;
- extract approximate dual solution

Influential work on the smoothing technique:

with $f \in C_L^{1,1}(Q)$. For simplicity, we assume that the constant $L > 0$ is known. Recall that the standard gradient projection method at this problem converges as $O(\frac{1}{k})$, where k is the iteration counter (see, e.g. [7]).

In our scheme we update recursively two sequences of points $\{x_k\}_{k=0}^{\infty} \subset Q$ and $\{y_k\}_{k=0}^{\infty} \subset Q$ in such a way that they satisfy the following relation:

$$A_k f(y_k) \le \psi_k \equiv \min_x \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] : \; x \in Q \right\}, \qquad (\mathcal{R}_k)$$

Note that

$$f_\mu(x) = \max_u \{\langle Ax, u\rangle_2 - \hat{\phi}(u) - \mu d_2(u) : \ u \in Q_2\}$$

$$= \langle Ax, u_\mu(x)\rangle_2 - \hat{\phi}(u_\mu(x)) - \mu d_2(u_\mu(x)),$$

$$\langle \nabla f_\mu(x), x\rangle_1 = \langle A^* u_\mu(x), x\rangle_1.$$

Therefore

$$f_\mu(x_i) - \langle \nabla f_\mu(x_i), x_i\rangle_1 = -\hat{\phi}(u_\mu(x_i)) - \mu d_2(u_\mu(x_i)), \quad i = 0, \ldots, N. \tag{4.6}$$

Thus, in view of (2.6) and (4.6) we have

$$\sum_{i=0}^{N}(i+1)[\bar{f}_\mu(x_i) + \langle \nabla \bar{f}_\mu(x_i), x - x_i\rangle_1]$$

$$\leq \sum_{i=0}^{N}(i+1)[f_\mu(x_i) - \langle \nabla f_\mu(x_i), x_i\rangle_1] + \frac{1}{2}(N+1)(N+2)(\hat{f}(x) + \langle A^*\hat{u}, x\rangle_1)$$

$$\leq -\sum_{i=0}^{N}(i+1)\hat{\phi}(u_\mu(x_i)) + \frac{1}{2}(N+1)(N+2)(\hat{f}(x) + \langle A^*\hat{u}, x\rangle_1)$$

$$\leq \frac{1}{2}(N+1)(N+2)[-\hat{\phi}(\hat{u}) + \hat{f}(x) + \langle Ax, \hat{u}\rangle_2].$$

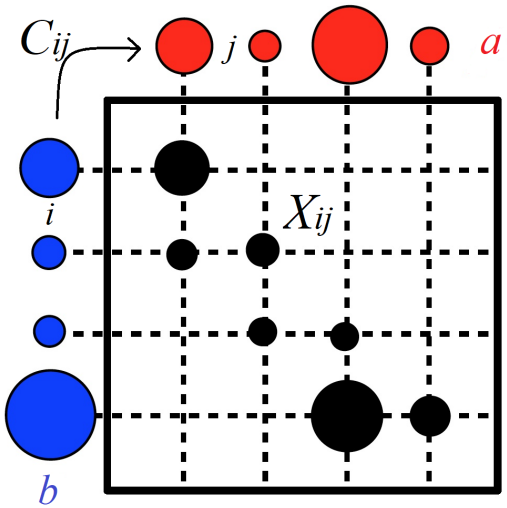
Hence, using (4.5), (2.3) and (2.7), we get the following bound:

$$\frac{L_\mu D_1}{\sigma_1 (N+1)^2} \geq \bar{f}_\mu(\hat{x}) - \phi(\hat{u}) \geq f(\hat{x}) - \phi(\hat{u}) - \mu D_2.$$

That is

$$0 \leq f(\hat{x}) - \phi(\hat{u}) \leq \mu D_2 + \frac{4\|A\|_{1,2}^2 D_1}{\mu\sigma_1\sigma_2(N+1)^2} + \frac{4MD_1}{\sigma_1(N+1)^2}. \tag{4.7}$$

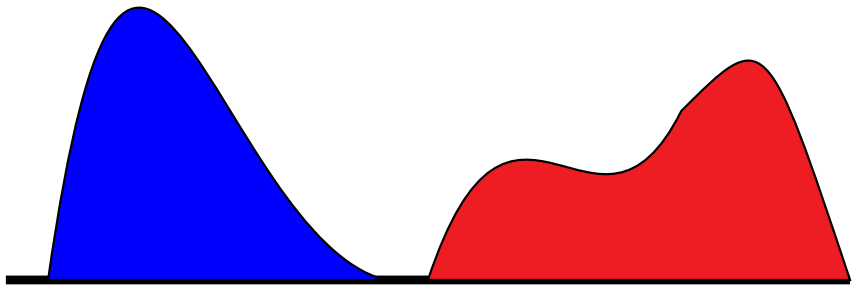
- $x_i \in \mathbb{R}^d$, $i = 1, ..., n$ – support of μ;
- $y_i \in \mathbb{R}^d$, $i = 1, ..., n$ – support of ν;
- $\mu = \sum_{i=1}^n a_i \delta(x_i), \quad a \in S_n(1)$;
- $\nu = \sum_{j=1}^n b_j \delta(y_j), \quad b \in S_n(1)$;
- $C_{ij} = C(x_i, y_j), \quad i, j = 1, ..., n$ – ground cost matrix;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



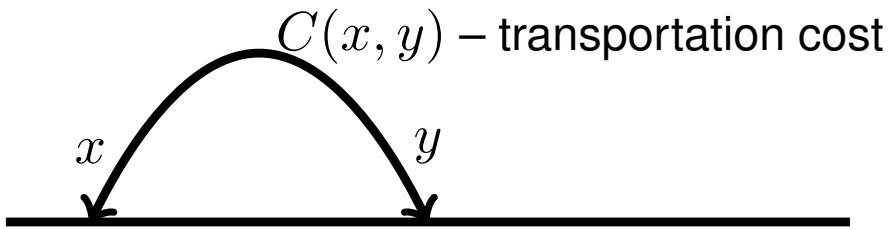
Optimal transport problem

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

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



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



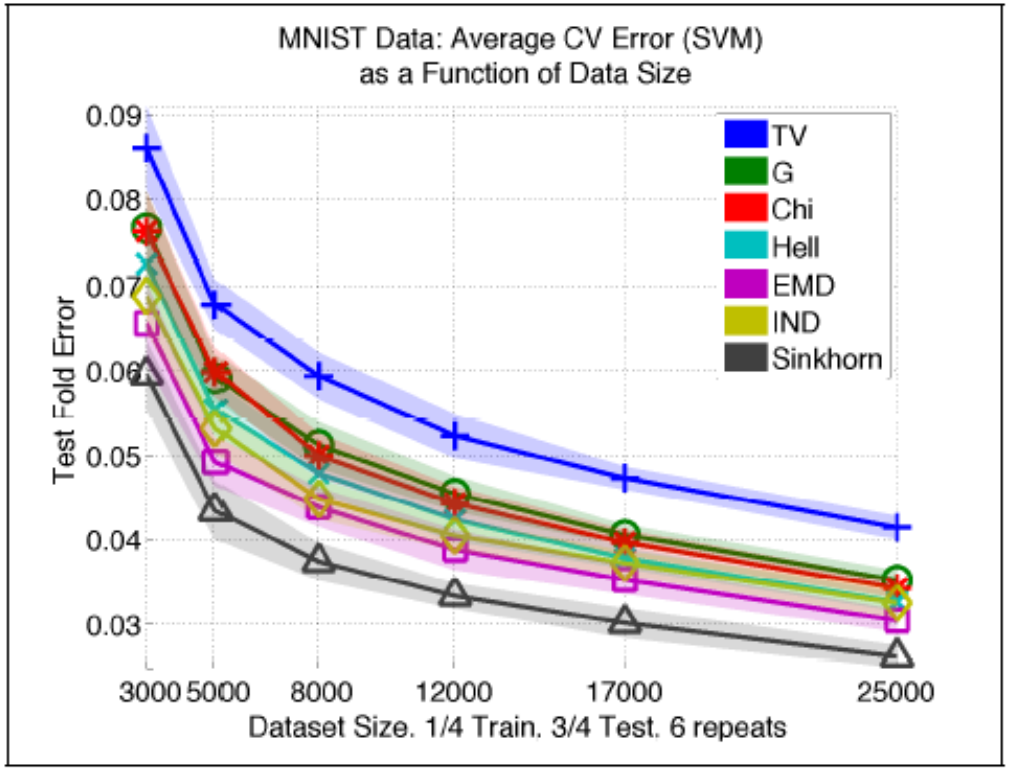
Goal: classify images from MNIST dataset 

Basis space – pixel grid

Cost – Squared Euclidean distance

Measures – histograms of pixel intensities

Run standard SVM based on distance between images



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

1 Motivation: Optimal Transport

2 **Numerical methods for OT distance**

- 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].
- Widespread approach [Cuturi, 2013]. Solve by Sinkhorn's algorithm an entropy-regularized optimal transport problem

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

Primal problem

$$\min_{X\in\mathcal{U}(a,b)} \langle C, X\rangle + \gamma\langle X, \ln X\rangle = \min_{X\in\mathcal{U}(a,b)} \gamma(-\langle X, \ln \mathrm{e}^{-\frac{C}{\gamma}}\rangle + \langle X, \ln X\rangle)$$
$$= \min_{X\in\mathcal{U}(a,b)} \gamma KL\left(X, \mathrm{e}^{-\frac{C}{\gamma}}\right).$$

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

Dual problem

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

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

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.

Bounds for the iterates and optimal solution

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

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

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

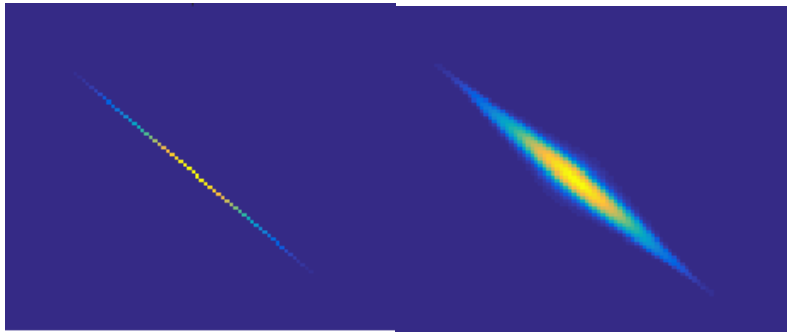
Small feasibility error turns out to be enough to make the objective small.

D., Gasnikov, Kroshnin, Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm. ICML 2018.

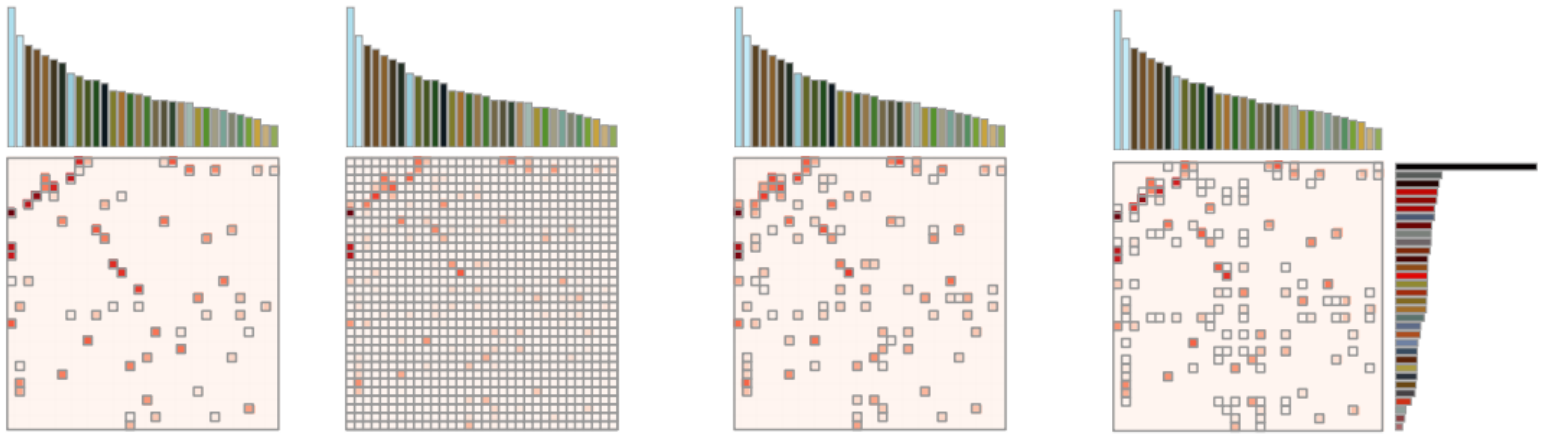
- Entropy-specific.
- Complexity $\frac{1}{\gamma\varepsilon}$ or rate $\frac{1}{\gamma k}$.
- Adaptivity.
- May be unstable for small γ.

Can we propose something else?

- Blurring in the transportation plan.



- Dense transportation plan.



Lower image: Blondel et al., 2017

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

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. $\|X\|_2^2$)
- $\{x : Ax = c\} = \{X : X\mathbf{1} = a,\ X^T\mathbf{1} = b\}$.

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	√	×	√
Patrascu et al., 2015	√	×	√
Gasnikov et al., 2016	√	×	√
Li et al., 2016	√	×	√
Lan et al., 2011	×	×	√
Ouyang et al., 2015	×	√	×
Our algorithm	√	√	√

A series of related works [Alacaoglu, Tran-Dinh, Fercoq, Cevher, 2017], [Tran-Dinh, Fercoq, Cevher, 2015-2018], [Tran-Dinh, Alacaoglu, Fercoq, Cevher, 2018] achieve similar results.

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

Dual problem

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

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

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

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

Many algorithms proposed by Yurii are primal-dual: a linear model of the objective is produced that gives information about the dual variables.

with $f \in C_L^{1,1}(Q)$. For simplicity, we assume that the constant $L > 0$ is known. Recall that the standard gradient projection method at this problem converges as $O(\frac{1}{k})$, where k is the iteration counter (see, e.g. [7]).

In our scheme we update recursively two sequences of points $\{x_k\}_{k=0}^{\infty} \subset Q$ and $\{y_k\}_{k=0}^{\infty} \subset Q$ in such a way that they satisfy the following relation:

$$A_k f(y_k) \le \psi_k \equiv \min_x \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] : \; x \in Q \right\}, \qquad (\mathcal{R}_k)$$

Main idea: apply Accelerated Gradient Method to the dual, use the primal-dual connection $x(\lambda)$ to reconstruct the primal variable.

Main obstacle: unbounded dual feasible set.

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

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

repeat {Main iterate}

  repeat {Line search}

    Set $M_k = 2^{i_k - 1} M_k$, find α_{k+1} s.t. $\beta_{k+1} := \beta_k + \alpha_{k+1} = M_k \alpha_{k+1}^2$. Set $\tau_k = \alpha_{k+1}/\beta_{k+1}$.

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

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

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

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

repeat {Main iterate}

  repeat {Line search}

    Set $M_k = 2^{i_k - 1} M_k$, find α_{k+1} s.t. $\beta_{k+1} := \beta_k + \alpha_{k+1} = M_k \alpha_{k+1}^2$. Set $\tau_k = \alpha_{k+1}/\beta_{k+1}$.

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

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

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

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

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

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

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

Complexity $O\left(\frac{1}{\sqrt{\gamma\varepsilon}}\right)$.

- 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.

- Sinkhorn's algorithm, [Altschuler, Weed, Rigollet, 2017]

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

- Sinkhorn's algorithm, [D., Gasnikov, Kroshnin, 2018]

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

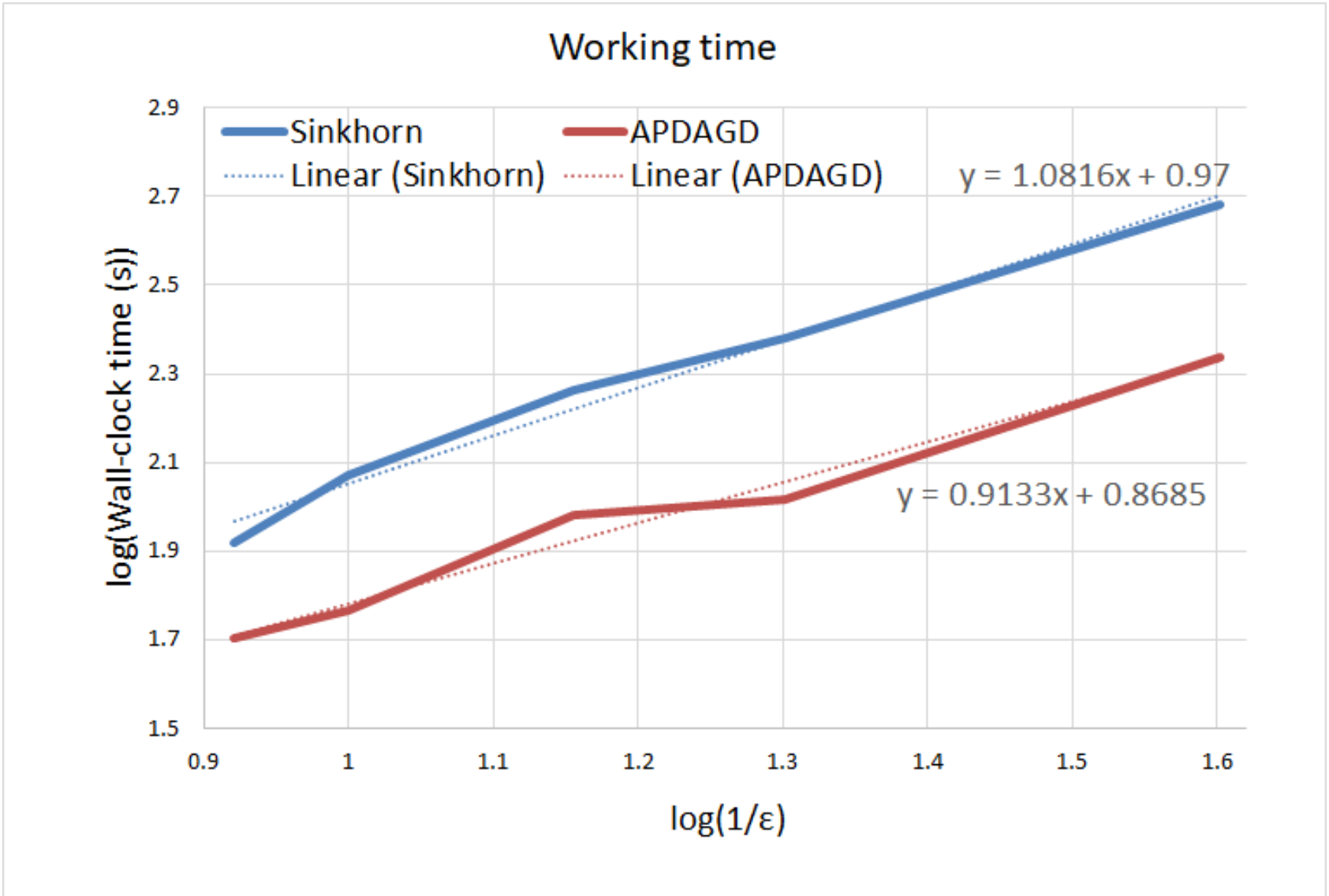
- Accelerated Gradient Descent, [D., Gasnikov, Kroshnin, 2018], [Guminov, 2019]

$$O\left(\frac{n^{2.5}\|C\|_\infty \sqrt{\ln n}}{\varepsilon}\right).$$

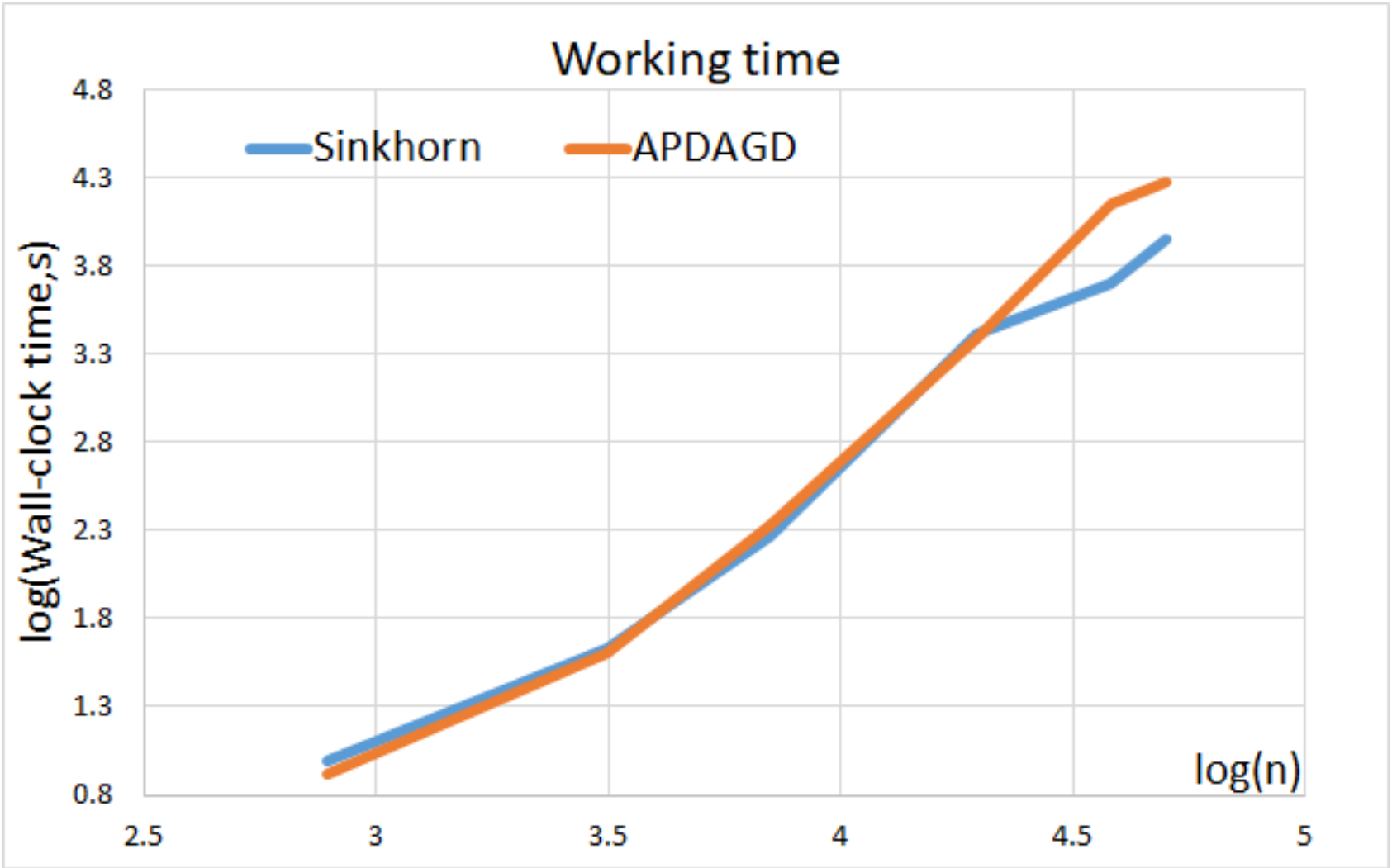
Altschuler, Weed, Rigollet, Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. NIPS 2017.

D., Gasnikov, Kroshnin, Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm. ICML 2018.

Comparison with Sinkhorn's algorithm



MNIST dataset, average in 10 randomly chosen images.



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 minimization problem

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

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

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 $\eta_{k+1} = \arg\min\limits_{\eta\in S_{i_k}(\lambda_{k+1})} \varphi(\eta)$.

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

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

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

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

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

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

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

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

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

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

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

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

Coupling step

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

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

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 $i_k = \arg\max_{i \in \{1,\ldots,n\}} \|\nabla_i \varphi(\lambda_k)\|_2^2$.

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

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

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

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

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

9: **end for**

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

Assume that the objective in the primal problem is γ-strongly convex and that the dual solution λ^* satisfies $\|\lambda^*\|_2 \leq R$. Then, for $k \geq 1$, the points $\hat{x}_k$, η_k

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

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

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

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

Assume that algorithm is applied for a non-convex and L_φ-smooth objective $\varphi(\cdot)$.

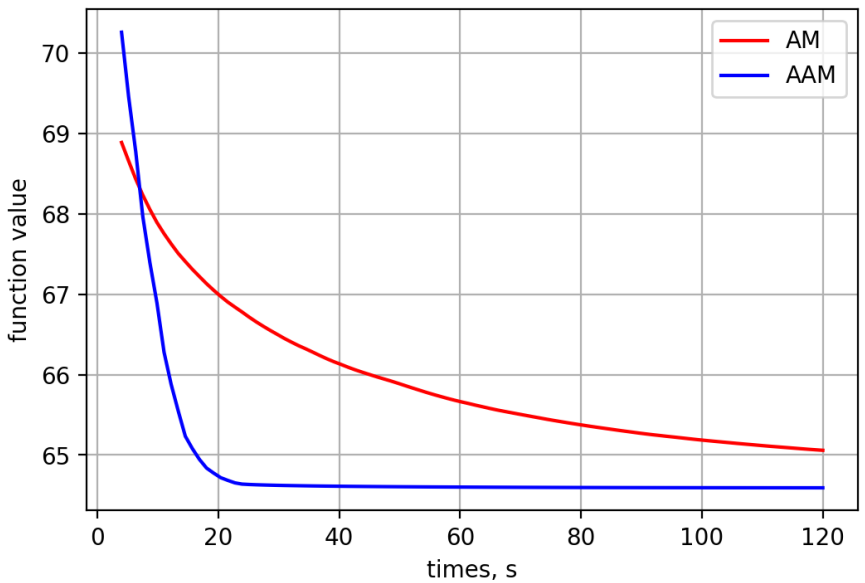
Then

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

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

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.

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



A smooth dual problem

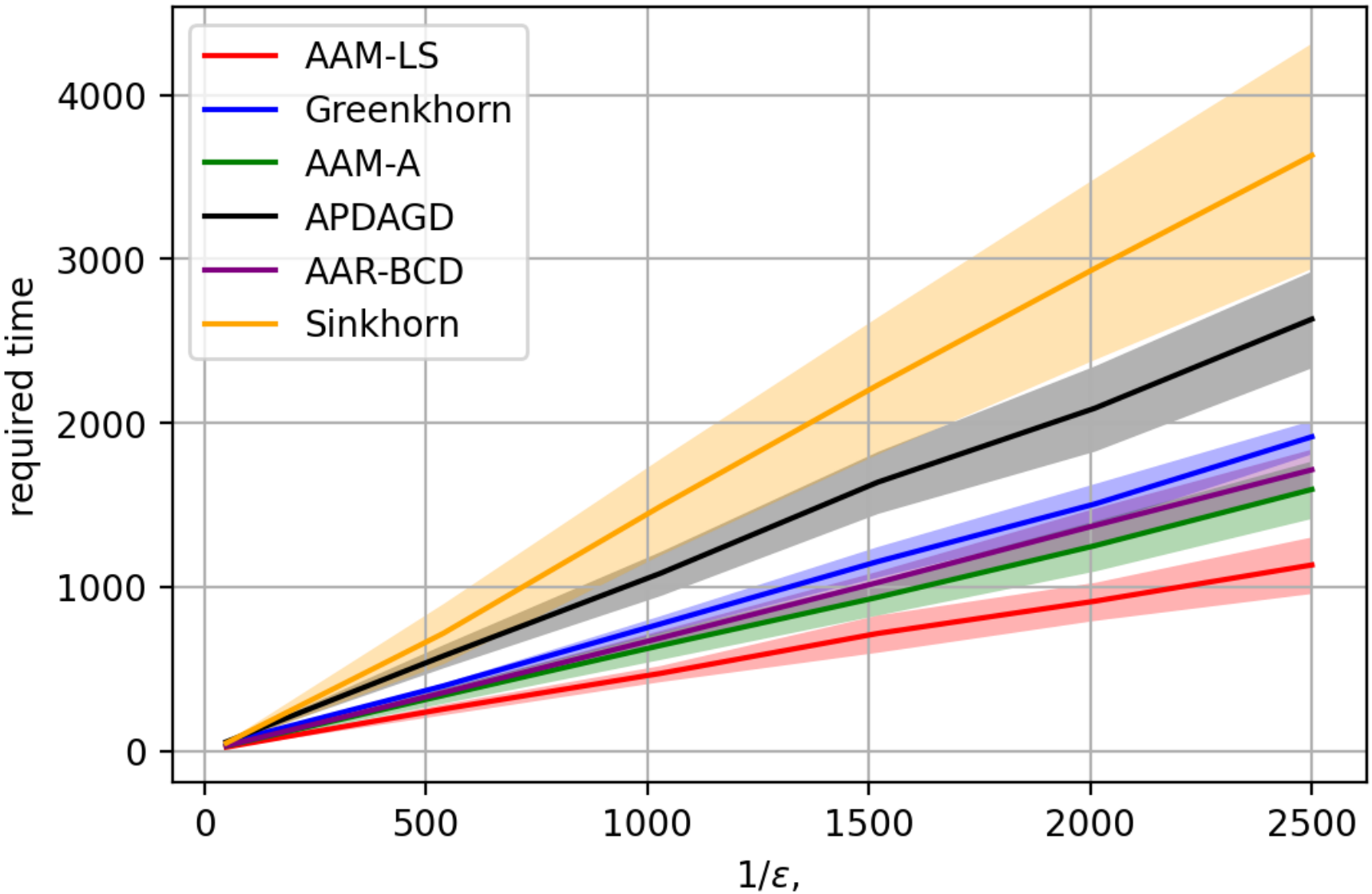
$$\min_{y,z\in\mathbb{R}^N} \varphi(y,z) = \gamma \ln\left(\sum_{i,j=1}^{N} \exp\left(-(y^i + z^j + C^{ij})/\gamma\right)\right) + \langle y, r\rangle + \langle z, c\rangle.$$

Estimate for R

$$\|(y^*, z^*)\|_2 \leqslant R := \sqrt{N/2}\left(\|C\|_\infty - \frac{\gamma}{2}\ln\min_{i,j}\{r_i, c_j\}\right).$$

Complexity for OT by AGD

$$O\left(\frac{N^{5/2}\sqrt{\ln N}\|C\|_\infty}{\varepsilon}\right) \qquad \text{cf. Sinkhorn} \quad O\left(\frac{N^2\sqrt{\ln N}\|C\|_\infty^2}{\varepsilon^2}\right).$$


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

Motivated by Optimal Transport we considered

- Alternating minimization
- Accelerated alternating minimization (AAM)

Obtained results

- AAM that is uniform for convex and non-convex optimization and adaptive to smoothness
- Complexity bounds for the OT problem

References

- 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

Primal-Dual methods

The idea of primal-dual methods turned out to be quite productive.

- Generalizations to Wasserstein barycenters
  - Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C. A., and Nedic, A. Decentralize and randomize: Faster algorithm for Wasserstein barycenters, NeurIPS 2018
  - Kroshnin, A., Tupitsa, N., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., and Uribe, C. 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
- Generalizations to multimarginal optimal transport
  - Tupitsa, N., Dvurechensky, P., Gasnikov, A., and Uribe, C. A. Multimarginal optimal transport by accelerated alternating minimization, CDC 2020
- Generalizations to Conjugate Gradient type of methods
  - Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. OMS 2020
- Stochastic primal-dual methods
  - Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C. A., and Nedic, A. Decentralize and randomize: Faster algorithm for Wasserstein barycenters, NeurIPS 2018
- Dual approach to decentralized distributed optimization
  - Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C. A., and Nedic, A. Decentralize and randomize: Faster algorithm for Wasserstein barycenters, NeurIPS 2018

Happy Burthday, Yurii Evgenievich!

Happy Burthday, Vladimir Yurievich!

Thank you!