On Some Problems Allowing Alternating Minimization

Nazarii Tupitsa

July 16, 2021

Outline

Alternating Minimization Problem

$$\min_{x \in \mathbb{R}^N} f(x), \tag{1}$$

where f is a L-smooth function and has a block structure, i.e.

$$f(x) = f(x_1, \ldots, x_n),$$

and we know exact expression for the minimizer over i-th block:

$$x_i^* = \operatorname*{argmin}_{z \in \subset \mathcal{L}_i} f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n),$$

where $\bigotimes_i^n \mathcal{L}_i = \mathbb{R}^N$.

Sample Problems

Regularized Discrete Optimal Transport

The problem is to find such a transportation plan X, that minimizes

$$\min_{X \in \mathcal{U}(r,c)} f(X) = \langle C, X \rangle + \gamma \langle X, \ln X \rangle, \tag{2}$$

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

where $C \in \mathbb{R}_+^{N \times N}$ is a given cost matrix.

- $\mathbf{1} \in \mathbb{R}^N$ is the vector of all ones,
- $r, c \in S_N(1) := \{s \in \mathbb{R}_+^N : \langle s, \mathbf{1} \rangle = 1\}$ are given discrete measures,
- $\langle A, B \rangle$ denotes the Frobenius product of matrices defined as $\langle A, B \rangle = \sum_{i,j=1}^{N} A_{ij} B_{ij}$.

OT Dual

The dual (minimization) problem

$$\min_{u,v\in\mathbb{R}^N} \varphi(u,v) = \gamma(\ln\left(\mathbf{1}^T B(u,v)\mathbf{1}\right) - \langle u, r\rangle - \langle v, c\rangle), \qquad (3)$$

where $[B(u,v)]^{ij} = \exp\left(u^i + v^j - \frac{C^{ij}}{\gamma}\right)$.

The tranportation plan is computed as follows

$$X = \frac{B(u,v)}{\mathbf{1}^T B(u,v)\mathbf{1}}$$

Sinkhorn's algorithm

Lemma

The iterations

$$u^{k+1} \in \underset{u \in \mathbb{R}^N}{\operatorname{argmin}} \varphi(u, v^k),\ v^{k+1} \in \underset{v \in \mathbb{R}^N}{\operatorname{argmin}} \varphi(u^{k+1}, v),$$

can be written explicitly as

$$u^{k+1} = u^k + \ln r - \ln\left(B\left(u^k, v^k\right)\mathbf{1}\right),$$
$$v^{k+1} = v^k + \ln c - \ln\left(B\left(u^{k+1}, v^k\right)^T \mathbf{1}\right).$$

This lemma implies that an alternating minimization method applied to the dual formulation is a natural algorithm. In fact, this is the celebrated Sinkhorn's algorithm [Sinkhorn, 1974, Cuturi, 2013].

Entropic Regularization for Wasserstein Barycenters

Using the entropic regularization we define the regularized OT-distance for $\gamma > 0$:

$$W_{C,\gamma}(p,q) = \min_{\pi \in \mathcal{U}(p,q)} \langle \pi, C \rangle + \gamma H(\pi),$$

where $H(\pi) := \sum_{i,j=1}^{n} \pi_{ij} \left(\ln \pi_{ij} - 1\right) = \left\langle \pi, \ln \pi - 11^T \right\rangle$. One may also consider the regularized barycenter which is the solution to the following problem:

$$\min_{q \in \Delta^n} \frac{1}{m} \sum_{l=1}^{m} \mathcal{W}_{C_i,\gamma}\left(p_l, q\right) \tag{4}$$

Dual for Wasserstein Barycenters

Lemma

Lemma 1 from [Kroshnin et al., 2019]. The dual (minimization) problem of (4) *is*

$$\min_{\substack{u,v \\ \sum_{l=1}^m w_l v_l = 0}} \varphi(\mathrm{u},\mathrm{v}) := \gamma \sum_{l=1}^{m} w_l \left\{ \ln\left(\mathbf{1}^T B_l\left(u_l, v_l\right)\mathbf{1}\right) - \langle u_l, p_l \rangle \right\} - m\gamma \tag{5}$$

$u = [u_1, \ldots, u_m], v = [v_1, \ldots, v_m], u_l, v_l \in \mathbb{R}^N$, *and*

$$B_l(u_l, v_l) := \mathrm{diag}(e^{u_l}) K_l \,\mathrm{diag}(e^{v_l}), \quad K_l = \exp\left(-\frac{C_l}{\gamma}\right)$$

Moreover, the solution π^*_γ *to* (4) *is given by the formula*

$$\left[\boldsymbol{\pi}^*_\gamma\right]_l = B_l\left(u_l^*, v_l^*\right) / \left(\mathbf{1}^T B_l(u_l^*, v_l^*)\mathbf{1}\right),$$

where (u^*, v^*) *is a solution to the problem* (5).

Explicit Solutions for AM

Lemma

Iterations

$$u^{k+1} = \underset{u}{\operatorname{argmin}}\, \varphi(u, v^k),\ v^{k+1} = \underset{v}{\operatorname{argmin}}\, \varphi(u^k, v),$$

may be written explicitly as

$$u_l^{k+1} = u_l^k + \ln p_l - \ln\left(B_l\left(u_l, v_l\right)\mathbf{1}\right),$$

$$v_l^{k+1} = v_l^k + \sum_{j=1}^{m} \ln(B_j(u_j^k, v_j^k)^T 1) - \ln B_l(u_l, v_l)^T \mathbf{1}.$$

Multi-marginal Optimal Transport Problem

Following the approach from [Cuturi, 2013, Benamou et al., 2015], we consider a regularized multi-marginal OT problem.

$$\min_{\substack{X\in\mathbb{R}_+^{n\times\ldots\times n},\\ p_k(X)=p_k,\quad \forall k\in\{1,\ldots,m\}\\ \sum_{i_1,\ldots,i_m} X_{i_1,\ldots,i_m}=1,\quad 1\leqslant i_j\leqslant n}} F(X) := \langle C, X\rangle - \gamma H(X), \qquad (6)$$

where

- $$[p_k(A)]_j \;=\; \sum_{1\leqslant i_l\leqslant n_l,\forall l\neq k} A_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_m}.$$
- $\gamma > 0$ is the regularization parameter,
- $H(X) \;:=\; -\langle X, \log(X)\rangle$ is the entropic regularization term
- logarithm of a tensor should be understood as component-wise.
- X belongs to a probability simplex of the size n^m. The entropy is strongly convex on the probability simplex w.r.t. the 1-norm, that implies that the dual function has a Lipschitz-continuous gradient.

MOT Dual

With the change of variable $u_k = -\frac{\lambda_k}{\gamma} - \frac{\mathbf{1}}{m}$ the dual objective becomes

$$\phi(U) \equiv \phi(u_1, \ldots, u_m) \equiv$$
$$\gamma \left[\ln \sum_{\substack{i_1,\ldots,i_m \\ 1 \leqslant i_j \leqslant n \\ 1 \leqslant j \leqslant m}} \exp \left\{ \sum_{k=1}^{m} [u_k]_{i_k} - \frac{C_{i_1 \ldots i_m}}{\gamma} \right\} - \sum_{k=1}^{m} u_k^T p_k \right], \quad (7)$$

where $U = (u_1^T, \ldots, u_m^T)^T \in \mathbb{R}^{mn}$.

$$X_{i_1 \ldots i_m}(\Lambda) = \frac{\exp\left(- \sum_{k=1}^{m} \frac{[\lambda_k]_{i_k}}{\gamma} - \frac{C_{i_1 \ldots i_m}}{\gamma} \right)}{\sum_{\substack{i_1,\ldots,i_m \\ 1 \leqslant i_j \leqslant n \\ 1 \leqslant j \leqslant m}} \exp \left\{ - \sum_{k=1}^{m} \frac{[\lambda_k]_{i_k}}{\gamma} - \frac{C_{i_1 \ldots i_m}}{\gamma} \right\}} \quad (8)$$

AM

Lemma

The iterations

$$u_k^{t+1} \in \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \phi(u_1^t, \ldots, u_{k-1}^t, u, u_{k+1}^t, \ldots, u_m^t),$$

can be written explicitly as

$$u_k^{t+1} = u_k^t + \ln p_k - \ln p_k(B(U^t)),$$

or entry-wise as

$$[u_k^{t+1}]_\eta = [u_k^t]_\eta + \ln[p_k]_\eta - \ln[p_k(B(U^t))]_\eta. \tag{9}$$

Low Rank Matrix Completion

Assume that

$$\hat{A} = UV^\top,$$

where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ consist of $\mathbf{u}_i$ and $\mathbf{v}_i$ respectively.

It turns out that assuming the matrix to have rank at most r is equivalent to assuming that the matrix $\hat{A}$ can be written as $\hat{A} = UV^\top$ with the matrices $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ having at most r columns.

$$\hat{A} = \underset{\substack{U \in \mathbb{R}^{m \times r} \\ V \in \mathbb{R}^{n \times r}}}{\arg\min} \sum_{(i,j) \in \Omega} \left\{ U_i^\top V_j - A_{ij} \right\}^2.$$

The problem

- has no constraints
- is formulated as alternating minimization in (U, V)
- is non-convex in (U, V).
- has Lipchitz continuous gradient (by Theorem 1 from [Khenissi and Nasraoui, 2019])

$$\min_{U,V} \quad F(U,V) = \sum_{\text{observed } i,j} c_{ij} \left(p_{ij} - U_i^\top V_j \right)^2 + \lambda \left(\sum_i ||U_i||_2^2 + \sum_j ||V_j||_2^2 \right). \quad (10)$$

Explicit minimization is possible

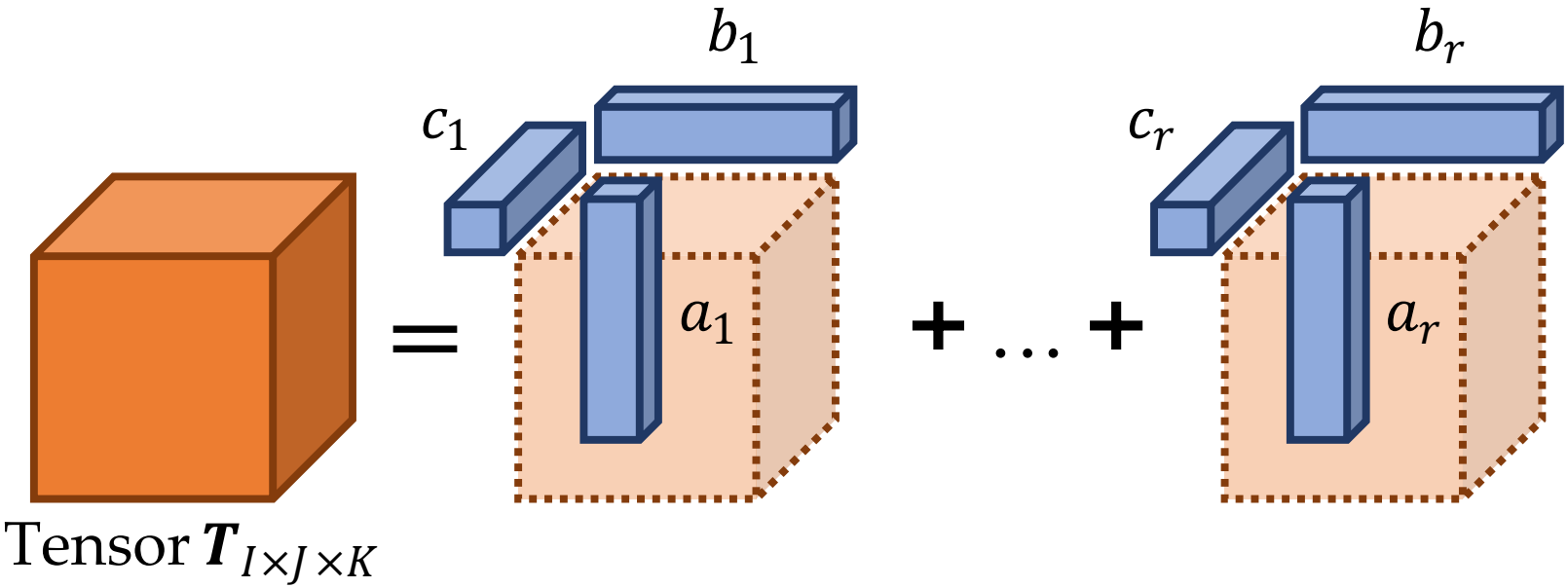
$$U_i = (V^\top C^i V + \lambda I)^{-1} V^\top C^i p(i)$$

where C_i is a diagonal with entries $[C^i]_{kk} = c_{ik}$
vector $p(i)$ contains preference by i-th user, e.g. p_{ik} values.

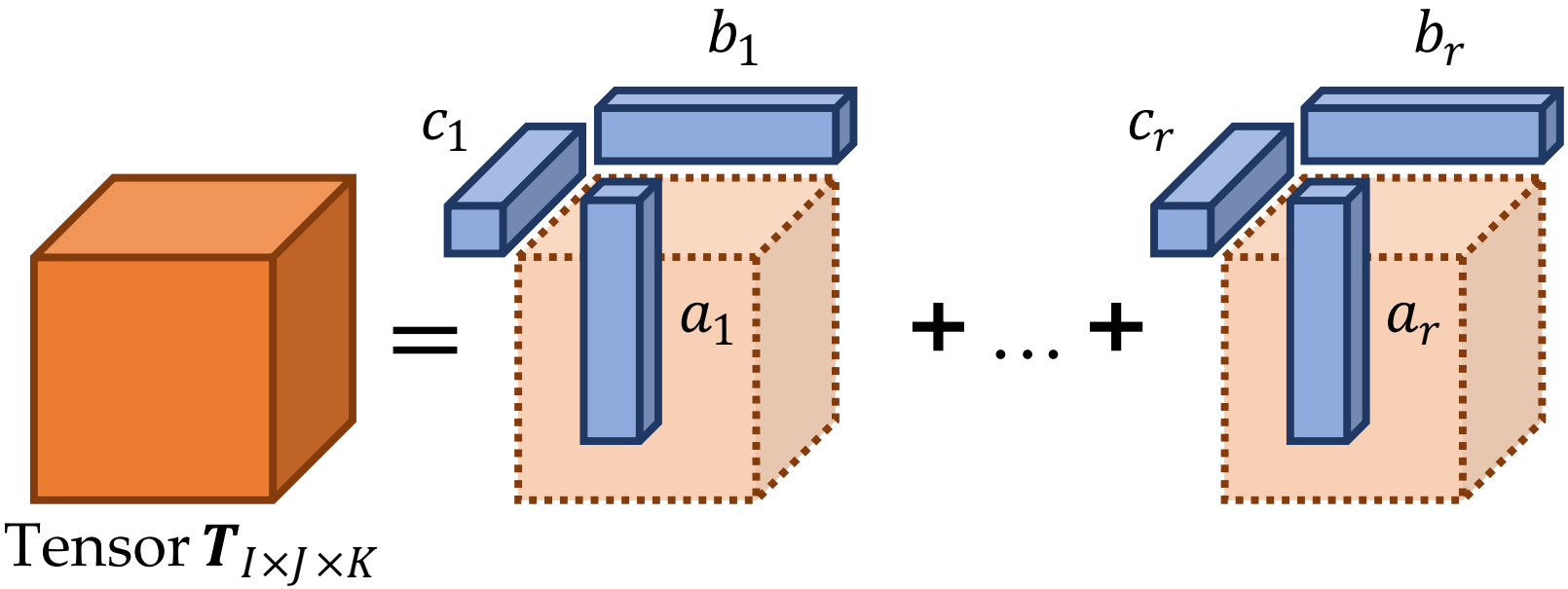
$$V_i = (U^\top C^j U + \lambda I)^{-1} U^\top C^j p(j)$$

where C_j is a diagonal with entries $[C^j]_{kk} = c_{kj}$
vector $p(j)$ contains preference by j-th item, e.g. p_{kj} values.

Canonical Polyadic Tensor Decomposition



Canonical Polyadic Tensor Decomposition



The problem is to find appropriate matrices A, B, C with given tensor $\mathbf{T}$ and CP-rank r.

CP Decomposition

CP decomposition allows us to calculate any element of the tensor in a following manner:

$$\hat{T}_{ijk} = \sum_{p=1}^{r} A_{ip} B_{jp} C_{kp}$$

Hence, we want to find matrices, that minimizes squared Frobenius norm of the residual:

$$f(x) = \frac{1}{2} \sum_{i,j,k} \left(T_{ijk} - \sum_{p=1}^{r} A_{ip} B_{jp} C_{kp} \right)^2 \rightarrow \min_{A \in \mathbb{R}^{I \times r}, B \in \mathbb{R}^{J \times r}, C \in \mathbb{R}^{K \times r}}, \tag{11}$$

where $x = (A, B, C)^\top$

AM minimization

$$A^{k+1} = \arg\min_{A} \frac{1}{2} \|\mathbf{T}_{(1)} - A(C^k \odot B^k)^T\|_F^2$$

$$B^{k+1} = \arg\min_{B} \frac{1}{2} \|\mathbf{T}_{(2)} - B(C^k \odot A^k)^T\|_F^2$$

$$C^{k+1} = \arg\min_{C} \frac{1}{2} \|\mathbf{T}_{(3)} - C(B^k \odot A^k)^T\|_F^2$$

($\mathbf{T}_{(i)}$ is the unfolded matrix of T in a current mode)

Algorithms

Standard Alternating Minimization

A very old and natural idea under this assumption is to use alternating minimization procedure
[Ortega and Rheinboldt, 2000, Bertsekas and Tsitsiklis, 1989], where the objective is minimized sequentially in each subspace.

An alternating minimization algorithm may be written as Algorithm 1 for the general case of number of blocks $n \geqslant 2$ and possible non-smooth composite term presence objective function

$$\min_{\substack{x_i \in Q_i \\ i=1,\ldots,n}} F(x_1, \ldots, x_2) \equiv f(x_1, \ldots, x_n) + \sum_{i=1}^{n} g_i(x_i), \tag{12}$$

Standard AM

Algorithm 1 Alternating Minimization

Input: Starting point x_0.
Output: x^k

1: Set x^0.
2: **for** $i \geqslant 0$ **do**
3: Select block $I = 1, \ldots, n$,
we use $I = \operatorname{argmax}_{i=1,\ldots,n} \left\| \frac{\partial F}{\partial x_I} \right\|$
4: $x_I^{k+1} = \operatorname{argmin}_{z \in Q_I} f(x_1, \ldots, x_{I-1}, z, x_{I+1}, \ldots, x_n) + g_I(z)$
5: $x_i^{k+1} = x_i^k, i \neq I$
6: **end for**

There are multiple common block selection rules, such as the cyclic rule or the Gauss–Southwell rule ($I = \operatorname{argmax}_{i=1,\ldots,n} \left\| \frac{\partial F}{\partial x_I} \right\|$), which uses further. More generally, it is also possible to update more than one block on each iteration [Hong et al., 2016].

Convergence w.r.t any norm

Theorem (n=2 for simplicity)

If F from (12) satisfies the proximal-PL inequality. Then the algorithm 1 has a linear convergence

$$F(x^{k+1}) - F^* \leqslant \left(1 - \frac{\mu_2}{L_2}\right)\left(1 - \frac{\mu_1}{L_1}\right)[F(x^k) - F^*]. \qquad (13)$$

The derivation does not specify what norm is used.

Accelerated Alternating Minimization

We proposed new accelerated method for alternating minimization which is originates in [Nesterov et al., 2020].

The Original Algorithm

Algorithm 2 AGMsDR

Input: Starting point x_0.
Output: x^N

1: Set $A_0 = 0$, $x^0 = v^0$.
2: **while** $k \leq N - 1$ **do**
3: $\beta_k = \arg\min_{\beta \in [0,1]} f\left(v^k + \beta(x^k - v^k)\right)$
$y^k = v^k + \beta_k(x^k - v^k)$
4: $h_{k+1} = \arg\min_{h \geq 0} f\left(y^k - h\nabla f(y^k)\right)$
$x^{k+1} = y^k - h_{k+1}\nabla f(y^k)$
5: Choose a_{k+1} from $f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}}\|\nabla f(y^k)\|_2^2 = f(x^{k+1})$
6: $A_{k+1} = A_k + a_{k+1}$
7: $v^{k+1} = v^k - a_{k+1}\nabla f(y^k)$
8: **end while**

Accelerated Alternating Minimization

Algorithm 3 Accelerated Alternating Minimization (AAM)

Input: Starting point x_0.
Output: x^N

1: Set $A_0 = 0$, $x^0 = v^0$.
2: **while** $k \leq N - 1$ **do**
3: $\beta_k = \operatorname{argmin}_{\beta \in [0,1]} f\left(x^k + \beta(v^k - x^k)\right)$
$y^k = x^k + \beta_k(v^k - x^k)$
4: $i_k = \underset{i \in \{1,\ldots,n\}}{\operatorname{argmax}} \|\nabla_i f(y^k)\|_2^2$
$x^{k+1} = \underset{x \in S_{i_k}(y^k)}{\operatorname{argmin}} f(x)$
5: Choose a_{k+1} from $f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}} \|\nabla f(y^k)\|_2^2 = f(x^{k+1})$
6: $A_{k+1} = A_k + a_{k+1}$
7: $v^{k+1} = v^k - a_{k+1} \nabla f(y^k)$
8: **end while**

Accelerated Alternating Minimization

Accelerated alternating minimization method Algorithm 3:

- Uses a univariate minimization
- Greedy approach to determine the block which is updated, unlike how it is usually done in random coordinate descent methods
- If f is convex, our method enjoys the accelerates $O(1/k^2)$ rate for the objective residual and, for a general setting of possibly non-convex functions it guarantees that the squared norm of the gradient decreases as $O(1/k)$
- The original method is the same for the both settings meaning that it is uniformly optimal for convex and non-convex optimization
- Also our method possesses linear convergence rate under PL condition

Theorem

a) Assume that f is (possibly non-convex) L-smooth function w.r.t. $\|\cdot\|_2$. Then, after k steps of Algorithm 3,

$$\min_{i=0,\dots,k} \|\nabla f(y^i)\|_2^2 \leqslant \frac{2nL(f(x^0) - f(x_*))}{k}. \tag{14}$$

b) If f additionally satisfies Polyak-Łojasiewicz condition Algorithm 3 started with $\mu = 0$ possesses linear convergence:

$$f(x^{k+1}) - f(x^*) \leqslant \Pi_{i=0}^{k-1} \left(1 - \frac{\mu}{L}\right) \cdot (f(x^0) - f(x^*)), \tag{15}$$

c) Assume additionally that f is strongly convex with $\mu \geqslant 0$. Then, after k steps of Algorithm 3,

$$f(x^k) - f(x_*) \leqslant nLR^2 \min\left\{\frac{4}{k^2}, \left(1 - \sqrt{\frac{\mu}{nL}}\right)^{k-1}\right\} \tag{16}$$

Implementation Details

But in practice the line search procedure can be computationally expensive, but the proof uses the line search for $\beta_k = \operatorname{argmin}_{\beta \in [0,1]} f\left(x^k + \beta(v^k - x^k)\right)$ only to satisfy

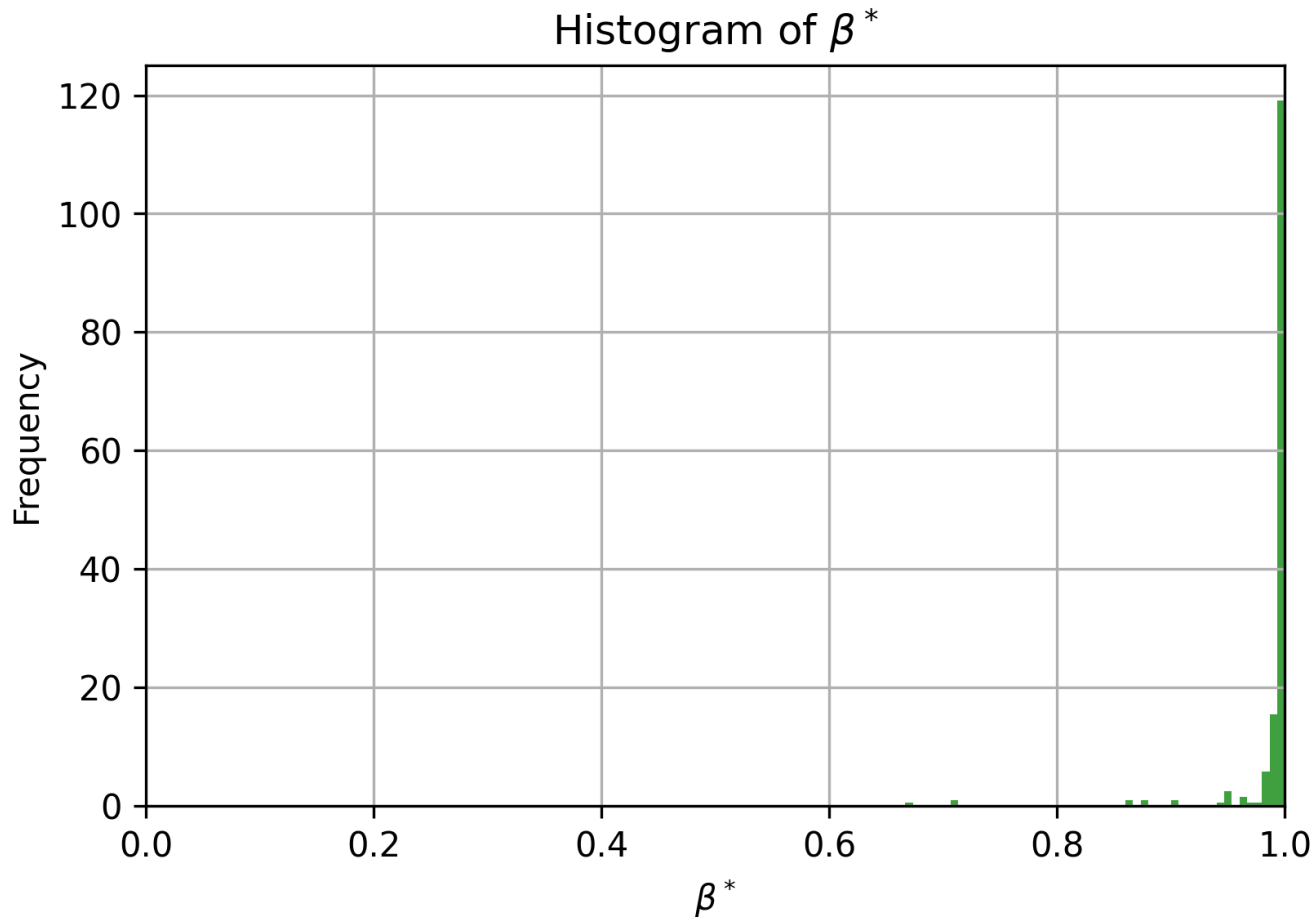
$$\langle \nabla f(y^k), v^k - y^k \rangle \geqslant 0$$

and

$$f(y^k) \leqslant f(x^k)$$

so these conditions are used as a stopping criterion.

Moreover, appropriate value of β^* are typically close to 1, $\frac{k}{k+3}$ – are used is a starting point



So the algorithm needs 2-3 iteration in average of the line search subroutine.

Numerical results

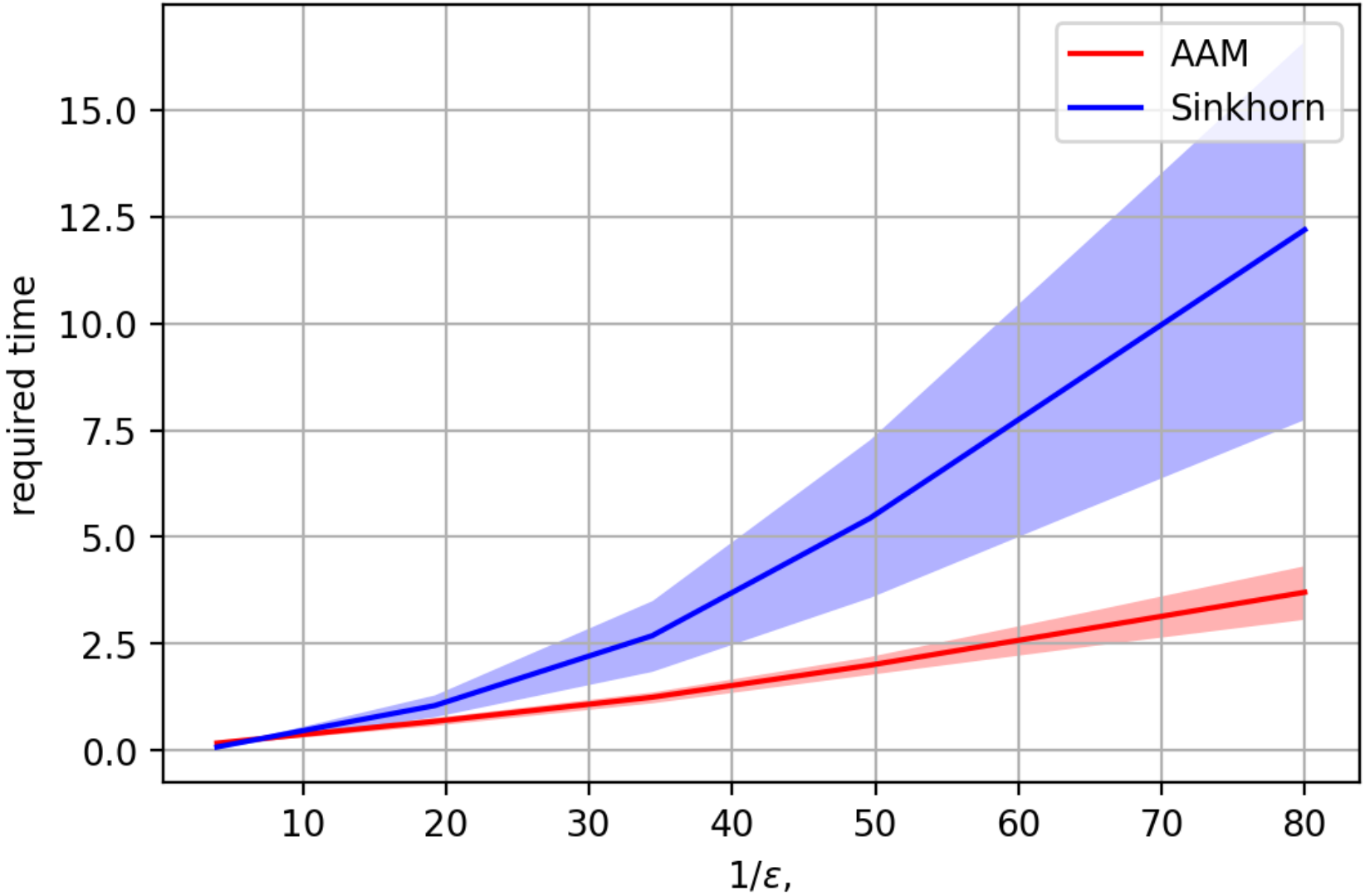


Figure: Performance comparison between multimarginal Sinkhorn's algorithm and Algorithm 3 ($n = 15$, $m = 4$)

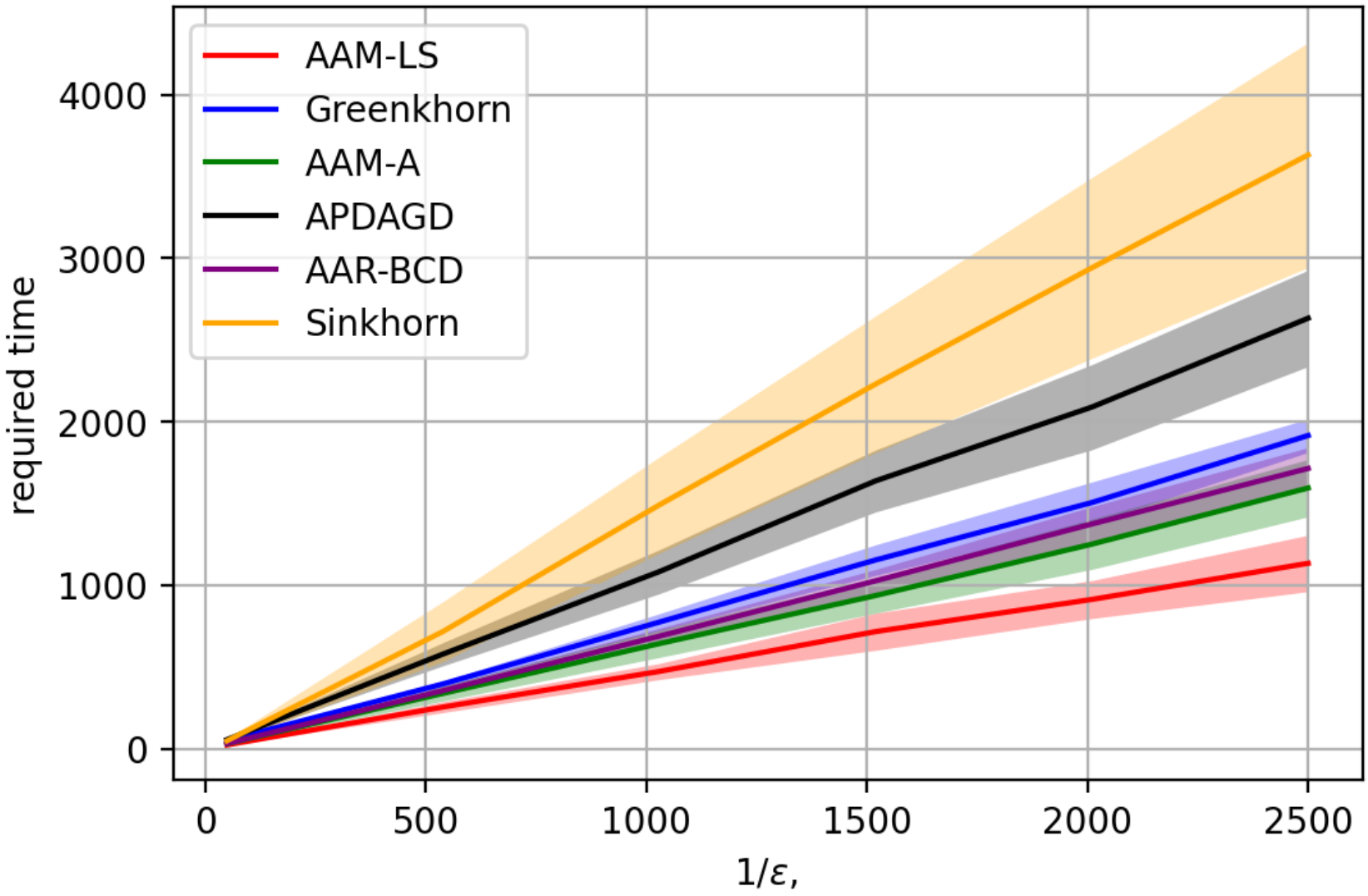


Figure: Performance comparison between on the MNIST dataset. Filled in area corresponds to 1 standard deviation.

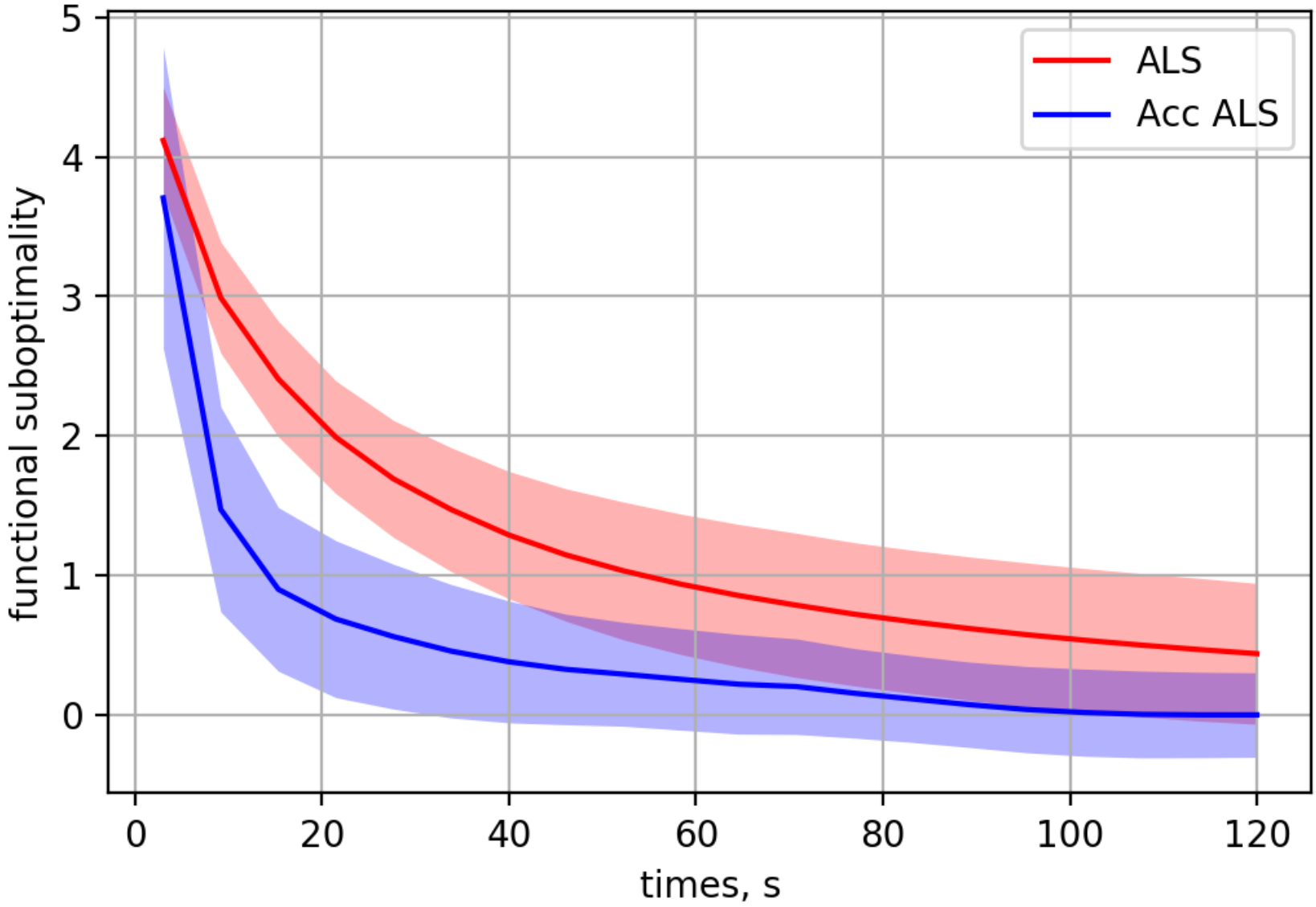


Figure: Performance of the Algorithm 1 and Algorithm 3 applied to the problem (10)

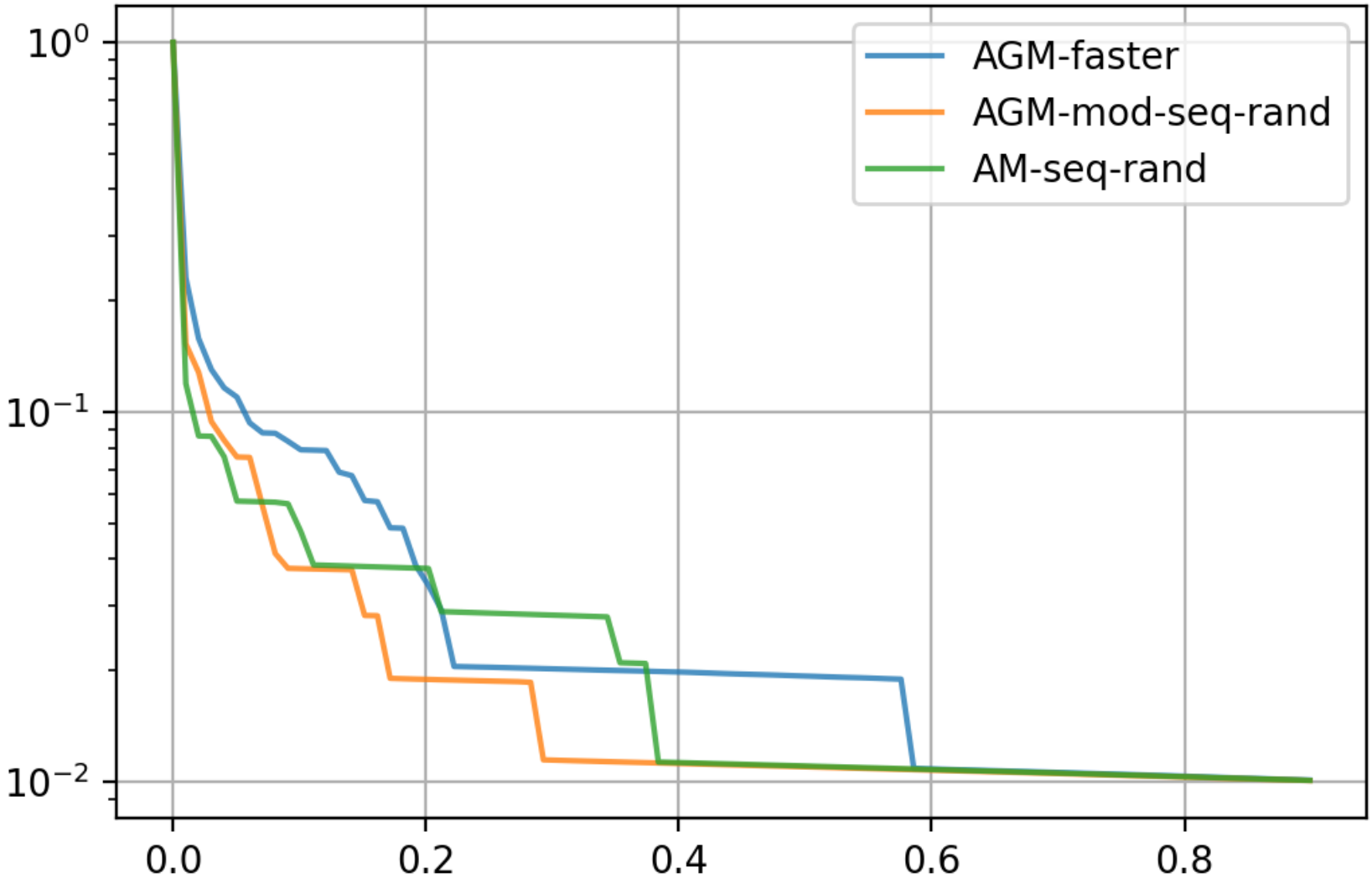


Figure: Performance of the Algorithm 1 and Algorithm 3 applied to the problem (11)

AAM Modification

Algorithm 4 Accelerated Alternating Minimization (AAM)

Input: Starting point x_0.

Output: x^N

1: Set $A_0 = 0$, $x^0 = v^0$.

2: **while** $k \leq N - 1$ **do**

3: $\beta_k = \operatorname{argmin}_{\beta \in [0,1]} f\left(x^k + \beta(v^k - x^k)\right)$

$y^k = x^k + \beta_k(v^k - x^k)$

4: $i_k = \operatorname{argmax}_{i \in \{1,\ldots,n\}} \|\nabla_i f(y^k)\|_2^2$ {Costs: $O(\cdot)$}

$x^{k+1} = \operatorname{argmin}_{x \in S_{i_k}(y^k)} f(x)$ {Costs: $O(\frac{\cdot}{n})$}

$x^{k+1} = \operatorname{argmin}_{x \in S_{i \neq i_k}(x^{k+1})} f(x)$ {Costs: $O(\frac{\cdot}{n})$}

5: Choose a_{k+1} from $f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}} \|\nabla f(y^k)\|_2^2 = f(x^{k+1})$

6: $A_{k+1} = A_k + a_{k+1}$

7: $v^{k+1} = v^k - a_{k+1} \nabla f(y^k)$

8: **end while**

Comparison of AGMsDR and its Alternating Modification

The test problem

$$\min_z f(z) = \|Wz - b\|_2^2, \tag{17}$$

and its alternating modification

$$\min_{x,y} \|Ax + By - c\|_2^2 + \|Cx + Dy - d\|_2^2, \tag{18}$$

where

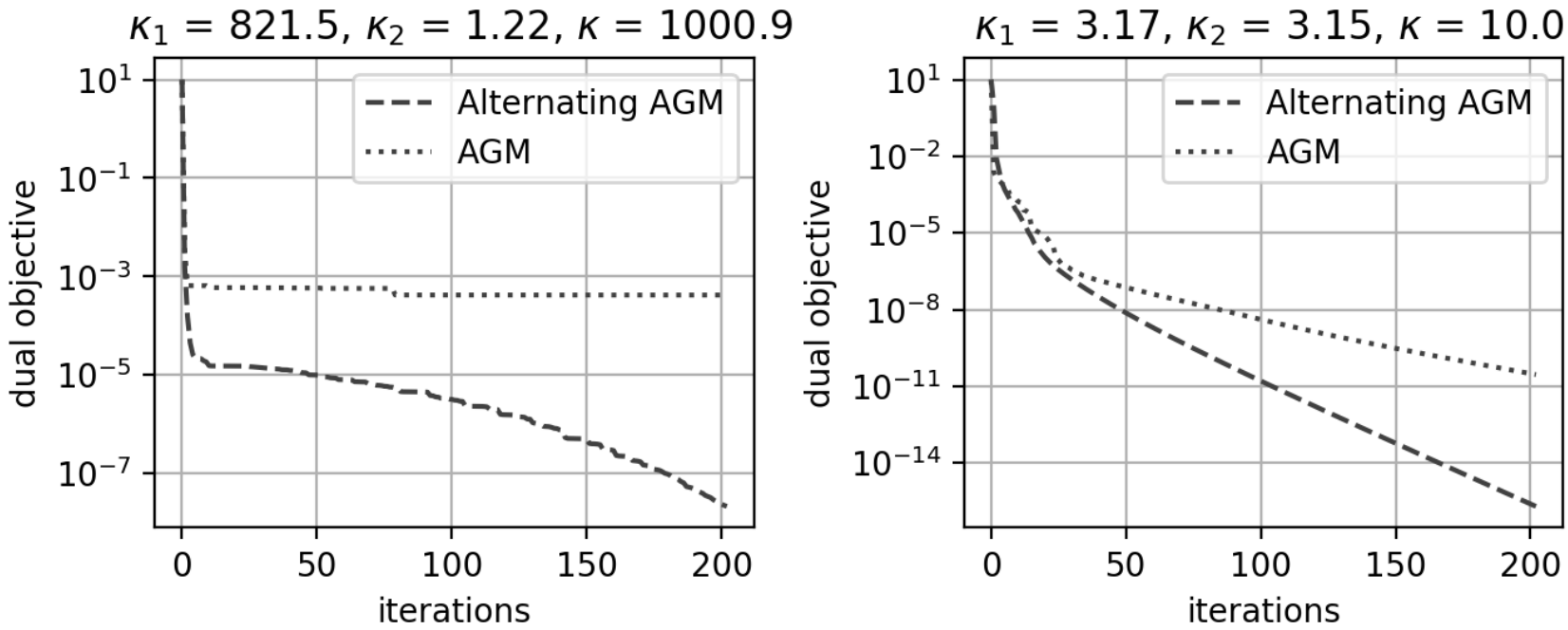
$$z = \begin{pmatrix} x \\ y \end{pmatrix} \quad W = \begin{pmatrix} AB \\ CD \end{pmatrix} \quad b = \begin{pmatrix} c \\ d \end{pmatrix}.$$

AM:

$$x^{k+1} = (A^T A + C^T C)^{-1}\left[A^T(c - By^k) + C^T(d - Dy^k)\right]$$
$$y^{k+1} = (B^T B + D^T D)^{-1}\left[B^T(c - Ax^k) + D^T(d - Cx^k)\right]$$

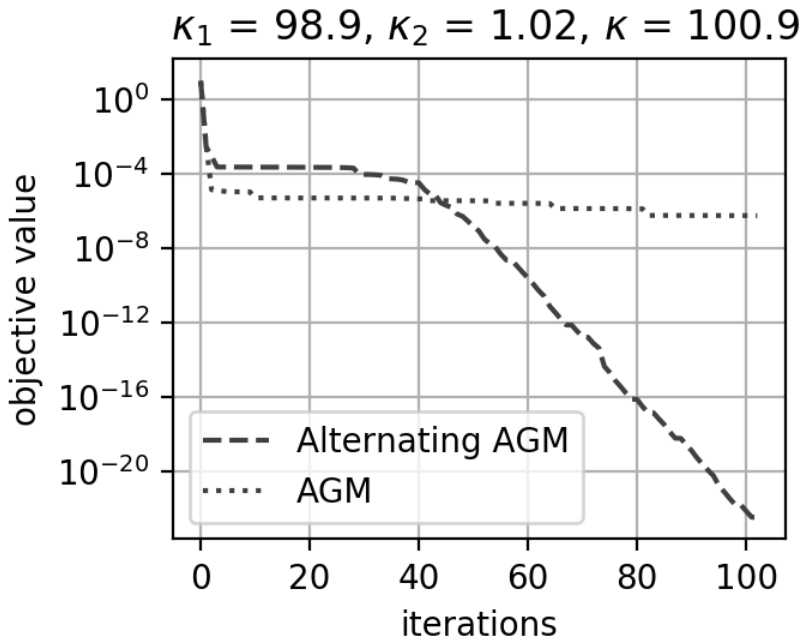
Comparison

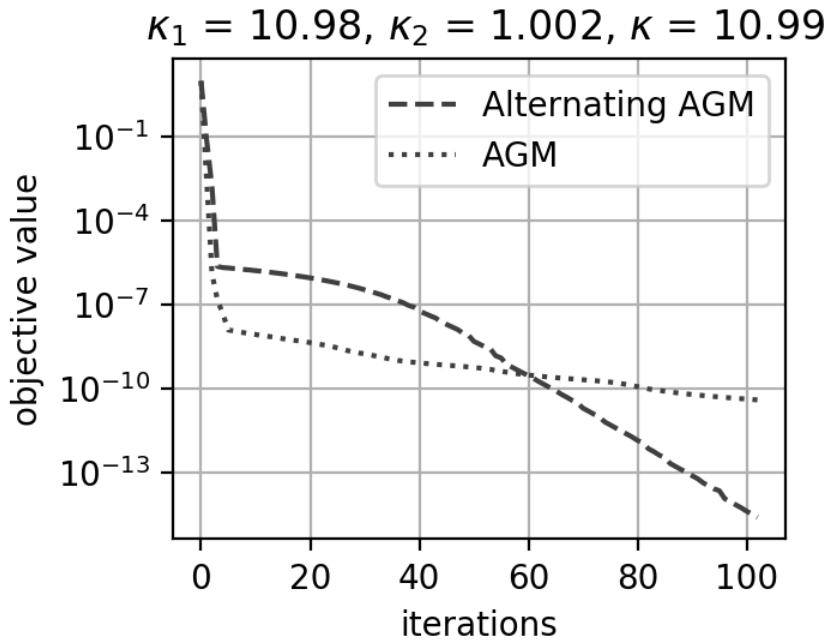
Comparison of AGMsDR and Alternating AGMsDR, started with $\mu = 0$



Comparison

Comparison of AGMsDR and Alternating AGMsDR, started with $\mu = 0$.





Thank you for your attention!

References I

[Benamou et al., 2015] Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015).
Iterative bregman projections for regularized transportation problems.
SIAM Journal on Scientific Computing, 37(2):A1111–A1138.

[Bertsekas and Tsitsiklis, 1989] Bertsekas, D. P. and Tsitsiklis, J. N. (1989).
Parallel and Distributed Computation: Numerical Methods.
Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Cuturi, 2013] Cuturi, M. (2013).
Sinkhorn distances: Lightspeed computation of optimal transport.
In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc.

References II

[Hong et al., 2016] Hong, M., Razaviyayn, M., Luo, Z., and Pang, J. (2016).
A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing.
IEEE Signal Processing Magazine, 33(1):57–77.

[Khenissi and Nasraoui, 2019] Khenissi, S. and Nasraoui, O. (2019).
Modeling and counteracting exposure bias in recommender systems.

References III

[Kroshnin et al., 2019] Kroshnin, A., Tupitsa, N., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., and Uribe, C. A. (2019).
On the complexity of approximating wasserstein barycenters.
In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3530–3540.

[Nesterov et al., 2020] Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. (2020).
Primal-dual accelerated gradient methods with small-dimensional relaxation oracle.
Optimization Methods and Software, pages 1–28.
arXiv:1809.05895.

References IV

[Ortega and Rheinboldt, 2000] Ortega, J. M. and Rheinboldt, W. C. (2000).
Iterative Solution of Nonlinear Equations in Several Variables.
Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[Sinkhorn, 1974] Sinkhorn, R. (1974).
Diagonal equivalence to matrices with prescribed row and column sums. II.
Proc. Amer. Math. Soc., 45:195–198.