Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Alternating minimization methods for strongly convex optimization

Nazarii Tupitsa[1,2], Pavel Dvurechensky[3], Alexander Gasnikov[1,2], Sergey Guminov[1,2]

submitted: February 19, 2020

[1] Moscow Institute of Physics and Technology
Institutskiy Pereulok, 9
Dolgoprudny, Moscow Region
141701 Russian Federation
E-Mail: tupitsa@phystech.edu
gasnikov@yandex.ru
sergey.guminov@phystech.edu

[2] Institute for Information Transmission Problems of RAS
Bolshoy Karetny per. 19, build.1
127051 Moscow
Russian Federation

[3] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: pavel.dvurechensky@wias-berlin.de

No. 2692
Berlin 2020

2010 *Mathematics Subject Classification.* 90C30, 90C25, 68Q25.

Key words and phrases. Convex optimization, alternating minimization, block-coordinate method, complexity analysis.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Alternating minimization methods for strongly convex optimization

Nazarii Tupitsa, Pavel Dvurechensky, Alexander Gasnikov, Sergey Guminov

Abstract

We consider alternating minimization procedures for convex optimization problems with variable divided in many block, each block being amenable for minimization with respect to its variable with freezed other variables blocks. In the case of two blocks, we prove a linear convergence rate for alternating minimization procedure under Polyak-Łojasiewicz condition, which can be seen as a relaxation of the strong convexity assumption. Under strong convexity assumption in many-blocks setting we provide an accelerated alternating minimization procedure with linear rate depending on the square root of the condition number as opposed to condition number for the non-accelerated method.

1 Introduction

In this paper we consider unconstrained minimization problem

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

where $f(x)$ is a smooth convex function with L-Lipschitz-continuous gradient. Further, our main assumption is that the space $\mathbb{R}^m$ can be divided into n disjoint subspaces $L_i \in \mathbb{R}^m$, s.t. $\cup L_i = \mathbb{R}^m$ and it is possible to minimize the objective f in each block if the variables in all other blocks are fixed. Moreover, we are mostly interested in obtaining linear convergence rate and sufficient conditions for it.

To be exact, we suppose that f has a block structure, i.e. $f(x) = f(x_1, \ldots, x_n)$, and we know exact expression for the minimizer
$x_i^* = \operatorname{argmin}_{z \in \mathbb{R}^{n_i}} f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)$, $\sum n_i = m$.

A very old and natural idea under this assumption is to use alternating minimization procedure [13, 3], where the objective is minimized sequentially in each subspace. First of all, we are interested in the convergence rate analysis of this type of algorithms. For smooth strongly convex problems under some additional technical assumptions, the linear rate was obtained in [9]. [1] analyze alternating minimization procedure for the case of two blocks in the general convex setting. The underlying assumption is presence of a smooth component in at least one block of variables. Also the non-smoothness is possible via composite terms which do not ruin the block minimizability property. Since there is no strong convexity assumption, the obtained convergence rate us sublinear, namely $O(1/k)$, where k is the iteration counter. Similar result, but for many-block setting wan obtained in [8, 15]. In the fully smooth setting under strong convexity assumption [12] obtain linear rate of convergence also for the many-block setting. This linear rate is proportional to κ – efficient condition number of the problem. [4] provide an accelerated alternating minimization method for a vary special problem with two block having the form of a sum of a quadratic function with two proximally friendly composite terms. The obtained convergence rate is $O(1/k^2)$ for convex setting and is linear with exponent $\sqrt{\kappa}$ in the strongly

DOI 10.20347/WIAS.PREPRINT.2692 Berlin 2020

convex case. [6] analyze a non-accelerated alternating minimization method and obtain $O(1/k)$ convergence rate in the convex setting and linear rate with exponent κ for strongly convex case. They also propose an accelerated method for general convex setting with rate $O(1/k^2)$ and conjecture that their analysis can be extended for the strongly convex case. We also mention the review [7].

In this paper we, firstly, focus on obtaining linear rate of convergence for non-accelerated method with the exponent κ in a more general setting of Polyak-Łojasiewicz condition [14]. This assumption is weaker than the strong convexity assumption since it follows from the strong convexity. Secondly, we propose an accelerated alternating minimization method for general smooth objective functions in the many-blocks setting. For this method we obtain accelerated convergence rate

$$O\left(\min\left\{\frac{1}{k^2}, \left(1-\sqrt{\kappa}\right)^k\right\}\right).$$

2 Simple alternating minimization algorithm and notation

Consider for simplicity alternating minimization algorithm for the problem with only two block structure. All the following results and myproofs can be easily extended for any number of blocks.

Algorithm 1 Alternating Minimization

Require: Starting point x_0.
Ensure: x^k
1: Set x^0.
2: **for** $k \geqslant 0$ **do**
3: **if** $k \bmod 2 = 0$ **then**
4: $x_1^{k+1} = \operatorname{argmin}_{z\in\mathbb{R}^{n_1}} f(z, x_2^k)$
5: **else**
6: $x_2^{k+1} = \operatorname{argmin}_{z\in\mathbb{R}^{n_2}} f(x_1^{k+1}, z)$
7: **end if**
8: **end for**

Optimality conditions for algorithm's minimization problems reads as follows:

$$\nabla_1 f(x_1^{k+1}, x_2^k) = 0 \tag{2}$$
$$\nabla_2 f(x_1^k, x_2^k) = 0 \tag{3}$$

Introduce the following notation:

$$x^k = (x_1^k, x_2^k), \quad x^{k+\frac{1}{2}} = (x_1^{k+1}, x_2^k)$$

$$T_M(x) = (T_M^1(x), T_M^2(x)) \quad G_M(x) = (G_M^1(x), G_M^2(x))$$

$$T_{M_i}^i(x) = \operatorname{prox}_{\frac{1}{M_i}}\left(x_i - \frac{1}{M_i}\nabla_i f(x)\right), \tag{4}$$

$$G_{M_i}^i(x) = M_i(x_i - T_{M_i}^i(x))$$

For the case $i = 1$

$$T^1_{M_1}(x^k) = \underset{u\in\mathbb{R}^{n_1}}{\operatorname{argmin}} \left(\frac{M_1}{2}\|u - (x^k_1 - \frac{1}{M_1}\nabla_1 f(x^k))\|_2^2 \right) = \\ = x^k_1 - \frac{1}{M_1}\nabla_1 f(x^k).$$

Lemma 1

$$G^1_M(x^{k+\frac{1}{2}}) = 0, \quad G^2_M(x^k) = 0$$

for all k*.*

proof:
$T^2_M(x^k) = \operatorname{argmin}_{v\in R}\left(\frac{M}{2}\|v - x^k_2\|_2^2 + \langle\nabla_2 f(x^k), v - x^k_2\rangle\right)$
$= x^k_2$, since $\nabla_2 f(x^k_1, x^k_2) = 0$ by (3).

$$T_M(x^k) = \left(T^1_M(x^k), T^2_M(x^k)\right) = \left(T^1_M(x^k), x^k_2\right) \\ = \left(T^1_M(x^k), x^k_2 - \frac{1}{M}G^2_M(x^k)\right),$$

where the last equality follows from the definition of $G^2_M(x^k)$.

3 Sufficient decrease-type result

The following result can be found in the 14-th chapter of [2] or in [10]

$$\|G^2_{M_2}(x^{k+\frac{1}{2}})\|_2^2 \leqslant 2L_2\left(f(x^{k+\frac{1}{2}}) - f(x^{k+1})\right) \tag{5}$$

$$\|G^1_{M_1}(x^k)\|_2^2 \leqslant 2L_1\left(f(x^k) - f(x^{k+\frac{1}{2}})\right) \tag{6}$$

where again we suppose that constant L_1 and L_2 can be different for different blocks:

$$f(u,v) \leqslant f(\xi,\eta) + \langle\nabla_1 f(\xi,\eta), u - \xi\rangle + \langle\nabla_2 f(\xi,\eta), v - \eta\rangle \\ + \frac{L_1}{2}\|u - \xi\|_2 + \frac{L_2}{2}\|v - \eta\|_2,$$

and the the constant in the regular definition of Lipshitz continuity of the gradient of f is described by $L = \max(L_1, L_2)$.

4 Polyak-Łojasiewicz condition

Our myproof of the convergence rate demands Polyak-Łojasiewicz (PL) condition, that can be satisfied for variety of problems. Next we show, that (PL) condition follows from the strong convexity of f

Lemma 2 *Strong convexity of f implies PL conditions:*

$$f(x^*) \geqslant f(x^k) - \frac{1}{2\mu_1}\|G^1_{M_1}(x^k)\|_2^2 \tag{7}$$

$$f(x^*) \geqslant f(x^{k+\frac{1}{2}}) - \frac{1}{2\mu_2}\|G^2_{M_2}(x^{k+\frac{1}{2}})\|_2^2 \tag{8}$$

proof: Since $T^1_{M_1}(x)$ is a minimizer of $\frac{M_1}{2}\|u - x_1^k\|_2^2 + \langle \nabla_1 f(x^k), u - x_1^k\rangle$, w.r.t. $u \in \mathbb{R}^{n_1}$

$$\nabla_1 f(x^k) + M_1(T^1_{M_1}(x^k) - x_1^k) = 0$$

or equivalently

$$\nabla_1 f(x^k) = G^1_{M_1}(x^k) \tag{9}$$

We suppose, that strong convexity parameter can different for subspase $\mathbb{R}^{n_1}$ and subspace $\mathbb{R}^{n_2}$

$$f(u,v) \geqslant f(\xi,\eta) + \langle \nabla_1 f(\xi,\eta), u - \xi\rangle + \langle \nabla_2 f(\xi,\eta), v - \eta\rangle + \\ \frac{\mu_1}{2}\|u - \xi\|_2 + \frac{\mu_2}{2}\|v - \eta\|_2,$$

but the regular definition can be written with $\mu = \min(\mu_1, \mu_2)$.

Strong convexity of f implies the first inequality in the following:

$$\begin{aligned} &f(u,v) \geqslant f(x_1^k, x_2^k) + \\ &+ \langle \nabla_1 f(x_1^k, x_2^k), u - x_1^k\rangle + \langle \nabla_2 f(x_1^k, x_2^k), v - x_2^k\rangle + \\ &\frac{\mu_1}{2}\|u - x_1^k\|_2 + \frac{\mu_2}{2}\|v - x_2^k\|_2 \overset{①}{\geqslant} \\ &\overset{①}{\geqslant} \min_u \left\{ f(x^k) + \langle G^1_{M_1}(x^k), u - x_1^k\rangle + \frac{\mu_1}{2}\|u - x_1^k\|_2 \right\} \overset{②}{=} \\ &= f(x^k) - \frac{1}{2\mu_1}\|G^1_{M_1}(x^k)\|_2, \end{aligned}$$

where ① since $\|\cdot\| \geqslant 0$, (3) and (9). Plugging in $(u,v) = (x_1^*, x_2^*)$

$$f(x^*) \geqslant f(x^k) - \frac{1}{2\mu_1}\|G^1_{M_1}(x^k)\|_2^2$$

The same derivation for the point $x^{k+\frac{1}{2}} = (x_1^{k+1}, x_2^k)$ gives the other condition

$$f(x^*) \geqslant f(x^{k+\frac{1}{2}}) - \frac{1}{2\mu_2}\|G^2_{M_2}(x^{k+\frac{1}{2}})\|_2^2$$

5 Convergence rate

Combining (5), (6) and (7), (8) and the definition of μ from the lemma 2 we get

$$\mu \left(f(x^k) - f(x^*)\right) \leqslant \mu_1 \left(f(x^k) - f(x^*)\right) \leqslant L_1 \left(f(x^k) - f(x^{k+\frac{1}{2}})\right)$$

$$\left(f(x^{k+\frac{1}{2}}) - f(x^*)\right) \leqslant (1 - \frac{\mu_1}{L_1}) \left(f(x^k) - f(x^*)\right)$$

The same we have for the second block

$$\left(f(x^{k+1}) - f(x^*)\right) \leqslant (1 - \frac{\mu_2}{L_2}) \left(f(x^{k+\frac{1}{2}}) - f(x^*)\right)$$

By combining these inequalities we get

$$\left(f(x^{k+1}) - f(x^*)\right) \leqslant (1 - \frac{\mu_2}{L_2}) \left(f(x^{k+\frac{1}{2}}) - f(x^*)\right) \left(f(x^{k+\frac{1}{2}}) - f(x^*)\right) \leqslant (1 - \frac{\mu_1}{L_1}) \left(f(x^k) - f(x^*)\right)$$

$$\left(f(x^{k+1}) - f(x^*)\right) \leqslant (1 - \frac{\mu_2}{L_2})(1 - \frac{\mu_1}{L_1}) \left(f(x^k) - f(x^*)\right)$$

or for regular definition of PL condition

$$\left(f(x^{k+1}) - f(x^*)\right) \leqslant (1 - \frac{\mu}{L_2})(1 - \frac{\mu}{L_1}) \left(f(x^k) - f(x^*)\right)$$

$$\text{notice that } (1 - \frac{\mu}{L_{max}}) \leqslant 1$$

$$\leqslant (1 - \frac{\mu}{L_{min}}) \left(f(x^k) - f(x^*)\right).$$

6 Accelerated Alternating Minimization

In this section we describe accelerated method for alternating minimization, which is originates in [11]. But before notice, that algorithm 1 does not use the constant of strong convexity and consequently adapts to strong convexity of the problem. If the problem is not strongly convex or PL condition is not satisfied the algorithm 1 will poses the following convergence rate $f(x^k) - f_{opt} \leqslant \max\left\{\frac{f(x_0)-f_{opt}}{2^{(k-1)/2}}, \frac{8\min(L_1,L_2)R^2}{k-1}\right\}$. The proof can be found in [2]. The following algorithm requires the knowing of the parameter μ of strong convexity. But it is possible to use this method with $\mu = 0$.

Algorithm 2 Accelerated Alternating Minimization (AAM)

Require: Starting point x_0.

Ensure: x^k

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

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

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

4: Set $y^k = x^k + \beta_k(v^k - x^k)$ {Extrapolation step}

5: Choose $i_k = \underset{i \in \{1,\dots,n\}}{\operatorname{argmax}} \|\nabla_i f(y^k)\|_2$

6: Set $x^{k+1} = \underset{x \in S_{i_k}(y^k)}{\operatorname{argmin}} f(x)$ {Block minimization}

7: If L is known choose a_{k+1} s.t. $\frac{a_{k+1}^2}{(A_k + a_{k+1})(\tau_k + \mu a_{k+1})} = \frac{1}{Ln}$
If L is unknown, find largest a_{k+1} from the equation

$$f(y^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})(\tau_k + \mu a_{k+1})}\|\nabla f(y^k)\|_2^2 + \frac{\mu \tau_k a_{k+1}}{2(A_k + a_{k+1})(\tau_k + \mu a_{k+1})}\|v^k - y^k\|_2^2 = f(x^{k+1})$$

8: Set $A_{k+1} = A_k + a_{k+1}$

9: Set $v^{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^N} \psi_{k+1}(x)$ {Update momentum term}

10: **end for**

We will begin with one key Lemma. Let us introduce an auxiliary functional sequence defined as

$$\psi_0(x) = \frac{1}{2}\|x - x^0\|_2^2,$$

$$\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2}\|x - y^k\|_2^2\}.$$

For

$$l_k(x) = \sum_{i=0}^{k} a_{i+1}\{f(y^i) + \langle \nabla f(y^i), x - y^i \rangle + \frac{\mu}{2}\|x - y^i\|_2^2\}$$

we can write

$$\psi_{k+1}(x) = \psi_0(x) + l_k(x)$$

It is easy to see that $\psi_k(x)$ is τ_k strongly convex function with $\tau_k = 1 + \mu \sum_{i=0}^{k} a_i = 1 + \mu A_k$.

Lemma 3 *After k steps of Algorithm 2 it holds that*

$$A_k f(x^k) \leqslant \min_{x \in \mathbb{R}^N} \psi_k(x) = \psi_k(v^k). \tag{10}$$

Moreover, if the objective is L-smooth and μ-strongly convex $A_k \geqslant \max\left\{\frac{k^2}{4Ln}, \frac{1}{nL}\left(1 - \sqrt{\frac{\mu}{nL}}\right)^{-k-1}\right\}$, where n is the number of blocks.

proof: First, we prove inequality (10) by induction over k. For $k = 0$, the inequality holds. Assume that

$$A_k f(x^k) \leqslant \min_{x\in\mathbb{R}^N} \psi_k(x) = \psi_k(v^k).$$

Then

$$\begin{aligned}\psi_{k+1}(v^{k+1}) = \min_{x\in\mathbb{R}^N}\Big\{\psi_k(x) + a_{k+1}\{f(y^k) + \langle\nabla f(y^k), x \\ - y^k\rangle + \frac{\mu}{2}\|x - y^k\|_2^2\}\Big\} \geqslant \min_{x\in\mathbb{R}^N}\Big\{\psi_k(v^k) + \frac{\tau_k}{2}\|x - v^k\|_2^2 + \\ a_{k+1}\{f(y^k) + \langle\nabla f(y^k), x - y^k\rangle + \frac{\mu}{2}\|x - y^k\|_2^2\}\Big\} \\ \geqslant \min_{x\in\mathbb{R}^N}\Big\{A_k f(x^k) + \frac{\tau_k}{2}\|x - v^k\|_2^2 + a_{k+1}\{f(y^k) + \langle\nabla f(y^k), x \\ - y^k\rangle + \frac{\mu}{2}\|x - y^k\}\Big\}\end{aligned}$$

Here we used that ψ_k is a strongly convex function with minimum at v^k and that $f(y^k) \leqslant f(x^k)$.

By the optimality conditions for the problem $\min_{\beta\in[0,1]} f\left(x^k + \beta(v^k - x^k)\right)$, either

1 $\beta_k = 1$, $\langle\nabla f(y^k), x^k - v^k\rangle \geqslant 0$, $y^k = v^k$;

2 $\beta_k \in (0, 1)$ and $\langle\nabla f(y^k), x^k - v^k\rangle = 0$, $y^k = v^k + \beta_k(x^k - v^k)$;

3 $\beta_k = 0$ and $\langle\nabla f(y^k), x^k - v^k\rangle \leqslant 0$, $y^k = x^k$.

In all three cases, $\langle\nabla f(y^k), v^k - y^k\rangle \geqslant 0$.

Thus

$$\begin{aligned}\psi_{k+1}(v^{k+1}) \geqslant \min_{x\in\mathbb{R}^N}\Big\{A_k f(y^k) + \frac{\tau_k}{2}\|x - v^k\|_2^2 + \\ a_{k+1}\{f(y^k) + \langle\nabla f(y^k), x - y^k\rangle + \frac{\mu}{2}\|x - y^k\|_2^2\}\Big\}.\end{aligned}$$

The explicit solution to the above quadratic optimization problem is

$$x = \frac{1}{\tau_{k+1}}(\tau_k v^k + \mu a_{k+1} y^k - a_{k+1}\nabla f(y^k))$$

By plugging in the solution and using $\langle\nabla f(y^k), v^k - y^k\rangle \geqslant 0$, we obtain

$$\begin{aligned}\psi_{k+1}(v^{k+1}) \geqslant A_{k+1} f(y^k) - \frac{a_{k+1}^2}{2\tau_{k+1}}\|\nabla f(y^k)\|_2^2 + \\ \frac{\mu\tau_k a_{k+1}}{2\tau_{k+1}}\|v^k - y^k\|_2^2.\end{aligned}$$

Our next goal is to show that

$$A_{k+1}f(y^k) - \frac{a_{k+1}^2}{2\tau_{k+1}}\|\nabla f(y^k)\|_2^2 + \frac{\mu\tau_k a_{k+1}}{2\tau_{k+1}}\|v^k - y^k\|_2^2$$
$$\geqslant A_{k+1}f(x^{k+1})$$

which proves the induction step.

To do this, by the L-smoothness of the objective, we have $\forall i$

$$f(y^k) - \frac{1}{2L}\|\nabla_i f(y^k)\|_2^2 \geqslant f(x_i^{k+1}),$$

where $x_i^{k+1} = \operatorname{argmin}_{x\in S_i} f(x)$. Since
$i_k = \operatorname{argmax}_i \|\nabla_i f(y^k)\|_2^2$,

$$\|\nabla_{i_k} f(y^k)\|_2^2 \geqslant \frac{1}{n}\|\nabla f(y^k)\|_2^2$$

and $f(y^k) - \frac{1}{2Ln}\|\nabla f(y^k)\|_2^2 \geqslant f(y^k) - \frac{1}{2L}\|\nabla_{i_k} f(y^k)\|_2^2 \geqslant f(x^{k+1})$, Choosing a_{k+1} such that $\frac{a_{k+1}^2}{2A_{k+1}\tau_{k+1}} \geqslant \frac{1}{2Ln}$ implies

$$A_{k+1}f(y^k) - \frac{a_{k+1}^2}{2\tau_{k+1}}\|\nabla f(y^k)\|_2^2 + \frac{\mu\tau_k a_{k+1}}{2\tau_{k+1}}\|v^k - y^k\|_2^2$$
$$\geqslant A_{k+1}f(y^k) - \frac{a_{k+1}^2}{2\tau_{k+1}}\|\nabla f(y^k)\|_2^2$$
$$\geqslant A_{k+1}f(y^k) - \frac{A_{k+1}}{2Ln}\|\nabla f(y^k)\|_2^2 \geqslant A_{k+1}f(x^{k+1})$$

which proves the induction step.

Rewriting the rule for choosing a_{k+1} gives
$\frac{a_{k+1}^2}{(A_k+a_{k+1})(\tau_k+\mu a_{k+1})} \geqslant \frac{1}{Ln}$. Let us estimate the rate of the growth for A_k. $\tau_k = 1 + \mu\sum_{i=0}^k a_i = 1 + \mu A_k$. $\frac{a_{k+1}^2}{2A_{k+1}\tau_{k+1}} \geqslant \frac{1}{2Ln}$

$$a_k^2 \geqslant \frac{A_k\tau_k}{nL} = \frac{A_k + \mu A_k^2}{nL}$$

$$a_k \geqslant \frac{1}{\sqrt{nL}}\sqrt{A_k + \mu A_k^2} \geqslant \sqrt{\frac{\mu}{2Ln}}A_k \tag{11}$$

$$\sqrt{A_i} - \sqrt{A_{i-1}} \geqslant \frac{A_i - A_{i-1}}{\sqrt{A_i} + \sqrt{A_{i-1}}} \geqslant \frac{a_i}{2\sqrt{A_i}} \geqslant \frac{\sqrt{1+\mu A_i}}{2\sqrt{Ln}}$$

Summing it up for $i = 1, \ldots, k$ we get

$$A_k \geqslant \frac{k^2}{4Ln}$$

We also have

$$A_{k+1} = A_k + a_{k+1} \geqslant A_k + \sqrt{\frac{\mu}{nL}} A_{k+1}$$

which leads to

$$A_{k+1} \geqslant \left(1 - \sqrt{\frac{\mu}{nL}}\right)^{-1} A_k$$

To use this bound we only need to estimate A1 , which we can do as follows:

$$A_1 = \frac{a_1^2}{A_1} \geqslant \frac{a_1^2}{(1+\mu A_1)A1} \geqslant \frac{a_1^2}{A_1 \tau_1} \geqslant \frac{1}{nL}$$

By recursively applying the last bound we reach the desired result:

$$A_k \geqslant \max\left\{\frac{k^2}{4Ln}, \frac{1}{nL}\left(1 - \sqrt{\frac{\mu}{nL}}\right)^{-k+1}\right\}$$

Theorem 1 *After k steps of Algorithm 2 it holds that*

$$f(x^k) - f(x_*) \leqslant nLR^2 \min\left\{\frac{4}{k^2}, \left(1 - \sqrt{\frac{\mu}{nL}}\right)^{k-1}\right\} \tag{12}$$

proof: From the convexity of $f(x)$ we have

$$l_k(x_*) = \sum_{i=0}^{k} a_{i+1}(f(y^i) + \langle \nabla f(y^i), x_* - y^i\rangle + \frac{\mu}{2}\|x_* - y^i\|_2^2) \leqslant A_{k+1} f(x_*).$$

From Lemma (3) we have

$$\begin{aligned} A_k f(x^k) &\leqslant \psi_k(v^k) \\ &\leqslant \psi_k(x_*) = \frac{1}{2}\|x_* - x^0\|_2^2 \\ &+ \sum_{i=0}^{k-1} a_{i+1}(f(y^i) + \langle \nabla f(y^i), x_* - y^i\rangle + \frac{\mu}{2}\|x_* - y^i\|_2^2) \\ &\leqslant A_k f(x_*) + \frac{1}{2}\|x_* - x^0\|_2^2 \end{aligned}$$

$$\begin{aligned} f(x^k) - f(x_*) &\leqslant \frac{R^2}{2A_k} \\ &\leqslant nLR^2 \min\left\{\frac{4}{k^2}, \left(1 - \sqrt{\frac{\mu}{nL}}\right)^{k-1}\right\}. \end{aligned}$$

7 Application to optimal transport

In this section we will be dealing with the discrete optimal transportation problem

$$f(X) = \langle C, X\rangle + \gamma\langle X, \ln X\rangle \to \min_{X\in\mathcal{U}(r,c)}, \tag{13}$$
$$\mathcal{U}(r,c) = \{X \in \mathbb{R}_+^{N\times N} : X\mathbf{1} = r, X^T\mathbf{1} = c\},$$

where X is the transportation plan, $C \in \mathbb{R}_+^{N\times N}$ is a given cost matrix, $r, c \in \mathbb{R}^N$ are given elements of the probability simplex, and $\langle A, B\rangle$ denotes the Frobenius product of matrices defined as $\langle A, B\rangle = \sum\limits_{i,j=1}^{N} A_{ij}B_{ij}$.

To ensure the smoothness of the dual problem, we must dualize the linear constraints $X\mathbf{1} = r, X^T\mathbf{1} = c$ while minimizing the Lagrangian over a closed convex set Q such that the primal function is strongly convex on it. Here $Q = \{X \in \mathbb{R}_+^{N\times N} : \mathbf{1}^T X\mathbf{1} = 1\}$. Then the dual problem is constructed as follows:

$$\begin{aligned}
&\min_{X\in Q\cap\mathcal{U}(r,c)} \langle C, X\rangle + \gamma\langle X, \ln X\rangle \\
&= \min_{X\in Q}\max_{y,z\in\mathbb{R}^n} \langle C, X\rangle + \gamma\langle X, \ln X\rangle + \langle y, X\mathbf{1} - r\rangle \\
&+ \left\langle z, X^T\mathbf{1} - c\right\rangle \\
&= \max_{y,z\in\mathbb{R}^n} -\langle y, r\rangle - \langle z, c\rangle \\
&+ \min_{X\in Q}\sum_{i,j=1}^{n} X^{ij}\left(C^{ij} + \gamma\ln X^{ij} + y^i + z^j\right)
\end{aligned}$$

First of all, we notice that for all i, j and some small ε

$$X^{ij}\left(C^{ij} + \gamma\ln X^{ij} + y^i + z^j\right) < 0$$

for $X^{ij} \in (0, \varepsilon)$ and approaches 0 as X^{ij} approaches 0. Hence, $X^{ij} > 0$ without loss of generality. Using Lagrange multipliers for the constraint $\mathbf{1}^T X\mathbf{1} = 1$, we obtain the problem

$$\min_{X^{ij}>0}\sum_{i,j=1}^{n}\left[X^{ij}\left(C^{ij} + \gamma\ln X^{ij} + y^i + z^j\right)\right] - \nu.$$

The solution to this problem is

$$X^{ij} = \frac{\exp\left(-\frac{1}{\gamma}\left(y^i + z^j + C^{ij}\right) - 1\right)}{\sum_{i,j=1}^{n}\exp\left(-\frac{1}{\gamma}\left(y^i + z^j + C^{ij}\right) - 1\right)}.$$

With a change of variables $u = -y/\gamma - \frac{1}{2}\mathbf{1}, v = -z/\gamma - \frac{1}{2}\mathbf{1}$ we arrive at the following expression for the dual (minimization) problem:

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

where $[B(u, v)]_{ij} = \exp\left(u^i + v^j - \frac{C^{ij}}{\gamma}\right)$ The variables in the dual problem naturally decompose into two blocks u and v. Moreover, minimization over any one block may be performed analytically:

Lemma 4 *Iterations*

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

may be written explicitly as

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

proof:

$$\nabla_u\varphi(u, v^k) = \frac{1}{\mathbf{1}^T B(u, v^k)\mathbf{1}}B(u, v^k)\mathbf{1} - r.$$

From optimality conditions, for u to be the optimal point it is sufficient to have

$$r - \frac{1}{\mathbf{1}^T B(u, v^k)\mathbf{1}}B(u, v^k)\mathbf{1} = 0.$$

Now we check that is, indeed, the case for $u = u^{k+1}$ from the statement of this lemma. We manually check that

$$B(u^{k+1}, v^k)\mathbf{1} = \operatorname{diag}(e^{(u^{k+1}-u^k)})B(u^k, v^k)\mathbf{1} = r,$$

and the conclusion then follows from the fact that

$$\mathbf{1}^T B(u^{k+1}, v^k)\mathbf{1} = \mathbf{1}r = 1.$$

The optimality of v^{k+1} can be proved in the exact same way.

The AM algorithm for this problem with $t = 0$ is the well-known Sinkhorn's algorithm ([5]).

Algorithm 3 Sinkhorn's Algorithm

Ensure: x^k
  for $k \geqslant 1$ **do**
    if $k \bmod 2 = 0$ **then**
      $u^{k+1} = u^k + \ln r - \ln\left(B\left(u^k, v^k\right)\mathbf{1}\right)$
      $v^{k+1} = v^k$
    else
      $u^{k+1} = u^k$
      $v^{k+1} = v^k + \ln c - \ln\left(B\left(u^k, v^k\right)^T\mathbf{1}\right)$
    end if
  end for

Sinkhorn's algorithm can be accelerated using the algorithm 2, since it alternates between two subspaces.

Sinkhorn's algorithm in practice shows the best convergence rate in time (see figure 1, 3), but convergence in iteration worse than AAM (see figure 2). In this paper we make an attempt to understand this behaviour. Sinkhorn's algorithm adapts to strong convexity of the problem, and demonstrate, in fact, linear convergence, in contrast with accelerated methods, which requires parameter μ to be initialized in order to poses linear convergence. So we experimentally found parameter μ with line search, which improved convergence rate of the accelerated algorithm. Such a search is computationally expensive and not applicable in practice, but allows to suppose that faster Sinkhorn's convergence is ensured by strong convexity.

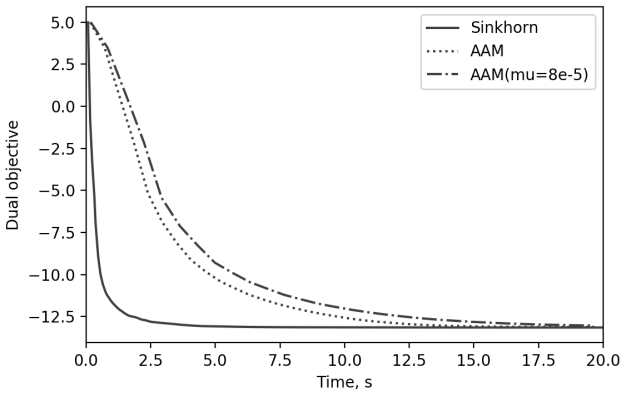

Figure 1: First iteration convergence

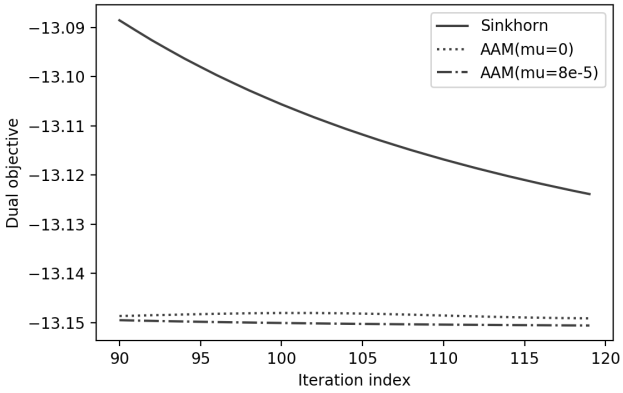

Figure 2: Empirical convergence rate(vs iteration)

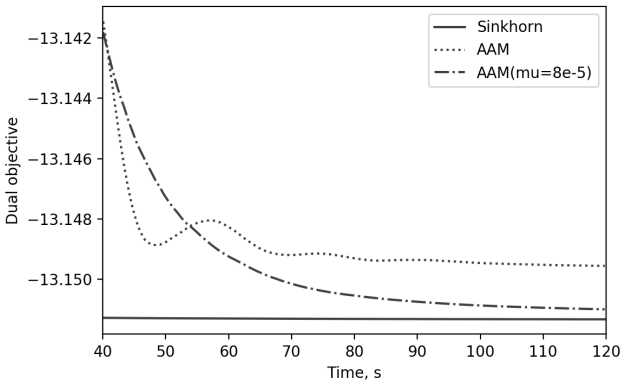

Figure 3: Empirical convergence rate (vs time)

References

[1] Amir. Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25(1):185–209, 2015.

[2] Amir. Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[3] Dimitri P. Bertsekas and John N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[4] Antonin Chambolle, Pauline Tan, and Samuel Vaiter. Accelerated alternating descent methods for dykstra-like problems. *Journal of Mathematical Imaging and Vision*, 59(3):481–497, Nov 2017.

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

[6] Jelena Diakonikolas and Lorenzo Orecchia. Alternating randomized block coordinate descent. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1224–1232, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[7] M. Hong, M. Razaviyayn, Z. Luo, and J. Pang. 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, Jan 2016.

[8] Mingyi Hong, Xiangfeng Wang, Meisam Razaviyayn, and Zhi-Quan Luo. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163(1):85–114, May 2017.

[9] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, Mar 1993.

[10] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014.

[11] Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle, 2018.

[12] Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the gauss-southwell rule than random selection. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1632–1641, Lille, France, 07–09 Jul 2015. PMLR.

[13] James M. Ortega and Werner C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[14] Boris Polyak. *Introduction to Optimization*. New York, Optimization Software, 1987.

[15] Ruoyu Sun and Mingyi Hong. Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1306–1314, Cambridge, MA, USA, 2015. MIT Press.