

# Alternating Minimization Methods for Strongly Convex Optimization

Nazarii Tupitsa, Pavel Dvurechensky, Alexander Gasnikov and  
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 freed other variables blocks. In the case of two blocks, we prove a linear convergence rate for alternating minimization procedure under Polyak-Åojasiewicz 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. We also mention an approximating non-negative solution to a linear system of equations  $Ax = y$  with alternating minimization of Kullback-Leibler (KL) divergence between  $Ax$  and  $y$ .

**Keywords.** convex optimization, alternating minimization, block-coordinate method, complexity analysis.

**2010 Mathematics Subject Classification.** 90C25.

## 1 Introduction

In this paper we consider the minimization problem

$$\min_{x \in Q \subset \mathbb{R}^m} f(x), \quad (1.1)$$

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  $\mathcal{L}_i \in \mathbb{R}^{n_i}$ ,  $\sum n_i = m$ , s.t.  $\cup \mathcal{L}_i = \mathbb{R}^m$  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.

---

This research was funded by Russian Science Foundation (project 18-71-10108) and by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) N°075-00337-20-03, project No. 0714-2020-0005.

To be exact, we suppose that  $f$  has a block structure, i.e.  $f(x) = f(x_1, \dots, x_n)$ , and we know exact expression for the minimizer over  $i$ -th block

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

where  $\cup Q_i = Q$ .

A very old and natural idea under this assumption is to use alternating minimization procedure [4, 22], where the objective is minimized sequentially in each block. 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 [17]. In [2] the author 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 non-smoothness is possible via composite terms which still allow the block minimization. Since there is no strong convexity assumption, the obtained convergence rate is sublinear, namely  $O(1/k)$ , where  $k$  is the iteration counter. Similar result, but for many-block setting was obtained in [13, 24]. In the fully smooth setting under strong convexity assumption [21] obtain linear rate of convergence also for the many-block setting. This linear rate is proportional to  $\kappa$  – efficient condition number of the problem. The authors of [7] provide an accelerated alternating minimization method for a very special problem with two blocks 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 convex case. The authors of [9] analyze a non-accelerated alternating minimization method and obtain  $O(1/k)$  convergence rate in the convex setting and linear rate with exponent  $\kappa$  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. Interested readers can look also into the review [12].

In this paper we, firstly, focus on obtaining linear rate of convergence for non-accelerated method with the exponent  $\kappa$  in a more general setting of Polyak-Łojasiewicz condition [23]. 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

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

From the perspective of applications, many existing statistical algorithms can be

derived as alternating minimization of Kullback–Leibler (KL) divergence [8]. These include the expectation maximization (EM) algorithm for likelihood maximization [1, 25], the Bayesian maximum *a posteriori* (MAP) method with gamma-distributed priors [16], the multiplicative algebraic reconstruction technique (MART) [10] and the "simultaneous" MART (SMART) algorithm [5]. Each of these algorithms can be viewed as an algorithm to find an approximate non-negative solution to a linear system  $Ax = y$ . For example, the SMART can be shown to minimize  $\text{KL}(Ax, y)$  [5]. Some other application of optimization to inverse problems can be found in [6, 26, 27].

Another example is a system of nonlinear equations  $g(x) = 0$ , where  $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ,  $m < n$  and there exists some  $\mu$  s.t. for any  $x \in \mathbb{R}^n$

$$\lambda_{\min} \left( \frac{\partial g(x)}{\partial x} \left[ \frac{\partial g(x)}{\partial x} \right]^T \right) \geq \mu.$$

Then the function  $f(x) = \|g(x)\|_2^2$  satisfies Polyak–Łojasiewicz condition [20], and the algorithms analyzed here can be applied and have linear convergence rate.

## 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 proofs can be easily extended for any number of blocks.

Consider alternating minimization Algorithm 1 for the problem

$$\min_{x_1 \in Q_1, x_2 \in Q_2} F(x_1, x_2) \equiv f(x_1, x_2) + g_1(x_1) + g_2(x_2), \quad (2.1)$$

where  $f(x)$  is a smooth convex function with  $L$ -Lipschitz-continuous gradient and each  $g_i(x)$  is a convex possibly non-smooth function.

---

### Algorithm 1 Alternating Minimization (AM)

---

**Input:** Starting point  $x_0$ .

**Output:**  $x^k$

- 1: Set  $x^0$ .
  - 2: **for**  $k \geq 0$  **do**
  - 3:   **if**  $k \bmod 2 = 0$  **then**
  - 4:      $x_1^{k+1} = \operatorname{argmin}_{z \in Q_1} f(z, x_2^k) + g_1(z)$
  - 5:   **else**
  - 6:      $x_2^{k+1} = \operatorname{argmin}_{z \in Q_2} f(x_1^{k+1}, z) + g_2(z)$
  - 7:   **end if**
  - 8: **end for**
-

We introduce the following notation:

$$\begin{aligned} x^k &= (x_1^k, x_2^k), \quad x^{k+\frac{1}{2}} = (x_1^{k+1}, x_2^k) \\ T_M(x) &= (T_M^1(x), T_M^2(x)) \quad G_M(x) = (G_M^1(x), G_M^2(x)) \\ T_M^i(x) &= \operatorname{prox}_{\frac{1}{M}g_i} \left( x_i - \frac{1}{M} \nabla_i f(x) \right), \quad G_M^i(x) = M(x_i - T_M^i(x)) \end{aligned} \quad (2.2)$$

For the case  $i = 1$

$$\begin{aligned} T_M^1(x^k) &= \operatorname{argmin}_{u \in Q_1} \left( g_1(u) + \frac{M}{2} \|u - (x_1^k - \frac{1}{M} \nabla_1 f(x^k))\|^2 \right) = \\ &= \operatorname{argmin}_{u \in Q_1} \left( g_1(u) + \frac{M}{2} \|u - x_1^k\|^2 + \langle \nabla_1 f(x^k), u - x_1^k \rangle \right). \end{aligned}$$

Next we write

$$\begin{aligned} \partial_1 F(x_1^{k+1}, x_2^k) &= \nabla_1 f(x_1^{k+1}, x_2^k) + \partial g_1(x_1^{k+1}) \\ \partial_2 F(x_1^k, x_2^k) &= \nabla_2 f(x_1^k, x_2^k) + \partial g_2(x_2^k), \end{aligned}$$

where  $\partial_1 F(x_1^{k+1}, x_2^k)$  denotes a subgradient of  $F$  w.r.t first block, e.g. such a set  $S$ , that for all  $s \in S$  the following holds

$$F(y, x_2^k) \geq F(x_1^{k+1}, x_2^k) + \langle s, y - x_1^{k+1} \rangle.$$

$\partial_2 F(x_1^k, x_2^k)$  is defined similarly.

Then, optimality conditions can be written as follows:

$$\begin{aligned} \langle \nabla_1 f(x_1^{k+1}, x_2^k), u - x_1^{k+1} \rangle &\geq \langle -\partial g_1(x_1^{k+1}), u - x_1^{k+1} \rangle \\ \langle \nabla_2 f(x_1^k, x_2^k), v - x_2^k \rangle &\geq \langle -\partial g_2(x_2^k), v - x_2^k \rangle \end{aligned} \quad (2.3)$$

for all  $u \in Q_1, v \in Q_2$ .

The following should clarify the notation.

**Lemma 2.1.** *For points, generated by Algorithm 1 the following holds*

$$\begin{aligned} G_M^1(x^{k+\frac{1}{2}}) &= 0, \quad G_M^2(x^k) = 0 \\ T_M^2(x^k) &= x_2^k, \quad T_M^1(x^{k+\frac{1}{2}}) = x_1^{k+1} \end{aligned}$$

for all  $k$ .

*Proof.*

$$T_M^2(x^k) = \operatorname{argmin}_{v \in Q_2} \left( g_2(v) + \frac{M}{2} \|v - x_2^k\|^2 + \langle \nabla_2 f(x^k), v - x_2^k \rangle \right) = x_2^k,$$

The first is that  $\langle \partial g_2(x_2^k) + \nabla_2 f(x_1^k, x_2^k), v - x_2^k \rangle \geq 0$  for all  $v \in Q_2$  by (2.3), second  $\langle \partial \|v - x_2^k\|^2; v - x_2^k \rangle \geq 0$  since  $x_2^k$  is a minimizer of  $\|v - x_2^k\|^2$ . Summing this two inequalities implies optimality condition for  $T_M^2(x^k)$  at the point  $x_2^k$ .

$$G_M^2(x^k) = M(x_2^k - T_M^2(x^k)) = M(x_2^k - x_2^k) = 0,$$

where the last equality follows from the definition of  $G_M^2(x^k)$ .  $\square$

Introduce also the following notation:

$$\mathcal{D}_1(x^k, M)$$

$$\equiv -2M \min_{u \in Q_1} [\langle \nabla_1 f(x^k), u - x_1^k \rangle + \frac{M}{2} \|u - x_1^k\|^2 + g_1(u) - g_1(x_1^k)] \quad (2.4)$$

$$\mathcal{D}_2(x^{k+\frac{1}{2}}, M)$$

$$\equiv -2M \min_{v \in Q_2} [\langle \nabla_2 f(x^{k+\frac{1}{2}}), v - x_2^k \rangle + \frac{M}{2} \|v - x_2^k\|^2 + g_2(v) - g_2(x_2^k)]. \quad (2.5)$$

Notice that  $T_M(x^k)$  and  $T_M(x^{k+\frac{1}{2}})$  are corresponding minimizers of these two above problems.

### 3 Proximal Polyak-Łojasiewicz condition

In this section we prove that strongly convex function satisfies the proximal-PL inequality condition [14].

We suppose, that strong convexity parameter can different for different variable blocks:

$$\begin{aligned} f(u, v) \geq f(\xi, \eta) + \langle \nabla_1 f(\xi, \eta), u - \xi \rangle + \langle \nabla_2 f(\xi, \eta), v - \eta \rangle \\ + \frac{\mu_1}{2} \|u - \xi\|^2 + \frac{\mu_2}{2} \|v - \eta\|^2, \end{aligned}$$

for any  $u, \xi \in Q_1$  and  $v, \eta \in Q_2$ . Notice, that the single variable definition can be written with  $\mu = \min(\mu_1, \mu_2)$ .

The main result of this section reads as follows.

**Theorem 3.1.** *If  $f$  is strongly convex and  $g$  is convex, then  $F(x) = f(x) + g(x)$  satisfies proximal PL-conditions*

$$F^* \geq F(x^k) - \frac{1}{2\mu_1} \mathcal{D}_1(x^k, \mu_1), \quad F^* \geq F(x^{k+\frac{1}{2}}) - \frac{1}{2\mu_2} \mathcal{D}_2(x^{k+\frac{1}{2}}, \mu_2), \quad (3.1)$$

for the points  $x^k$  and  $x^{k+\frac{1}{2}}$ , generated by Algorithm 1.

*Proof.* By the strong convexity of  $f$  we have

$$\begin{aligned} f(u, v) &\geq f(x_1^k, x_2^k) + \langle \nabla_1 f(x_1^k, x_2^k), u - x_1^k \rangle + \langle \nabla_2 f(x_1^k, x_2^k), v - x_2^k \rangle \\ &\quad + \frac{\mu_1}{2} \|u - x_1^k\|^2 + \frac{\mu_2}{2} \|v - x_2^k\|^2 \stackrel{\textcircled{1}}{\geq} \\ &\stackrel{\textcircled{1}}{\geq} f(x_1^k, x_2^k) + \langle \nabla_1 f(x_1^k, x_2^k), u - x_1^k \rangle - \langle \partial g_2(x_2^k), v - x_2^k \rangle + \frac{\mu_1}{2} \|u - x_1^k\|^2 \stackrel{\textcircled{2}}{\geq} \\ &\stackrel{\textcircled{2}}{\geq} f(x_1^k, x_2^k) + \langle \nabla_1 f(x_1^k, x_2^k), u - x_1^k \rangle + g_2(x_2^k) + g_2(v) + \frac{\mu_1}{2} \|u - x_1^k\|^2 \end{aligned}$$

where

- $\textcircled{1}$  by (2.3)
- $\textcircled{2}$  by convexity of  $g_2$

which leads to

$$\begin{aligned} F(u, v) &\geq F(x_1^k, x_2^k) + \langle \nabla_1 f(x_1^k, x_2^k), u - x_1^k \rangle - g_1(x_1^k) + g_1(u) + \frac{\mu_1}{2} \|u - x_1^k\|^2. \end{aligned}$$

Minimizing both sides respect to  $u \in Q_1, v \in Q_2$ ,

$$\begin{aligned} F^* &\geq F(x^k) + \min_u [\langle \nabla_1 f(x^k), u - x_1^k \rangle + \frac{\mu_1}{2} \|u - x_1^k\|^2 + g_1(u) - g_1(x_1^k)] \\ &= F(x^k) - \frac{1}{2\mu_1} \mathcal{D}_1(x^k, \mu_1). \quad (3.2) \end{aligned}$$

Rearranging, we have our result.

The similar result holds for the point  $x^{k+\frac{1}{2}}$ :

$$F^* \geq F(x^{k+\frac{1}{2}}) - \frac{1}{2\mu_2} \mathcal{D}_2(x^{k+\frac{1}{2}}, \mu_2).$$

□

We also need the Corollary 1 from [14], which proof is almost the same as the proof of the following lemma:

**Lemma 3.2.** *For any differentiable function  $f$  and any convex function  $g$ , given  $\lambda_2 > \lambda_1 > 0$  we have*

$$\begin{aligned}\mathcal{D}_1(x^k, \lambda_2) &\geq \mathcal{D}_1(x^k, \lambda_1), \\ \mathcal{D}_2(x^{k+\frac{1}{2}}, \lambda_2) &\geq \mathcal{D}_2(x^{k+\frac{1}{2}}, \lambda_1).\end{aligned}$$

*Proof.* By convexity of  $g$  for  $0 < \alpha < 1$

$$g(\alpha z) = g(\alpha z + (1 - \alpha) \cdot 0) \leq \alpha g(z) + (1 - \alpha)g(0).$$

Then with  $z = \frac{\zeta}{\lambda_1}$  and  $\alpha = \frac{\lambda_1}{\lambda_2}$

$$\begin{aligned}g\left(\frac{\zeta}{\lambda_2}\right) - g(0) &\leq \frac{\lambda_1}{\lambda_2} \left(g\left(\frac{\zeta}{\lambda_1}\right) - g(0)\right) \\ \lambda_2 \cdot \left(g\left(\frac{\zeta}{\lambda_2}\right) - g(0)\right) &\leq \lambda_1 \cdot \left(g\left(\frac{\zeta}{\lambda_1}\right) - g(0)\right)\end{aligned}$$

Then move values of our function:

$$\lambda_2 \cdot \left(g\left(\frac{\zeta}{\lambda_2} + x_1^k\right) - g(0 + x_1^k)\right) \leq \lambda_1 \cdot \left(g\left(\frac{\zeta}{\lambda_1} + x_1^k\right) - g(0 + x_1^k)\right)$$

and add to both sides

$$h(\zeta) = \langle \nabla_1 f(x^k), \zeta \rangle + \frac{1}{2} \|\zeta\|^2$$

we have

$$\begin{aligned}\min_{\zeta \in Q} \langle \nabla_1 f(x^k), \zeta \rangle + \frac{1}{2} \|\zeta\|^2 + \lambda_2 \cdot \left(g\left(\frac{\zeta}{\lambda_2} + x_1^k\right) - g(x_1^k)\right) \\ \leq \min_{\zeta \in Q} \langle \nabla_1 f(x^k), \zeta \rangle + \frac{1}{2} \|\zeta\|^2 + \lambda_1 \cdot \left(g\left(\frac{\zeta}{\lambda_1} + x_1^k\right) - g(x_1^k)\right)\end{aligned}$$

Or with the change of variables  $\zeta = \lambda_i(u - x_1^k)$

$$\begin{aligned} \lambda_2 \min_{u \in \frac{Q}{\lambda_2} + x_1^k} \langle \nabla_1 f(x^k), u - x_1^k \rangle + \frac{\lambda_2}{2} \|u - x_1^k\|^2 + g(u) - g(x_1^k) \\ \leq \lambda_1 \min_{u \in \frac{Q}{\lambda_1} + x_1^k} \langle \nabla_1 f(x^k), u - x_1^k \rangle + \frac{\lambda_1}{2} \|u - x_1^k\|^2 + g(u) - g(x_1^k) \end{aligned}$$

which holds if

$$\frac{Q}{\lambda_2} + x_1^k \subset \frac{Q}{\lambda_1} + x_1^k. \quad (3.3)$$

For example it holds if  $Q = \mathbb{R}^n$ .

□

## 4 Convergence

In this section we prove convergence rate of Algorithm 1. If proximal PL-condition hold for  $F$ , then one can guarantee linear convergence rate, if not, the convergence is polynomial. The two following subsections contain proofs of that.

### 4.1 Linear convergence

Lipschitz continuity of the gradient of function  $f$  w.r.t. a  $\|\cdot\|$  implies

$$\begin{aligned} f(u, v) \leq f(\xi, \eta) + \langle \nabla_1 f(\xi, \eta), u - \xi \rangle + \langle \nabla_2 f(\xi, \eta), v - \eta \rangle \\ + \frac{L_1}{2} \|u - \xi\|^2 + \frac{L_2}{2} \|v - \eta\|^2, \quad (4.1) \end{aligned}$$

where again we suppose that constant  $L_1$  and  $L_2$  can be different for different blocks, and the the constant in the regular definition of Lipschitz continuity of the gradient of  $f$  is described by  $L = \max(L_1, L_2)$ .

**Theorem 4.1.** *If  $F$  from (2.1) satisfies the proximal-PL inequality (2.4). Then the algorithm 1 has a linear convergence.*

$$F(x^{k+1}) - F^* \leq \left(1 - \frac{\mu_2}{L_2}\right) \left(1 - \frac{\mu_1}{L_1}\right) [F(x^k) - F^*]. \quad (4.2)$$



*Proof.* By using Lipschitz continuity of the gradient of  $f$  we have

$$\begin{aligned}
 F(T_{L_1}(x^k)) &= F(T_{L_1}^1(x^k), x_2^k) = f(T_{L_1}^1(x^k), x_2^k) + g_1(T_{L_1}^1(x^k)) + g_2(x_2^k) \\
 &= f(T_{L_1}^1(x^k), x_2^k) + g_1(T_{L_1}^1(x^k)) + g_2(x_2^k) + g_1(x_1^k) - g_1(x_1^k) \\
 &\leq F(x^k) + \langle \nabla_1 f(x^k), T_{L_1}^1(x^k) - x_1^k \rangle + \frac{L_1}{2} \|T_{L_1}^1(x^k) - x_1^k\|^2 \\
 &\quad + g_1(T_{L_1}^1(x^k)) - g_1(x_1^k) \\
 &\leq F(x^k) - \frac{1}{2L_1} \mathcal{D}^1(x^k, L_1) \leq F(x^k) - \frac{\mu_1}{L_1} [F(x^k) - F^*],
 \end{aligned}$$

which uses the definition of  $T_M(x_k)$  and  $\mathcal{D}_1$  followed by the proximal-PL inequality (3.1). This subsequently implies that

$$F(x^{k+\frac{1}{2}}) - F^* \leq F(T_{L_1}(x^k)) - F^* \leq \left(1 - \frac{\mu_1}{L_1}\right) [F(x^k) - F^*],$$

The same derivation for the point  $x^{k+\frac{1}{2}}$  gives

$$F(x^{k+1}) - F^* \leq F(T_{L_2}(x^{k+\frac{1}{2}})) - F^* \leq \left(1 - \frac{\mu_2}{L_2}\right) [F(x^{k+\frac{1}{2}}) - F^*],$$

as well as the result of the theorem.  $\square$

Notice that above derivation does not require specification of what norm is used, so the above theorem guarantees that alternating minimization is better than the gradient methods w.r.t. any norm. In other words, alternating minimization pick up the geometric structure of the problem automatically and convergence rate of AM algorithm is not worse than the convergence rate of the gradient method in the basis with the best possible condition number.

**Example with**  $\|x\|_A = \sqrt{\langle Ax, x \rangle}$

As an example we consider here a norm endowed with a matrix. The following result can be found in the 14-th chapter of [3] or in [18]

$$\|G_{M_2}^2(x^{k+\frac{1}{2}})\|_2^2 \leq 2L_2 \left(f(x^{k+\frac{1}{2}}) - f(x^{k+1})\right) \quad (4.3)$$

$$\|G_{M_1}^1(x^k)\|_2^2 \leq 2L_1 \left(f(x^k) - f(x^{k+\frac{1}{2}})\right) \quad (4.4)$$

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

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

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

Let also consider a norm endowed with a matrix:

$$\|x\|_A^2 = \langle Ax, x \rangle.$$

We can guarantee that the following holds for any matrix  $A$

$$\begin{aligned} \|\nabla_2 f(x^{k+\frac{1}{2}})\|_{A^{-1}}^2 &\leq 2L_2^A \left( f(x^{k+\frac{1}{2}}) - f(x^{k+1}) \right) \\ \|\nabla_1 f(x^k)\|_{A^{-1}}^2 &\leq 2L_1^A \left( f(x^k) - f(x^{k+\frac{1}{2}}) \right) \end{aligned}$$

where

Let suppose that PL-conditions can be satisfied in the other basis

$$\begin{aligned} \mu_1^{B_1} \left( f(x^{k+\frac{1}{2}}) - f(x^*) \right) &\leq \|\nabla_1 f(x^k)\|_{B_1^{-1}}^2 \\ \mu_2^{B_2} \left( f(x^{k+1}) - f(x^*) \right) &\leq \|\nabla_2 f(x^{k+\frac{1}{2}})\|_{B_2^{-1}}^2, \end{aligned}$$

for all  $B_1 \in \mathbf{B}_1$  and  $B_2 \in \mathbf{B}_2$ . Then

$$f(x^{k+1}) - f(x^*) \leq \min_{B_2 \in \mathbf{B}_2} \left( 1 - \frac{\mu_2^{B_2}}{L_2^{B_2}} \right) \times \min_{B_1 \in \mathbf{B}_1} \left( 1 - \frac{\mu_1^{B_1}}{L_1^{B_1}} \right) \times \left( f(x^k) - f(x^*) \right).$$

## 4.2 Polynomial convergence

Our analysis mainly relies on the fact that alternating step is not worse than any step of any method w.r.t the only block of variables, e.g.

$$f(x_1^{k+1}, x_2^k) = \min_{z \in Q_1} f(z, x_2^k) \leq f(\text{Step}(x^k), x_2^k),$$

since  $x_1^{k+1} = \operatorname{argmin}_{z \in Q_1} f(z, x_2^k)$ .

In particular if  $\text{Step}(x)$  defined as gradient step w.r.t.  $p$  norm

$$\text{Step}^1(x) = \underset{u \in \mathbb{R}^{n_1}}{\operatorname{argmin}} f(x) + \langle \nabla_1 f(x), u - x_1 \rangle + \frac{L_p}{2} \|u - x_1\|_p^2,$$

and

$$\text{Step}^2(x) = \underset{v \in \mathbb{R}^{n_2}}{\operatorname{argmin}} f(x) + \langle \nabla_2 f(x), v - x_2 \rangle + \frac{L_p}{2} \|v - x_2\|_p^2,$$

[15] guarantee that

$$f(x^k) - f(\text{Step}(x^k)) \geq \frac{1}{2L_p} \|\nabla f(x^k)\|_{p^*}^2,$$

and

$$f(x^N) - f^* \lesssim \frac{L_p R_p^2}{N},$$

where

$$R_p^2 = \max_{x: f(x) \leq f(x_0)} \|x - x^*\|_p.$$

So for alternating minimization we can guarantee

$$f(x^N) - f^* \leq \min_{p \in [1, \infty]} \frac{2L_p R_p^2}{N}$$

## 5 Accelerated Alternating Minimization

In this section we describe accelerated method for alternating minimization, which is originates in [19]. 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 non-strongly convex or PL condition is not satisfied the algorithm 1 possesses the following convergence rate

$$f(x^N) - f_{opt} \leq \max \left\{ \frac{f(x_0) - f_{opt}}{2^{(N-1)/2}}, \frac{8 \min(L_1, L_2) R^2}{N-1} \right\}.$$

The proof can be found in [3]. The following algorithm requires the knowing of the parameter  $\mu$  of strong convexity. But it is possible to use this method with  $\mu = 0$ . In this case the algorithm turns exactly into algorithm 1 from [11]. The other interesting result that in the case method started with  $\mu = 0$  method automatically adapts to strong convexity of the problem and poses at least the same linear convergence rate as a gradient descent (see Lemma 5.3).

The set  $\{1, \dots, m\}$  of indices of the orthonormal basis vectors  $\{e_i\}_{i=1}^m$  is divided into  $n$  disjoint subsets (blocks)  $I_k$ ,  $k \in \{1, \dots, n\}$ . Let  $S_k(x) = x + \text{span}\{e_i : i \in I_k\}$ , i.e. the affine subspace containing  $x$  and all the points differing from  $x$  only over the block  $k$ . We use  $x_i$  to denote the components of  $x$  corresponding to the block  $i$  and  $\nabla_i f(x)$  to denote the gradient corresponding to the block  $i$ . We will further require that for any  $k \in \{1, \dots, n\}$  and any  $z \in \mathbb{R}^m$  the problem  $f(x) \rightarrow \min_{x \in S_i(z)}$  has a solution, and this solution is easily computable.

---

**Algorithm 2** Accelerated Alternating Minimization (AAM)
 

---

**Input:** Starting point  $x_0$

**Output:**  $x^k$

1: Set  $A_0 = 0$ ,  $x^0 = v^0$ ,  $\tau_0 = 1$

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

3: Set

$$\beta_k = \operatorname{argmin}_{\beta \in [0,1]} f\left(x^k + \beta(v^k - x^k)\right) \quad (5.1)$$

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

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

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

$$f(y^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})(\tau_k + \mu a_{k+1})} \|\nabla f(y^k)\|_2^2 + \frac{\mu \tau_k a_{k+1}}{2(A_k + a_{k+1})(\tau_k + \mu a_{k+1})} \|v^k - y^k\|_2^2 = f(x^{k+1}) \quad (5.2)$$

8: Set  $A_{k+1} = A_k + a_{k+1}$ ,  $\tau_{k+1} = \tau_k + \mu a_{k+1}$

9: Set  $v^{k+1} = v^k - a_{k+1} \nabla f(y^k)$ . {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  $\tau_k$  strongly convex function with

$$\tau_k = 1 + \mu \sum_{i=0}^k a_i = 1 + \mu A_k.$$

**Lemma 5.1.** *After  $k$  steps of Algorithm 2 it holds that*

$$A_k f(x^k) \leq \min_{x \in \mathbb{R}^m} \psi_k(x) = \psi_k(v^k). \quad (5.3)$$

Moreover, if the objective is  $L$ -smooth and  $\mu$ -strongly convex

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

where  $n$  is the number of blocks.

*Proof.* First, we prove inequality (5.3) by induction over  $k$ . For  $k = 0$ , the inequality holds. Assume that

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

Then

$$\begin{aligned} \psi_{k+1}(v^{k+1}) &= \min_{x \in \mathbb{R}^m} \left\{ \psi_k(x) + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|_2^2\} \right\} \\ &\geq \min_{x \in \mathbb{R}^m} \left\{ \psi_k(v^k) + \frac{\tau_k}{2} \|x - v^k\|_2^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle \right. \\ &\quad \left. + \frac{\mu}{2} \|x - y^k\|_2^2 \right\} \\ &\geq \min_{x \in \mathbb{R}^m} \left\{ A_k f(x^k) + \frac{\tau_k}{2} \|x - v^k\|_2^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle \right. \\ &\quad \left. + \frac{\mu}{2} \|x - y^k\|_2^2 \right\} \end{aligned}$$

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

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

- (i)  $\beta_k = 1$ ,  $\langle \nabla f(y^k), x^k - v^k \rangle \geq 0$ ,  $y^k = v^k$ ;
- (ii)  $\beta_k \in (0, 1)$  and  $\langle \nabla f(y^k), x^k - v^k \rangle = 0$ ,  $y^k = v^k + \beta_k(x^k - v^k)$ ;
- (iii)  $\beta_k = 0$  and  $\langle \nabla f(y^k), x^k - v^k \rangle \leq 0$ ,  $y^k = x^k$ .

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

Thus

$$\psi_{k+1}(v^{k+1}) \geq \min_{x \in \mathbb{R}^m} \left\{ A_k f(y^k) + \frac{\tau_k}{2} \|x - v^k\|_2^2 + a_{k+1} \{ f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|_2^2 \} \right\}$$

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  $\langle \nabla f(y^k), v^k - y^k \rangle \geq 0$ , we obtain

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

Our next goal is to show that

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

which proves the induction step.

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

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

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

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

and

$$f(y^k) - \frac{1}{2Ln} \|\nabla f(y^k)\|_2^2 \geq f(y^k) - \frac{1}{2L} \|\nabla_{i_k} f(y^k)\|_2^2 \geq f(x^{k+1}),$$

Choosing  $a_{k+1}$  such that  $\frac{a_{k+1}^2}{2A_{k+1}\tau_{k+1}} \geq \frac{1}{2Ln}$  implies

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

which proves the induction step.

Rewriting the rule for choosing  $a_{k+1}$  gives  $\frac{a_{k+1}^2}{(A_k + a_{k+1})(\tau_k + \mu a_{k+1})} \geq \frac{1}{Ln}$ .

Let us estimate the rate of the growth for  $A_k$ .  $\tau_k = 1 + \mu \sum_{i=0}^k a_i = 1 + \mu A_k$ .

$$\frac{a_{k+1}^2}{2A_{k+1}\tau_{k+1}} \geq \frac{1}{2Ln}$$

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

$$a_k \geq \frac{1}{\sqrt{nL}} \sqrt{A_k + \mu A_k^2} \geq \sqrt{\frac{\mu}{2Ln}} A_k \quad (5.4)$$

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

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

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

We also have

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

which leads to

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

To use this bound we only need to estimate  $A_1$ , which we can do as follows:

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

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

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

□

**Theorem 5.2.** *After  $k$  steps of Algorithm 2 it holds that*

$$f(x^k) - f(x_*) \leq nLR^2 \min \left\{ \frac{4}{k^2}, \left( 1 - \sqrt{\frac{\mu}{nL}} \right)^{k-1} \right\} \quad (5.5)$$

*Proof.* From the convexity of  $f(x)$  we have

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

From Lemma (5.1) we have

$$\begin{aligned} A_k f(x^k) &\leq \psi_k(v^k) \leq \psi_k(x_*) = \frac{1}{2} \|x_* - x^0\|_2^2 \\ + \sum_{i=0}^{k-1} a_{i+1} (f(y^i) + \langle \nabla f(y^i), x_* - y^i \rangle) + \frac{\mu}{2} \|x_* - y^i\|_2^2 &\leq A_k f(x_*) + \frac{1}{2} \|x_* - x^0\|_2^2 \\ f(x^k) - f(x_*) &\leq \frac{R^2}{2A_k} \leq nLR^2 \min \left\{ \frac{4}{k^2}, \left( 1 - \sqrt{\frac{\mu}{nL}} \right)^{k-1} \right\}. \end{aligned}$$

□

The other observation explains behaviour of this method when  $\mu$  is unknown.

**Lemma 5.3.** *Algorithm 2 started with  $\mu = 0$  automatically adapts to strong convexity of the problem and has linear convergence:*

$$f(x^{k+1}) - f(x^*) \leq \prod_{i=0}^{k-1} \left( 1 - \frac{\mu}{\hat{L}_i} \right) \cdot (f(x^0) - f(x^*)),$$

where  $\hat{L}_i = \frac{A_i + a_{i+1}}{a_{i+1}^2}$  is the upper bound on  $L$  at the  $i$ -th iteration.



*Proof.* (5.2) with  $\mu = 0$  implies sufficient decrease result:

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

since (5.1) implies that  $f(y^k) \leq f(x^k)$ . By combining this result with PL-condition

$$\|\nabla f(y^k)\|_2^2 \geq 2\mu (F(y^k) - F(x^*))$$

we have linear convergence

$$\begin{aligned} (f(y^{k+1}) - f(x^*)) &\leq \left(1 - \frac{\mu a_{k+1}^2}{A_k + a_{k+1}}\right) (f(y^k) - f(x^*)) \\ &\leq \prod_{i=0}^k \left(1 - \frac{\mu a_{i+1}^2}{A_i + a_{i+1}}\right) (f(x^0) - f(x^*)) \end{aligned}$$

And finally block minimization step guarantees that  $f(x^{k+1}) \leq f(y^k)$ , so we have

$$f(x^{k+1}) - f(x^*) \leq \prod_{i=0}^{k-1} \left(1 - \frac{\mu a_{i+1}^2}{A_i + a_{i+1}}\right) (f(x^0) - f(x^*))$$

□

## 6 Application

Consider the following problem of minimizing a quadratic function

$$f(z) = \|Wz - b\|_2^2 \rightarrow \min_z \tag{6.1}$$

this is a strongly convex function with  $\mu = \lambda_{\min}(W^T W)$ .

This problem can be solved with algorithm 1 by splitting the vector variable  $z$  into two vector variables with the dimension:

$$z = \begin{pmatrix} x \\ y \end{pmatrix}.$$

Then split matrix  $W$  into four blocks with the same size

$$W = \begin{pmatrix} AB \\ CD \end{pmatrix}.$$

and vector  $b$

$$b = \begin{pmatrix} d \\ c \end{pmatrix}.$$

The equivalent problem to (6.1) is

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

and the iterations of the algorithm 1 can be written explicitly

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

Next we provide comparison between Algorithm 1, Algorithm 2 started with  $\mu = 0$  and  $\mu = \mu^*$ , and the following accelerated algorithm

---

**Algorithm 3** Fast Gradient Method (FGM)

---

**Input:** Starting point  $z_0$ .

**Output:**  $z^k$

- 1: Set  $v^0 = z^0$ .
  - 2: **for**  $k \geq 0$  **do**
  - 3:    $z^{k+1} = v^k - \frac{1}{L} \nabla f(v^k)$
  - 4:    $v^{k+1} = z^k + \frac{k}{k+3} (z^{k+1} - z^k)$
  - 5: **end for**
- 

## 7 Linear convergence under general convex constraint sets

The proof of linear convergence relies on Lemma 3.2, which requires special structure of constraint sets (3.3). For general convex constraints, we were able to prove only a weaker result, which is presented in this section.

The following lemma is used instead of Theorem 4.1 and Lemma 3.2.

**Lemma 7.1.** *Strong convexity of  $f$  implies "nearly" PL-condition:*

$$\mu_1 \left( F(x^{k+\frac{1}{2}}) - F(x^*) \right) \leq \frac{1}{2} \|G_{M_1}^1(x^k)\|_2^2 \quad (7.1)$$

$$\mu_2 \left( F(x^{k+1}) - F(x^*) \right) \leq \frac{1}{2} \|G_{M_2}^2(x^{k+\frac{1}{2}})\|_2^2 \quad (7.2)$$

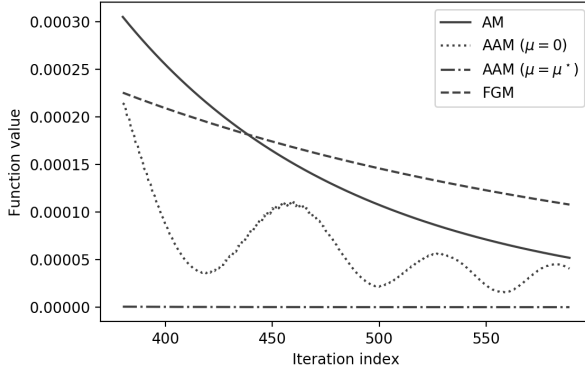


Figure 1. Comparison for quadratic function

*Proof.* Consider

$$T_M^2(x^{k+\frac{1}{2}}) = \operatorname{argmin}_{u \in Q} \left( g_2(u) + \frac{M_2}{2} \|u - (x_2^k - \frac{1}{M_2} \nabla f(x^{k+\frac{1}{2}}))\|_2^2 \right)$$

Since  $T_{M_2}^2(x^{k+\frac{1}{2}})$  is a minimizer, optimality condition gives for all  $v \in Q$

$$\langle \partial g_2(T_{M_2}^2(x^{k+\frac{1}{2}})) + \nabla_2 f(x^{k+\frac{1}{2}}) + M(T_{M_2}^2(x^{k+\frac{1}{2}}) - x_2^k), v - T_{M_2}^2(x^{k+\frac{1}{2}}) \rangle \geq 0$$

or

$$\begin{aligned} \langle \nabla_2 f(x^{k+\frac{1}{2}}), v - T_{M_2}^2(x^{k+\frac{1}{2}}) \rangle &\geq \langle -\partial g_2(T_{M_2}^2(x^{k+\frac{1}{2}})), v - T_{M_2}^2(x^{k+\frac{1}{2}}) \rangle \\ &\quad + \langle G_{M_2}^2(x^{k+\frac{1}{2}}), v - T_{M_2}^2(x^{k+\frac{1}{2}}) \rangle \end{aligned} \quad (7.3)$$

Since  $f$  is strongly convex

$$\begin{aligned}
f(u, v) &\geq f(x_1^{k+1}, x_2^k) + \langle \nabla_1 f(x_1^{k+1}, x_2^k), u - x_1^{k+1} \rangle + \langle \nabla_2 f(x_1^{k+1}, x_2^k), v - x_2^k \rangle \\
&\quad + \frac{\mu_1}{2} \|u - x_1^{k+1}\|_2^2 + \frac{\mu_2}{2} \|v - x_2^k\|_2^2 \geq \\
&\stackrel{\textcircled{1}}{\geq} f(x_1^{k+1}, x_2^k) - \langle \partial g_1(x_1^{k+1}), u - x_1^{k+1} \rangle + \langle \nabla_2 f(x_1^{k+1}, x_2^k), v - x_2^k \rangle + \frac{\mu_2}{2} \|v - x_2^k\|_2^2 \\
&\stackrel{\textcircled{2}}{\geq} f(x_1^{k+1}, x_2^k) + g_1(x_1^{k+1}) - g_1(u) + \langle \nabla_2 f(x_1^{k+1}, x_2^k), v - x_2^k \rangle + \frac{\mu_2}{2} \|v - x_2^k\|_2^2 \\
&\stackrel{\textcircled{3}}{\geq} f(x_1^{k+1}, x_2^k) + g_1(x_1^{k+1}) + \langle \nabla_2 f(x_1^{k+1}, x_2^k), T_{M_2}^2(x^{k+\frac{1}{2}}) - x_2^k \rangle + \frac{\mu_2}{2} \|v - x_2^k\|_2^2 \\
&\quad - g_1(u) - \langle \partial g_2(T_{M_2}^2(x^{k+\frac{1}{2}})), v - T_{M_2}^2(x^{k+\frac{1}{2}}) \rangle + \langle G_{M_2}^2(x^{k+\frac{1}{2}}), v - T_{M_2}^2(x^{k+\frac{1}{2}}) \rangle \\
&\stackrel{\textcircled{4}}{\geq} f(x_1^{k+1}, x_2^k) + g_1(x_1^{k+1}) - g_1(u) + \langle \nabla_2 f(x_1^{k+1}, x_2^k), T_{M_2}^2(x^{k+\frac{1}{2}}) - x_2^k \rangle - g_2(v) \\
&\quad + g_2(T_{M_2}^2(x^{k+\frac{1}{2}})) + \frac{\mu_2}{2} \|v - x_2^k\|_2^2 + \langle G_{M_2}^2(x^{k+\frac{1}{2}}), v - T_{M_2}^2(x^{k+\frac{1}{2}}) \rangle \\
&\stackrel{\textcircled{5}}{\geq} f(x_1^{k+1}, T_{M_2}^2(x^{k+\frac{1}{2}})) - \frac{M_2}{2} \left\| \frac{-1}{M_2} G_{M_2}^2(x^{k+\frac{1}{2}}) \right\|_2^2 + g_1(x_1^{k+1}) - g_1(u) - g_2(v) \\
&\quad + g_2(T_{M_2}^2(x^{k+\frac{1}{2}})) + \frac{\mu_2}{2} \|v - x_2^k\|_2^2 + \langle G_{M_2}^2(x^{k+\frac{1}{2}}), v - T_{M_2}^2(x^{k+\frac{1}{2}}) \rangle \quad (7.4)
\end{aligned}$$

where we used:

- $\textcircled{1}$  by (2.3)
- $\textcircled{2}$  by convexity  $-\langle \partial g_1(x_1^{k+1}), u - x_1^{k+1} \rangle \geq g_1(x_1^{k+1}) - g_1(u)$
- $\textcircled{3}$  by (7.3)
- $\textcircled{4}$  by convexity  $-\langle \partial g_2(T_{M_2}^2(x^{k+\frac{1}{2}})), v - T_{M_2}^2(x^{k+\frac{1}{2}}) \rangle \geq g_2(T_{M_2}^2(x^{k+\frac{1}{2}})) - g_2(v)$
- $\textcircled{5}$  since  $\nabla_2 f$  is  $L_2$ -Lipschitz continuous, for  $M_2 \geq L_2$  the following holds

$$\begin{aligned}
f(x^{k+\frac{1}{2}}) + \langle \nabla_2 f(x^{k+\frac{1}{2}}), T_{M_2}^2(x^{k+\frac{1}{2}}) - x_2^k \rangle &\geq \\
&\geq f(x_1^{k+1}, T_{M_2}^2(x^{k+\frac{1}{2}})) - \frac{M_2}{2} \left\| \frac{-1}{M_2} G_{M_2}^2(x^{k+\frac{1}{2}}) \right\|_2^2
\end{aligned}$$

Above inequality gives

$$\begin{aligned}
 F(u, v) &\geq F(x_1^{k+1}, T_{M_2}^2(x^{k+\frac{1}{2}})) \\
 &\quad + \frac{M_2}{2} \left\| \frac{-1}{M_2} G_{M_2}^2(x^{k+\frac{1}{2}}) \right\|_2^2 + \frac{\mu_2}{2} \|v - x_2^k\|_2^2 + \langle G_{M_2}^2(x^{k+\frac{1}{2}}), v - x_2^k \rangle \\
 &\geq F(x_1^{k+1}, T_{M_2}^2(x^{k+\frac{1}{2}})) + \frac{\mu_2}{2} \|v - x_2^k\|_2^2 + \langle G_{M_2}^2(x^{k+\frac{1}{2}}), v - x_2^k \rangle \longrightarrow \min_{v \in \mathbb{R}^n}
 \end{aligned} \tag{7.5}$$

Plugging in  $(u, v) = x^*$  we get one of the desired inequalities:

$$\begin{aligned}
 F(x^*) &\geq F(x_1^{k+1}, T_{M_2}^2(x^{k+\frac{1}{2}})) - \frac{1}{2\mu_2} \|G_{M_2}^2(x^{k+\frac{1}{2}})\|_2^2 \\
 &\geq F(x^{k+1}) - \frac{1}{2\mu_2} \|G_{M_2}^2(x^{k+\frac{1}{2}})\|_2^2
 \end{aligned} \tag{7.6}$$

$$\|G_{M_1}^1(x^k)\|_2^2 \geq 2\mu_1 \left( F(x^{k+\frac{1}{2}}) - F(x^*) \right) \tag{7.7}$$

The other inequality can be obtained the same way for the point  $x^k$ :

$$\|G_{M_2}^2(x^{k+\frac{1}{2}})\|_2^2 \geq 2\mu_2 \left( F(x^{k+1}) - F(x^*) \right) \tag{7.8}$$

□

Combining the result of the above lemma with (4.3), (4.4), we obtain convergence rate:

$$\mu_1 \left( F(x^{k+\frac{1}{2}}) - F(x^*) \right) \leq L_1 \left( F(x^k) - F(x^{k+\frac{1}{2}}) \right) \tag{7.9}$$

$$\mu_2 \left( F(x^{k+1}) - F(x^*) \right) \leq L_2 \left( F(x^{k+\frac{1}{2}}) - F(x^{k+1}) \right) \tag{7.10}$$

$$\begin{aligned}
 \left( F(x^{k+1}) - F(x^*) \right) &\leq \left( 1 - \frac{\mu_2}{L_2 + \mu_2} \right) \left( F(x^{k+\frac{1}{2}}) - F(x^*) \right) \\
 \left( F(x^{k+\frac{1}{2}}) - F(x^*) \right) &\leq \left( 1 - \frac{\mu_1}{L_1 + \mu_1} \right) \left( F(x^k) - F(x^*) \right)
 \end{aligned}$$

$$\begin{aligned}
 \left( F(x^{k+1}) - F(x^*) \right) &\leq \\
 &\leq \left( 1 - \frac{\mu_2}{L_2 + \mu_2} \right) \left( 1 - \frac{\mu_1}{L_1 + \mu_1} \right) \left( F(x^k) - F(x^*) \right)
 \end{aligned} \tag{7.11}$$

## Bibliography

- [1] Andreas Andresen and Vladimir Spokoiny, Convergence of an Alternating Maximization Procedure, *Journal of Machine Learning Research* **17** (2016), 1–53.
- [2] 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** (2015), 185–209.
- [3] Amir. Beck, *First-Order Methods in Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.
- [4] Dimitri P. Bertsekas and John N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.
- [5] Charles Byrne, Iterative Reconstruction Algorithms Based on Cross-Entropy Minimization, in: *Image Models (and their Speech Model Cousins)* (Stephen E. Levinson and Larry Shepp, eds.), pp. 1–11, Springer New York, New York, NY, 1996.
- [6] Charles L Byrne, *Iterative optimization in inverse problems*, CRC Press, 2014.
- [7] Antonin Chambolle, Pauline Tan and Samuel Vaiter, Accelerated Alternating Descent Methods for Dykstra-Like Problems, *Journal of Mathematical Imaging and Vision* **59** (2017), 481–497.
- [8] Imre Csiszár and Gábor E. Tusnády, Information geometry and alternating minimization procedures, 1984.
- [9] Jelena Diakonikolas and Lorenzo Orecchia, Alternating Randomized Block Coordinate Descent, in: *Proceedings of the 35th International Conference on Machine Learning* (Jennifer Dy and Andreas Krause, eds.), Proceedings of Machine Learning Research 80, pp. 1224–1232, PMLR, Stockholm, Sweden, 10–15 Jul 2018.
- [10] Richard Gordon, Robert Bender and Gabor T. Herman, Algebraic Reconstruction Techniques (ART) for three-dimensional electron microscopy and X-ray photography, *Journal of Theoretical Biology* **29** (1970), 471 – 481.
- [11] Sergey Guminov, Pavel Dvurechensky, Nazarii Tupitsa and Alexander Gasnikov, Accelerated Alternating Minimization, Accelerated Sinkhorn’s Algorithm and Accelerated Iterative Bregman Projections, *arXiv e-prints* (2019), arXiv:1906.03622.
- [12] 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** (2016), 57–77.
- [13] Mingyi Hong, Xiangfeng Wang, Meisam Razaviyayn and Zhi-Quan Luo, Iteration complexity analysis of block coordinate descent methods, *Mathematical Programming* **163** (2017), 85–114.
- [14] Hamed Karimi, Julie Nutini and Mark Schmidt, *Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition*, 2016.

- 
- [15] Jonathan A. Kelner, Yin Tat Lee, Lorenzo Orecchia and Aaron Sidford, An Almost-Linear-Time Algorithm for Approximate Max Flow in Undirected Graphs, and its Multicommodity Generalizations, *arXiv e-prints* (2013), arXiv:1304.2338.
- [16] K. Lange, M. Bahn and R. Little, A Theoretical Study of Some Maximum Likelihood Algorithms for Emission and Transmission Tomography, *IEEE Transactions on Medical Imaging* **6** (1987), 106–114.
- [17] Zhi-Quan Luo and Paul Tseng, Error bounds and convergence analysis of feasible descent methods: a general approach, *Annals of Operations Research* **46** (1993), 157–178.
- [18] Yurii Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1 ed, Springer Publishing Company, Incorporated, 2014.
- [19] Yurii Nesterov, Alexander Gasnikov, Sergey Guminov and Pavel Dvurechensky, *Primal-dual accelerated gradient methods with small-dimensional relaxation oracle*, 2018.
- [20] Yurii Nesterov and Boris Polyak, Cubic regularization of Newton method and its global performance, *Math. Program.* **108** (2006), 177–205.
- [21] Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander and Hoyt Koepke, Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection, in: *Proceedings of the 32nd International Conference on Machine Learning* (Francis Bach and David Blei, eds.), Proceedings of Machine Learning Research 37, pp. 1632–1641, PMLR, Lille, France, 07–09 Jul 2015.
- [22] 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.
- [23] Boris Polyak, *Introduction to Optimization*, New York, Optimization Software, 1987.
- [24] 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*, pp. 1306–1314, MIT Press, Cambridge, MA, USA, 2015.
- [25] Y. Vardi, L. A. Shepp and L. Kaufman, A Statistical Model for Positron Emission Tomography, *Journal of the American Statistical Association* **80** (1985), 8–20.
- [26] Curtis R Vogel, *Computational methods for inverse problems*, 23, Siam, 2002.
- [27] Nan Ye, Farbod Roosta-Khorasani and Tiangang Cui, *Optimization Methods for Inverse Problems*, pp. 121–140, 01 2019.

Received ???.

**Author information**

Nazarii Tupitsa, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow, Russia;

Institute for Information Transmission Problems RAS, Moscow, Russia;

National Research University Higher School of Economics, Moscow, Russia.

E-mail: [tupitsa@phystech.edu](mailto:tupitsa@phystech.edu)

Pavel Dvurechensky, Weierstrass Institute for Applied Analysis and Stochastics, Berlin;

Institute for Information Transmission Problems RAS, Moscow, Russia.

E-mail: [pavel.dvurechensky@wias-berlin.de](mailto:pavel.dvurechensky@wias-berlin.de)

Alexander Gasnikov, National Research University Higher School of Economics, Russia;

Moscow Institute of Physics and Technology, Dolgoprudny, Russia;

Institute for Information Transmission Problems RAS, Moscow, Russia.

E-mail: [gasnikov@yandex.ru](mailto:gasnikov@yandex.ru)

Sergey Guminov, Moscow Institute of Physics and Technology, Dolgoprudny, Russia;

Institute for Information Transmission Problems RAS, Moscow,.

E-mail: [sergey.guminov@phystech.edu](mailto:sergey.guminov@phystech.edu)