Alternating Minimization Methods for Strongly Convex Optimization

Nazarii Tupitsa∗ Pavel Dvurechensky∗∗
Alexander Gasnikov∗∗∗ Sergey Guminov∗∗∗∗

∗ Moscow Institute of Physics and Technology, Dolgoprudny, Russia,
Institute for Information Transmission Problems RAS, Moscow
(e-mail: tupitsa@phystech.edu)
∗∗ Weierstrass Institute for Applied Analysis and Stochastics, Berlin,
Institute for Information Transmission Problems RAS, Moscow
(e-mail: pavel.dvurechensky@wias-berlin.de)
∗∗∗ Higher school of economics,
Moscow Institute of Physics and Technology, Dolgoprudny, Russia,
Institute for Information Transmission Problems RAS, Moscow
(e-mail: gasnikov@yandex.ru)
∗∗∗∗ Moscow Institute of Physics and Technology, Dolgoprudny, Russia,
Institute for Information Transmission Problems RAS, Moscow
(e-mail: sergey.guminov@phystech.edu)

Abstract: We consider alternating minimization procedures for convex optimization problems with variable divided in many blocks, each block being amenable for minimization with freezed other variables blocks. In the case of two blocks, we prove a linear convergence rate for alternating minimization procedure under Polyak-ojasiewicz 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.

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

1. INTRODUCTION

In this paper we consider unconstrained minimization problem
\[ \min_{x \in \mathbb{R}^m} f(x), \]
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 \( L_i \subseteq \mathbb{R}^m \), s.t. \( \bigcup L_i = \mathbb{R}^m \) 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
\[ x^*_i = \text{argmin}_{z \in \mathbb{R}^{n_i}} f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n), \]
where \( n_i \) is the size of the \( i \)-th block.

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. 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 Luo and Tseng (1993). Beck (2015) 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 is sublinear, namely \( O(1/k) \), where \( k \) is the iteration counter. Similar result, but for many-block setting was obtained in Hong et al. (2017); Sun and Hong (2015). In the fully smooth setting under strong convexity assumption Nutini et al. (2015) obtain linear rate of convergence also for the many-block setting. This linear rate is proportional to \( \kappa \) – efficient condition number of the problem. Chambolle et al. (2017) provide an accelerated alternating minimization method for a very 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 convex case. Diakonikolas and Orecchia (2018) 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. We also mention the review Hong et al. (2016).

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-ojasiewicz condition Polyak (1987). 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

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

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 my proofs 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 $k \geq 0$ do
3: if $k \mod 2 = 0$ then
4: \[x_{k+1} = \arg\min_{x \in \mathbb{R}^n} f(z, x_k^2)\]
5: else
6: \[x_{k+1} = \arg\min_{x \in \mathbb{R}^n} f(x_{k+1}, z)\]
7: end if
8: end for

Optimality conditions for algorithm’s minimization problems read as follows:

\[
\begin{align*}
\nabla_1 f(x_{k+1}, x_{k}^2) &= 0 \quad (2) \\
\nabla_2 f(x_{k+1}, x_{k}^2) &= 0 \quad (3)
\end{align*}
\]

Introduce the following notation:

\[T_M(x) = (T_M^1(x), T_M^2(x))\]

\[G_M(x) = (G_M^1(x), G_M^2(x))\]

\[T_{M_i}^i(x) = \arg\min_{x \in \mathbb{R}^n} \frac{1}{M_i} \nabla_i f(x) = \frac{1}{M_i} \nabla_i f(x)\]

For the case $i = 1$

\[
T_{M_1}^1(x_k) = \arg\min_{x \in \mathbb{R}^n} \left(\frac{M_1}{2} \|u - x_1 - \frac{1}{M_1} \nabla_1 f(x)\|^2 + \langle \nabla_1 f(x), u - x_1 \rangle\right) = \arg\min_{x \in \mathbb{R}^n} \left(\frac{M_1}{2} \|u - x_1\|^2 + \langle \nabla_1 f(x), u - x_1 \rangle\right).
\]

Lemma 1.

\[G_M^1(x_{k+\frac{1}{2}}) = 0, \quad G_M^2(x_k) = 0\]

for all $k$.

Proof:

\[T_M^1(x_k) = \arg\min_{u \in \mathbb{R}^n} \left(\frac{M_1}{2} \|v - x_1\|^2 + \langle \nabla_2 f(x_k), v - x_1 \rangle\right) = x_2, \quad \text{since } \nabla_2 f(x_1^2, x_2^2) = 0 \text{ by (3)}.
\]

\[T_M(x_k) = (T_M^1(x_k), T_M^2(x_k)) = \left(\frac{T_M^1(x_k), x_2^2 - \frac{1}{M} G_M^2(x_k)\right),
\]

where the last equality follows from the definition of $G_M^2(x_k)$.

3. SUFFICIENT DECREASE-TYPE RESULT

The following result can be found in the 14-th chapter of Beck (2017) or in Nesterov (2014)

\[\|G_M(x_{k+\frac{1}{2}})\|^2 \leq 2L_2 \left(f(x_{k+\frac{1}{2}}) - f(x_{k+1})\right) \quad (5)\]

\[\|G_M(x_k)\|^2 \leq 2L_1 \left(f(x_k) - f(x_{k+\frac{1}{2}})\right) \quad (6)\]

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

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

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-OJASIEWICZ CONDITION

Our myproof of the convergence rate demands Polyak-ojasiewicz (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:

\[f(x^*) \geq f(x_k) - \frac{1}{2\mu_1} \|G_M^1(x_k)\|^2 \quad (7)\]

\[f(x^*) \geq f(x_{k+\frac{1}{2}}) - \frac{1}{2\mu_2} \|G_M^2(x_{k+\frac{1}{2}})\|^2 \quad (8)\]

Proof:

Since $T_{M_1}^1(x)$ is a minimizer of $\frac{M_1}{2} \|u - x_1\|^2 + \langle \nabla_1 f(x_k), u - x_1 \rangle$, w.r.t. $u \in \mathbb{R}^n$

\[\nabla_1 f(x_k) + M_1 (T_{M_1}^1(x_k) - x_1) = 0\]

or equivalently

\[\nabla_1 f(x_k) = G_{M_1}^1(x_k) \quad (9)\]

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

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

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:
\[ f(u, v) - f(x^k) - f(x^*) \leq f \left( x^{k+\frac{1}{2}} \right) - f(x^*) \leq \left( 1 - \frac{\mu_1}{L_1} \right) f(x^k) - f(x^*) \]

By combining these inequalities we get

\[ f(x^{k+1}) - f(x^*) \leq (1 - \frac{\mu_2}{L_2}) f(x^k) - f(x^*) \]

or for regular definition of PL condition

\[ f(x^{k+1}) - f(x^*) \leq (1 - \frac{\mu}{L}) f(x^k) - f(x^*) \]

\[ \text{notice that } (1 - \frac{\mu}{L_{\text{max}}}) \leq 1 \]

\[ \leq (1 - \frac{\mu}{L_{\text{min}}}) f(x^k) - f(x^*) \]

6. ACCELERATED ALTERNATING MINIMIZATION

In this section we describe accelerated method for alternating minimization, which is originates in Nesterov et al. (2018). 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 posses the following convergence rate \( f(x^k) - f(x^*) \)
Then

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

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

$$\min_{\beta \in [0, 1]} f \left( x^k + \beta(v^k - x^k) \right),$$

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

(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 $\langle \nabla f(y^k), x^k - v^k \rangle \leq 0$, $y^k = x^k$.

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

Thus

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

The explicit solution to the above quadratic optimization problem is

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

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

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

Our next goal is to show that

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

which proves the induction step.

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

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

where $x^{k+1} = \arg\max_{x \in S_k} f(x)$. Since $i_k = \arg\max_{i \in \mathbb{R}^N} \|\nabla_i f(y^k)\|^2$,

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

and $f(y^k) - \frac{n-1}{2L} \|\nabla f(y^k)\|^2 \geq f(y^k) - \frac{n-1}{2} \|\nabla_i f(y^k)\|^2 \geq f(x^{k+1})$. Choosing $a_{k+1}$ such that $\frac{\mu^2}{2 \tau_{k+1}} \|y^k - v^k\|^2 \geq \frac{1}{2\tau_{k+1}}$ implies

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

which proves the induction step.

Rewriting the rule for choosing $a_{k+1}$ gives

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

Let us estimate the rate of the growth for $A_k$. $\tau_{k} = 1 + \mu \sum_{i=0}^{k} a_i = 1 + \mu A_k$, $\frac{a_{k+1}^2}{A_{k+1} \tau_{k+1} + \mu a_{k+1}} \geq \frac{1}{2 \tau_{k+1}}$.

$$a_k \geq \frac{\sqrt{A_k + \frac{\mu^2}{2L} A_k^2}}{\sqrt{2 \Lambda L}}$$

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

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

We also have

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

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:

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

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

$$f(x^k) - f(x^*) \leq nL \frac{R^2}{k^2}, \left( 1 - \sqrt{\frac{\mu}{nL}} \right)^{k-1}$$

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

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

From Lemma (3) we have
The solution to this problem is \( x^* \) for \( i,j \)

First of all, we notice that for all \( X \) such that the primal function is strongly convex on it. Here minimizing the Lagrangian over a closed convex set dualize the linear constraints where \( \langle A, B \rangle = \sum_{i,j=1}^n A_{ij} B_{ij} \).

To ensure the smoothness of the dual problem, we must dualize the linear constraints \( \mathbf{X} = \mathbf{r}, \mathbf{X}^T \mathbf{1} = \mathbf{c} \) 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:

\[
\min_{X \in \mathbb{R}^{N \times N}_+} \langle C, X \rangle + \gamma \langle X, \ln X \rangle
\]

\[
= \min_{X \in Q} \max_{y,v \in \mathbb{R}^N} \langle C, X \rangle + \gamma \langle X, \ln X \rangle + \langle y, X \mathbf{1} - \mathbf{r} \rangle + \langle z, \mathbf{X}^T \mathbf{1} - \mathbf{c} \rangle
\]

\[
= \max_{y,v \in \mathbb{R}^N} \min_{z \in \mathbb{R}^N} \langle y, \mathbf{r} \rangle - \langle z, \mathbf{c} \rangle + \sum_{i,j=1}^n X^i_j \left( C^i_j + \gamma \ln X^i_j + y^i_j + z^j_i \right)
\]

First of all, we notice that for all \( i,j \) and some small \( \varepsilon \)

for \( X^i_j \in (0, \varepsilon) \) and approaches 0 as \( X^i_j \) approaches 0.

Hence, \( X^i_j > 0 \) without loss of generality. Using Lagrange multipliers for the constraint \( \mathbf{1}^T X \mathbf{1} = 1 \), we obtain the problem

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

The solution to this problem is

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

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

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

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

\[
u^{k+1} = \arg\min_{\nu \in \mathbb{R}^N} \varphi(u^n, v^n), \quad u^{k+1} = \arg\min_{u \in \mathbb{R}^N} \varphi(u^n, v^n),
\]

may be written explicitly as

\[
u^{k+1} = u^k + \ln r - \ln (B(u, v) \mathbf{1}), \quad u^{k+1} = v^k + \ln c - \ln (B(u, v)^T \mathbf{1}).
\]

**Proof:**

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

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

\[
r - \frac{1}{1^T B(u, v^n) \mathbf{1}} B(u, v^n) \mathbf{1} = 0.
\]

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

\[
B(u^{k+1}, v^n) \mathbf{1} = \text{diag}(e^{(u^{k+1} - u^n)}) B(u^n, v^n) \mathbf{1} = r,
\]

and the conclusion then follows from the fact that \( 1^T B(v^{k+1}, v^n) \mathbf{1} = 1 \).

The optimality of \( u^{k+1} \) can be proved in the exact same way.

The AM algorithm for this problem with \( t = 0 \) is the well-known Sinkhorn’s algorithm (Cuturi (2013)).

**Algorithm 3 Sinkhorn’s Algorithm**

**Ensure:** \( x^k \)

**for** \( k \geq 1 \) **do**

**if** \( k \mod 2 = 0 \) **then**

\[
u^{k+1} = u^k + \ln r - \ln (B(u^k, v^k) \mathbf{1})
\]

\[
u^{k+1} = v^k
\]

**else**

\[
u^{k+1} = u^k
\]

\[
u^{k+1} = v^k + \ln c - \ln (B(u^k, v^k)^T \mathbf{1})
\]

**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), but convergence in iteration worse then 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 \( \mu \) to be
initialized in order to pose linear convergence. So we experimentally found parameter $\mu$ 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.

REFERENCES

Beck, A. (2015). 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.

Beck, A. (2017). *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA. doi:10.1137/1.9781611974997.

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

Chambolle, A., Tan, P., and Vaiter, S. (2017). Accelerated alternating descent methods for dykstra-like problems. *Journal of Mathematical Imaging and Vision*, 59(3), 481–497.

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

Diakonikolas, J. and Orecchia, L. (2018). Alternating randomized block coordinate descent. In J. Dy and A. Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1224–1232. PMLR, Stockholmsmssan, Stockholm Sweden.

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. doi: 10.1109/MSP.2015.2481563.

Hong, M., Wang, X., Razaviyayn, M., and Luo, Z.Q. (2017). Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163(1), 85–114.

Luo, Z.Q. and Tseng, P. (1993). Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1), 157–178.

Nesterov, Y. (2014). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition.

Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. (2018). Primal-dual accelerated gradient methods with small-dimensional relaxation oracle.

Nutini, J., Schmidt, M., Laradji, I., Friedlander, M., and Koepe, H. (2015). Coordinate descent converges faster with the gauss-southwell rule than random selection. In F. Bach and D. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 1632–1641. PMLR, Lille, France.

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.

Polyak, B. (1987). *Introduction to Optimization*. New York, Optimization Software.

Sun, R. and Hong, M. (2015). 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, 1306–1314. MIT Press, Cambridge, MA, USA.