On the Complexity Analysis of Randomized Block-Coordinate Descent Methods

Zhaosong Lu* Lin Xiao †

May 20, 2013

Abstract

In this paper we analyze the randomized block-coordinate descent (RBCD) methods proposed in [8, 11] for minimizing the sum of a smooth convex function and a block-separable convex function. In particular, we extend Nesterov's technique developed in [8] for analyzing the RBCD method for minimizing a smooth convex function over a block-separable closed convex set to the aforementioned more general problem and obtain a sharper expected-value type of convergence rate than the one implied in [11]. Also, we obtain a better high-probability type of iteration complexity, which improves upon the one in [11] by at least the amount $O(n/\epsilon)$, where ϵ is the target solution accuracy and n is the number of problem blocks. In addition, for unconstrained smooth convex minimization, we develop a new technique called *randomized estimate sequence* to analyze the accelerated RBCD method proposed by Nesterov [8] and establish a sharper expected-value type of convergence rate than the one given in [8].

Key words: Randomized block-coordinate descent, accelerated coordinate descent, iteration complexity, convergence rate, composite minimization.

1 Introduction

Block-coordinate descent (BCD) methods and their variants have been successfully applied to solve various large-scale optimization problems (see, for example, [22, 4, 18, 19, 20, 21, 9, 23]). At each iteration, these methods choose one block of coordinates to sufficiently reduce the objective value while keeping the other blocks fixed. One common and simple approach for choosing such a block is by means of a *cyclic* strategy. The global and local convergence of the cyclic BCD method have been well studied in the literature (see, for example, [17, 5]) though its global convergence rate still remains unknown except for some special cases [13].

*Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada. (email: zhaosong@sfu.ca). This author was supported in part by NSERC Discovery Grant.

†Machine Learning Groups, Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA. (email: lin.xiao@microsoft.com).

Instead of using a deterministic cyclic order, recently many researchers proposed randomized strategies for choosing a block to update at each iteration of the BCD methods [1, 2, 14, 3, 8, 10, 11, 15, 12, 16]. The resulting methods are called randomized BCD (RBCD) methods. Numerous experiments have demonstrated that the RBCD methods are very powerful for solving large- and even huge-scale optimization problems arising in machine learning [1, 2, 14, 15]. In particular, Chang et al. [1] proposed a RBCD method for minimizing several smooth functions appearing in machine learning and derived its iteration complexity. Shalev-Shwartz and Tewari [14] studied a RBCD method for minimizing l_1-regularized smooth convex problems. They first transformed the problem into a box-constrained smooth problem by doubling the dimension and then applied a block-coordinate gradient descent method in which each block was chosen with equal probability. Leventhal and Lewis [3] proposed a RBCD method for minimizing a convex quadratic function and established its iteration complexity. Nesterov [8] analyzed some RBCD methods for minimizing a smooth convex function over a closed block-separable convex set and established its iteration complexity, which in effect extends and improves upon some of the results in [1, 3, 14] in several aspects. Richtárik and Takáč [11] generalized the RBCD methods proposed in [8] to the problem of minimizing a composite objective (i.e., the sum of a smooth convex function and a block-separable convex function) and derived some improved complexity results than those given in [8]. More recently, Shalev-Shwartz and Zhang [15] studied a randomized proximal coordinate ascent method for solving the dual of a class of large-scale convex minimization problems arising in machine learning and established iteration complexity for obtaining a pair of approximate primal-dual solutions.

Inspired by the recent work [8, 11], we consider the problem of minimizing the sum of two convex functions:

$$\min_{x\in\Re^N} \left\{F(x) \stackrel{\text{def}}{=} f(x) + \Psi(x)\right\}, \tag{1}$$

where f is differentiable on $\Re^N$, and Ψ has a block separable structure. More specifically,

$$\Psi(x) = \sum_{i=1}^{n} \Psi_i(x_i),$$

where each x_i denotes a subvector of x with cardinality N_i, the collection $\{x_i : i = 1, \ldots, n\}$ form a partition of the components of x, and each $\Psi_i : \Re^{N_i} \to \Re \cup \{+\infty\}$ is a closed convex function. Given the current iterate x^k, the RBCD method [11] picks a block $i \in \{1, \ldots, n\}$ uniformly at random and solves a block-wise proximal subproblem in the form of

$$d_i(x^k) := \arg\min_{d_i\in\Re^{N_i}} \left\{\langle \nabla_i f(x^k), d_i\rangle + \frac{L_i}{2}\|d_i\|^2 + \Psi_i(x_i^k + d_i)\right\},$$

and then it sets the next iterate as $x_i^{k+1} = x_i^k + d_i(x)$ and $x_j^{k+1} = x_j^k$ for all $j \neq i$. Here $\nabla_i f(x)$ denotes the *partial gradient* of f with respect to x_i, and L_i is the Lipschitz constant of the partial gradient (which will be defined precisely later).

Under the assumption that the partial gradients of f with respect to each block coordinate are Lipschitz continuous, Nesterov [8] studied RBCD methods for solving some *special*

cases of problem (1). In particular, for $\Psi \equiv 0$, he proposed a RBCD method in which a random block is chosen per iteration according to a uniform or certain non-uniform probability distributions and established an expected-value type of convergence rate. In addition, he proposed a RBCD method for solving (1) with each Ψ_i being the indicator function of a closed convex set, in which a random block is chosen uniformly at each iteration. He also derived an expected-value type of convergence rate for this method. It can be observed that the techniques used by Nesterov to derive these two convergence rates substantially differ from each other, and moreover, for $\Psi \equiv 0$ the second rate is much better than the first one. (However, the second technique can only work with uniform distribution.) Recently, Richtárik and Takáč [11] extended Nesterov's RBCD methods to the *general* form of problem (1) and established a high-probability type of iteration complexity. Although the expected-value type of convergence rate is not presented explicitly in [11], it can be readily obtained from some intermediate result developed in [11] (see Section 3 for a detailed discussion). Their results can be considered as a generalization of Nesterov's first technique mentioned above. Given that for $\Psi \equiv 0$ Nesterov's second technique can produce a better convergence rate than his first one, a natural question is whether his second technique can be extended to work with the general setting of problem (1) and obtain a sharper convergence rate than the one implied in [11] .

In addition, Nesterov [8] proposed an accelerated RBCD (ARCD) method for solving problem (1) with $\Psi \equiv 0$ and established an expected-value type of convergence rate for his method. When $n = 1$, this method becomes a deterministic accelerated full gradient method for minimizing smooth convex functions. When f is a strongly convex function, the convergence rate given in [8] for $n = 1$ is, however, worse than the well-known optimal rate shown in [6, Theorem 2.2.2]. Then the question is whether a sharper convergence rate for the ARCD method than the one given in [8] can be established (which would match the optimal rate for $n = 1$).

In this paper, we successfully address the above two questions by obtaining some sharper convergence rates for the RBCD method for solving problem (1) and for the ARCD method in the case $\Psi \equiv 0$. First, we extend Nesterov's second technique [8] developed for a special case of (1) to analyze the RBCD method in the general setting, and obtain a sharper expected-value type of convergence rate than the one implied in [11]. We also obtain a better high-probability type of iteration complexity, which improves upon the one in [11] at least by the amount $O(n/\epsilon)$, where ϵ is the target solution accuracy.

For unconstrained smooth convex minimization (i.e., $\Psi \equiv 0$), we develop a new technique called *randomized estimate sequence* to analyze Nesterov's ARCD method and establish a sharper expected-value type of convergence rate than the one given in [8]. Especially, for $n = 1$, our rate becomes the same as the well-known optimal rate achieved by accelerated full gradient method [6, Section 2.2].

This paper is organized as follows. In Section 2, we develop some technical results that are used to analyze the RBCD methods. In Section 3, we analyze the RBCD method for problem (1) by extending Nesterov's second technique [8], and establish a sharper expected-value type of converge rate as well as improved high-probability iteration complexity. In

Section 4, we develop the randomized estimate sequence technique and use it to derive a sharper expected-value type of converge rate for the ARCD method for solving unconstrained smooth convex minimization.

2 Technical preliminaries

In this section we develop some technical results that will be used to analyze the RBCD and ARCD methods subsequently. Throughout this paper we assume that problem (1) has a minimum ($F^\star > -\infty$) and its set of optimal solutions, denoted by X^*, is nonempty.

For any partition of $x \in \Re^N$ into $\{x_i \in \Re^{N_i} : i = 1, \ldots, n\}$, there is an $N \times N$ permutation matrix U partitioned as $U = [U_1 \cdots U_n]$, where $U_i \in \Re^{N \times N_i}$, such that

$$x = \sum_{i=1}^{n} U_i x_i, \qquad \text{and} \quad x_i = U_i^T x, \quad i = 1, \ldots, n.$$

For any $x \in \Re^N$, the *partial gradient* of f with respect to x_i is defined as

$$\nabla_i f(x) = U_i^T \nabla f(x), \quad i = 1, \ldots, n.$$

For simplicity of presentation, we associate each subspace $\Re^{N_i}$, for $i = 1, \ldots, n$, with the standard Euclidean norm, denoted by $\|\cdot\|$. We make the following assumption which is used in [8, 11] as well.

Assumption 1. *The gradient of function f is block-wise Lipschitz continuous with constants L_i, i.e.,*

$$\|\nabla_i f(x + U_i h_i) - \nabla_i f(x)\| \leq L_i \|h_i\|, \quad \forall\, h_i \in \mathbb{R}^{N_i}, \quad i = 1, \ldots, n, \quad x \in \mathbb{R}^N.$$

Following [8], we define the following pair of norms in the whole space $\Re^N$:

$$\begin{aligned} \|x\|_L &= \left(\sum_{i=1}^{n} L_i \|x_i\|^2 \right)^{1/2}, \quad \forall\, x \in \Re^N, \\ \|g\|_L^* &= \left(\sum_{i=1}^{n} \frac{1}{L_i} \|g_i\|^2 \right)^{1/2}, \quad \forall\, g \in \Re^N. \end{aligned}$$

Clearly, they satisfy the Cauchy-Schwartz inequality:

$$\langle g, x \rangle \leq \|x\|_L \cdot \|g\|_L^*, \quad \forall\, x, g \in \Re^N.$$

The convexity parameter of a convex function $\phi : \Re^N \to \Re \cup \{+\infty\}$ with respect to the norm $\|\cdot\|_L$, denoted by μ_ϕ, is the largest $\mu \geq 0$ such that for all $x, y \in \operatorname{dom} \phi$,

$$\phi(y) \geq \phi(x) + \langle s, y - x \rangle + \frac{\mu}{2} \|y - x\|_L^2, \quad \forall\, s \in \partial\phi(x).$$

Clearly, ϕ is strongly convex if and only if $\mu_\phi > 0$.

Assume that f and Ψ have convexity parameters $\mu_f \geq 0$ and $\mu_\Psi \geq 0$ with respect to the norm $\|\cdot\|_L$, respectively. Then the convexity parameter of $F = f + \Psi$ is at least $\mu_f + \mu_\Psi$. Moreover, by Assumption 1, we have

$$f(x + U_i h_i) \leq f(x) + \langle \nabla_i f(x), h_i \rangle + \frac{L_i}{2}\|h_i\|^2, \quad \forall\, h_i \in \mathbb{R}^{N_i}, \quad i = 1, \ldots, n, \quad x \in \mathbb{R}^N, \quad (2)$$

which immediately implies that $\mu_f \leq 1$.

The following lemma concerns the expected value of a block-separable function when a random block of coordinate is updated.

Lemma 1. *Suppose that $\Phi(x) = \sum_{i=1}^n \Phi_i(x_i)$. For any $x, d \in \Re^N$, if we pick $i \in \{1, \ldots, n\}$ uniformly at random, then*

$$\mathbf{E}_i\big[\Phi(x + U_i d_i)\big] = \frac{1}{n}\Phi(x + d) + \frac{n-1}{n}\Phi(x).$$

Proof. Since each i is picked randomly with probability $1/n$, we have

$$\begin{aligned}
\mathbf{E}_i\big[\Phi(x + U_i d_i)\big] &= \frac{1}{n}\sum_{i=1}^n \left(\Phi_i(x_i + d_i) + \sum_{j \neq i} \Phi_j(x_j)\right) \\
&= \frac{1}{n}\sum_{i=1}^n \Phi_i(x_i + d_i) + \frac{1}{n}\sum_{i=1}^n \sum_{j \neq i} \Phi_j(x_j) \\
&= \frac{1}{n}\Phi(x + d) + \frac{n-1}{n}\Phi(x).
\end{aligned}$$

□

For notational convenience, we define

$$H(x, d) \;:=\; f(x) + \langle \nabla f(x), d \rangle + \frac{1}{2}\|d\|_L^2 + \Psi(x + d). \quad (3)$$

The following result is equivalent to [11, Lemma 2].

Lemma 2. *Suppose $x, d \in \Re^N$. If we pick $i \in \{1, \ldots, n\}$ uniformly at random, then*

$$\mathbf{E}_i\big[F(x + U_i d_i)\big] - F(x) \;\leq\; \frac{1}{n}\big(H(x, d) - F(x)\big).$$

We next develop some results regarding the *block-wise composite gradient mapping*. Composite gradient mapping was introduced by Nesterov [7] for the analysis of full gradient methods for solving problem (1). Here we extend the concept and several associated properties to the block-coordinate case.

As mentioned in the introduction, the RBCD methods studied in [11] solves in each iteration a block-wise proximal subproblem in the form of:

$$d_i(x) := \arg\min_{d_i \in \Re^{N_i}} \left\{ \langle \nabla_i f(x), d_i \rangle + \frac{L_i}{2}\|d_i\|^2 + \Psi_i(x_i + d_i) \right\},$$

for some $i \in \{1, \ldots, n\}$. By the first-order optimality condition, there exists a subgradient $s_i \in \partial\Psi_i(x_i + d_i(x))$ such that

$$\nabla_i f(x) + L_i d_i(x) + s_i = 0. \tag{4}$$

Let $d(x) = \sum_{i=1}^n U_i d_i(x)$. By (3), the definition of $\|\cdot\|_L$ and separability of Ψ, we then have

$$d(x) = \arg\min_{d \in \Re^N} H(x, d).$$

We define the block-wise composite gradient mappings as

$$g_i(x) \stackrel{\text{def}}{=} -L_i d_i(x), \quad i = 1, \ldots, n.$$

From the optimality conditions (4), we conclude

$$-\nabla_i f(x) + g_i(x) \in \partial\Psi_i(x_i + d_i(x)), \quad i = 1, \ldots, n.$$

Let

$$g(x) = \sum_{i=1}^n U_i g_i(x).$$

Then we have

$$-\nabla f(x) + g(x) \in \partial\Psi(x + d(x)). \tag{5}$$

Moreover,

$$\|d(x)\|_L^2 = \sum_{i=1}^n L_i\|d_i(x)\|^2 = \sum_{i=1}^n \frac{1}{L_i}\|g_i(x)\|^2 = \big(\|g(x)\|_L^*\big)^2,$$

and

$$\langle g(x), d(x)\rangle = -\|d(x)\|_L^2 = -\big(\|g(x)\|_L^*\big)^2. \tag{6}$$

The following result establishes a lower bound of the function value $F(y)$, where y is arbitrary in $\Re^N$, based on the composite gradient mapping at another point x.

Lemma 3. *For any fixed $x, y \in \Re^N$, if we pick $i \in \{1, \ldots, n\}$ uniformly at random, then*

$$\begin{aligned}\frac{1}{n}F(y) + \frac{n-1}{n}F(x) \;\geq\; & \mathbf{E}_i\big[F(x + U_i d_i(x))\big] + \frac{1}{n}\left(\langle g(x), y - x\rangle + \frac{1}{2}\big(\|g(x)\|_L^*\big)^2\right) \\ & + \frac{1}{n}\left(\frac{\mu_f}{2}\|x - y\|_L^2 + \frac{\mu_\Psi}{2}\|x + d(x) - y\|_L^2\right).\end{aligned}$$

Proof. By (5) and convexity of f and Ψ, we have

$$\begin{aligned} H(x,d(x)) &= f(x)+\langle\nabla f(x),d(x)\rangle+\frac{1}{2}\|d(x)\|_L^2+\Psi(x+d(x)) \\ &\le f(y)+\langle\nabla f(x),x-y\rangle-\frac{\mu_f}{2}\|x-y\|_L^2+\langle\nabla f(x),d(x)\rangle+\frac{1}{2}\|d(x)\|_L^2 \\ &\quad +\Psi(y)+\langle-\nabla f(x)+g(x),x+d(x)-y\rangle-\frac{\mu_\Psi}{2}\|x+d(x)-y\|_L^2 \\ &= F(y)+\langle g(x),x-y\rangle+\langle g(x),d(x)\rangle+\frac{1}{2}\|d(x)\|_L^2-\frac{\mu_f}{2}\|x-y\|_L^2 \\ &\quad -\frac{\mu_\Psi}{2}\|x+d(x)-y\|_L^2 \\ &= F(y)+\langle g(x),x-y\rangle-\frac{1}{2}\big(\|g(x)\|_L^*\big)^2-\frac{\mu_f}{2}\|x-y\|_L^2-\frac{\mu_\Psi}{2}\|x+d(x)-y\|_L^2, \end{aligned}$$

where the last inequality holds due to (6). This together with Lemma 2 yields the desired result. □

Using Lemma 1 with $\Phi(\cdot)=\|\cdot\|_L^2$, we can rewrite the conclusion of Lemma 3 in an equivalent form:

$$\begin{aligned} \frac{1}{n}F(y)+\frac{n-1}{n}F(x)\ +\ \frac{\mu_\Psi}{2}\|x-y\|_L^2\ \ge\ \mathbf{E}_i\left[F(x+U_id_i(x))+\frac{\mu_\Psi}{2}\|x+U_id_i-y\|_L^2\right] \\ +\ \frac{1}{n}\left(\langle g(x),y-x\rangle+\frac{1}{2}\big(\|g(x)\|_L^*\big)^2+\frac{\mu_f+\mu_\Psi}{2}\|x-y\|_L^2\right). \end{aligned} \tag{7}$$

This is the form we will actually use in our subsequent convergence analysis.

Letting $y=x$ in Lemma 3, we obtain the following corollary.

Corollary 1. *Given $x\in\Re^N$. If we pick $i\in\{1,\dots,n\}$ uniformly at random, then*

$$F(x)-\mathbf{E}_i\big[F(x+U_id_i(x))\big]\ \ge\ \frac{1+\mu_\Psi}{2n}\big(\|g(x)\|_L^*\big)^2\ =\ \frac{1+\mu_\Psi}{2n}\big(\|d(x)\|_L\big)^2.$$

By similar arguments as in the proof of Lemma 3, it can be shown that a similar result as Lemma 3 also holds block-wise without taking expectation:

$$F(x)-F(x+U_id_i(x))\ge\frac{1+\mu_\Psi}{2}L_i\|d_i(x)\|^2.$$

The following (trivial) corollary is useful when we do not have knowledge on μ_f or μ_Ψ.

Corollary 2. *For any fixed $x,y\in\Re^N$, if we pick $i\in\{1,\dots,n\}$ uniformly at random, then*

$$\frac{1}{n}F(y)+\frac{n-1}{n}F(x)\ \ge\ \mathbf{E}_i\big[F(x+U_id_i(x))\big]+\frac{1}{n}\left(\langle g(x),y-x\rangle+\frac{1}{2}\big(\|g(x)\|_L^*\big)^2\right).$$

3 Randomized block-coordinate descent

In this section we analyze the following randomized block coordinate descent (RBCD) method for solving problem (1), which was proposed in [11]. In particular, we extend Nesterov's technique [8] developed for a special case of problem (1) to work with the general setting and establish some sharper expected-value type of converge rate, as well as improved high-probability iteration complexity, than those given or implied in [11].

Algorithm: RBCD(x^0)

Repeat for $k = 0, 1, 2, \ldots$

1. Choose $i_k \in \{1, \ldots, n\}$ randomly with a uniform distribution.
2. Update $x^{k+1} = x^k + U_{i_k} d_{i_k}(x^k)$.

After k iterations, the RBCD method generates a random output x^k, which depends on the observed realization of the random variable

$$\xi_{k-1} \stackrel{\text{def}}{=} \{i_0, i_1, \ldots, i_{k-1}\}.$$

The following quantity measures the distance between x^0 and the optimal solution set of problem (1) that will appear in our complexity results:

$$R_0 \stackrel{\text{def}}{=} \min_{x^\star \in X^*} \|x^0 - x^\star\|_L, \tag{8}$$

where X^* is the set of optimal solutions of problem (1).

3.1 Convergence rate of expected values

The following theorem is a generalization of [8, Theorem 5], where the function Ψ in (1) is restricted to be the indicator function of a block-separable closed convex set. Here we extend it to the general case of Ψ being block-separable convex functions by employing the machinery of block-wise composite gradient mapping developed in Section 2.

Theorem 1. *Let R_0 be defined in (8), $F^\star$ be the optimal value of problem (1), and $\{x^k\}$ be the sequence generated by the RBCD method. Then for any $k \geq 0$, the iterate x^k satisfies*

$$\mathbf{E}_{\xi_{k-1}}\big[F(x^k)\big] - F^\star \;\leq\; \frac{n}{n+k}\left(\frac{1}{2}R_0^2 + F(x^0) - F^\star\right). \tag{9}$$

Furthermore, if at least one of f and Ψ is strongly convex, i.e., $\mu_f + \mu_\Psi > 0$, then

$$\mathbf{E}_{\xi_{k-1}}\big[F(x^k)\big] - F^\star \;\leq\; \left(1 - \frac{2(\mu_f + \mu_\Psi)}{n(1 + \mu_f + 2\mu_\Psi)}\right)^k \left(\frac{1 + \mu_\Psi}{2}R_0^2 + F(x^0) - F^\star\right). \tag{10}$$

Proof. Let $x^\star$ be an arbitrary optimal solution of (1). Denote

$$r_k^2 \;=\; \|x^k - x^\star\|_L^2 \;=\; \sum_{i=1}^n L_i \langle x_i^k - x_i^\star, x_i^k - x_i^\star \rangle.$$

Notice that $x^{k+1} = x^k + U_{i_k} d_{i_k}(x^k)$. Thus we have

$$r_{k+1}^2 \;=\; r_k^2 + 2L_{i_k}\langle d_{i_k}(x^k), x_{i_k}^k - x_{i_k}^\star \rangle + L_{i_k}\|d_{i_k}(x^k)\|^2.$$

Multiplying both sides by $1/2$ and taking expectation with respect to i_k yield

$$\begin{aligned}
\mathbf{E}_{i_k}\left[\frac{1}{2}r_{k+1}^2\right] &= \frac{1}{2}r_k^2 + \frac{1}{n}\left(\sum_{i=1}^n L_i\langle d_i(x^k), x_i^k - x_i^\star\rangle + \frac{1}{2}\sum_{i=1}^n \frac{1}{L_i}\|g_i(x^k)\|^2\right) \\
&= \frac{1}{2}r_k^2 + \frac{1}{n}\left(\langle g(x^k), x^\star - x^k\rangle + \frac{1}{2}\big(\|g(x^k)\|_L^*\big)^2\right). \qquad (11)
\end{aligned}$$

Using Corollary 2, we obtain

$$\mathbf{E}_{i_k}\left[\frac{1}{2}r_{k+1}^2\right] \;\le\; \frac{1}{2}r_k^2 + \frac{1}{n}F^\star + \frac{n-1}{n}F(x^k) - \mathbf{E}_{i_k}F(x^{k+1}).$$

By rearranging terms, we obtain that for each $k \ge 0$,

$$\mathbf{E}_{i_k}\left[\frac{1}{2}r_{k+1}^2 + F(x^{k+1}) - F^\star\right] \;\le\; \left(\frac{1}{2}r_k^2 + F(x^k) - F^\star\right) - \frac{1}{n}\left(F(x^k) - F^\star\right).$$

Taking expectation with respect to ξ_{k-1} on both sides of the above inequality, we have

$$\mathbf{E}_{\xi_k}\left[\frac{1}{2}r_{k+1}^2 + F(x^{k+1}) - F^\star\right] \;\le\; \mathbf{E}_{\xi_{k-1}}\left[\frac{1}{2}r_k^2 + F(x^k) - F^\star\right] - \frac{1}{n}\mathbf{E}_{\xi_{k-1}}\left[F(x^k) - F^\star\right].$$

Applying this inequality recursively and using the fact that $\mathbf{E}_{\xi_k}\big[F(x^j)\big]$ is monotonically decreasing for $j = 0, \ldots, k+1$ (see Corollary 1), we further obtain that

$$\begin{aligned}
\mathbf{E}_{\xi_k}\big[F(x^{k+1})\big] - F^\star &\le \mathbf{E}_{\xi_k}\left[\frac{1}{2}r_{k+1}^2 + F(x^{k+1}) - F^\star\right] \\
&\le \frac{1}{2}r_0^2 + F(x^0) - F^\star - \frac{1}{n}\sum_{j=0}^k \left(\mathbf{E}_{\xi_k}\big[F(x^j)\big] - F^\star\right) \\
&\le \frac{1}{2}r_0^2 + F(x^0) - F^\star - \frac{k+1}{n}\left(\mathbf{E}_{\xi_k}\big[F(x^{k+1})\big] - F^\star\right).
\end{aligned}$$

This leads to

$$\mathbf{E}_{\xi_k}\big[F(x^{k+1})\big] - F^\star \;\le\; \frac{n}{n+k+1}\left(\frac{1}{2}\|x^0 - x^\star\|_L^2 + F(x^0) - F^\star\right),$$

which together with the arbitrariness of $x^\star$ and the definition of R_0 yields (9).

Next we prove (10) under the strong convexity assumption $\mu_f + \mu_\Psi > 0$. Using (7) and (11), we obtain that

$$\begin{aligned}\mathbf{E}_{i_k}\left[\frac{1+\mu_\Psi}{2}r_{k+1}^2 + F(x^{k+1}) - F^\star\right] \;\leq\; & \left(\frac{1+\mu_\Psi}{2}r_k^2 + F(x^k) - F^\star\right) \\ & -\frac{1}{n}\left(\frac{\mu_f+\mu_\Psi}{2}r_k^2 + F(x^k) - F^\star\right). \qquad (12)\end{aligned}$$

By strong convexity of F, we have

$$\frac{\mu_f+\mu_\Psi}{2}r_k^2 + F(x^k) - F^\star \;\geq\; \frac{\mu_f+\mu_\Psi}{2}r_k^2 + \frac{\mu_f+\mu_\Psi}{2}r_k^2 \;=\; (\mu_f+\mu_\Psi)r_k^2.$$

Define

$$\beta = \frac{2(\mu_f+\mu_\Psi)}{1+\mu_f+2\mu_\Psi}.$$

We have $0 < \beta \leq 1$ due to $\mu_f + \mu_\Psi > 0$ and $\mu_f \leq 1$. Then

$$\begin{aligned}\frac{\mu_f+\mu_\Psi}{2}r_k^2 + F(x^k) - F^\star \;&\geq\; \beta\left(\frac{\mu_f+\mu_\Psi}{2}r_k^2 + F(x^k) - F^\star\right) + (1-\beta)(\mu_f+\mu_\Psi)r_k^2 \\ &=\; \beta\left(\frac{1+\mu_\Psi}{2}r_k^2 + F(x^k) - F^\star\right).\end{aligned}$$

Combining the above inequality with (12) gives

$$\mathbf{E}_{i_k}\left[\frac{1+\mu_\Psi}{2}r_{k+1}^2 + F(x^{k+1}) - F^\star\right] \;\leq\; \left(1-\frac{\beta}{n}\right)\left(\frac{1+\mu_\Psi}{2}r_k^2 + F(x^k) - F^\star\right)$$

Taking expectation with respect ξ_{k-1} on both sides of the above relation, we have

$$\mathbf{E}_{\xi_k}\left[\frac{1+\mu_\Psi}{2}r_{k+1}^2 + F(x^{k+1}) - F^\star\right] \;\leq\; \left(1-\frac{\beta}{n}\right)^{k+1}\left(\frac{1+\mu_\Psi}{2}r_0^2 + F(x^0) - F^\star\right),$$

which together with the arbitrariness of $x^\star$ and the definition of R_0 leads to (10). □

We have the following remarks on comparing the results in Theorem 1 with those in [11].

- For the *general* setting of problem (1), expected-value type of convergence rate is not presented explicitly in [11]. Nevertheless, it can be derived straightforwardly from the following relation that was proved in [11, Theorem 5]:

$$\mathbf{E}_{i_k}[\Delta_{k+1}] \;\leq\; \Delta_k - \frac{\Delta_k^2}{2nc}, \qquad \forall k \geq 0, \qquad (13)$$

where $\Delta_k := F(x^k) - F^\star$, and

$$c := \max\{\bar{R}_0^2,\ F(x^0) - F^\star\}, \tag{14}$$

$$\bar{R}_0 := \max_x \Big\{ \max_{x^\star \in X^*} \|x - x^\star\|_L : F(x) \le F(x^0) \Big\}. \tag{15}$$

Taking expectation with respect to ξ_{k-1} on both sides of (13), one can have

$$\mathbf{E}_{\xi_k}[\Delta_{k+1}] \ \le\ \mathbf{E}_{\xi_{k-1}}[\Delta_k] - \frac{1}{2nc}\left(\mathbf{E}_{\xi_{k-1}}[\Delta_k]\right)^2, \qquad \forall k \ge 0.$$

By this relation and a similar argument as used in the proof of [8, Theorem 1], one can obtain that

$$\mathbf{E}_{\xi_{k-1}}[F(x^k)] - F^\star \ \le\ \frac{2nc(F(x^0) - F^\star)}{k(F(x^0) - F^\star) + 2nc}, \qquad \forall k \ge 0. \tag{16}$$

Let a and b denote the right-hand side of (9) and (16), respectively. By the definition of c and the relation $\bar{R}_0 \ge R_0$, we can see that when k is sufficiently large,

$$\frac{b}{a} \ \approx\ \frac{2c}{\frac{1}{2}R_0^2 + F(x^0) - F^\star} \ \ge\ \frac{4}{3}.$$

Therefore, our expected-value type of convergence rate is better by at least a factor of 4/3 asymptotically, and the improvement can be much larger if $\bar{R}_0$ is much larger than R_0.

- For the *special* case of (1) where at least one of f and Ψ is strongly convex, i.e., $\mu_f + \mu_\Psi > 0$, Richtárik and Takáč [11, Theorem 7] showed that for all $k \ge 0$, there holds

$$\mathbf{E}_{\xi_{k-1}}\big[F(x^k)\big] - F^\star \ \le\ \left(1 - \frac{\mu_f + \mu_\Psi}{n(1+\mu_\Psi)}\right)^k \big(F(x^0) - F^\star\big).$$

It is not hard to observe that

$$\frac{2(\mu_f + \mu_\Psi)}{n(1 + \mu_f + 2\mu_\Psi)} \ >\ \frac{\mu_f + \mu_\Psi}{n(1+\mu_\Psi)}. \tag{17}$$

It then follows that for sufficiently large k, one has

$$\begin{aligned}
&\left(1 - \frac{2(\mu_f + \mu_\Psi)}{n(1 + \mu_f + 2\mu_\Psi)}\right)^k \left(\frac{1+\mu_\Psi}{2} R_0^2 + F(x^0) - F^\star\right) \\
\le\ &\left(1 - \frac{2(\mu_f + \mu_\Psi)}{n(1 + \mu_f + 2\mu_\Psi)}\right)^k \left(\frac{1 + \mu_f + \mu_\Psi}{\mu_f + \mu_\Psi}\right) \big(F(x^0) - F^\star\big) \\
\ll\ &\left(1 - \frac{\mu_f + \mu_\Psi}{n(1+\mu_\Psi)}\right)^k \big(F(x^0) - F^\star\big).
\end{aligned}$$

Therefore, our convergence rate (10) is much sharper than their rate for sufficiently large k.

3.2 High probability complexity bound

By virtue of Theorem 1 we can also derive a sharper iteration complexity for a *single run* of the RBCD method for obtaining an ϵ-optimal solution with high probability than the one given in [11, Theorems 5 and 7].

Theorem 2. *Let R_0 be defined in (8) and $\{x^k\}$ be the sequence generated by the RBCD method. Let $0 < \epsilon < F(x^0) - F^\star$ and $\rho \in (0,1)$ be chosen arbitrarily.*

(i) For all $k \geq K$, there holds

$$\mathbf{P}(F(x^k) - F^\star \leq \epsilon) \;\geq\; 1 - \rho, \tag{18}$$

where

$$K := \frac{2nc}{\epsilon}\left(1 + \log\left(\frac{R_0^2 + 2[F(x^0) - F^\star]}{4c\rho}\right)\right) + 2 - n. \tag{19}$$

(ii) Furthermore, if at least one of f and Ψ is strongly convex, i.e., $\mu_f + \mu_\Psi > 0$, then (18) holds when $k \geq \tilde{K}$, where

$$\tilde{K} := \frac{n(1 + \mu_f + 2\mu_\Psi)}{2(\mu_f + \mu_\Psi)} \log\left(\frac{\frac{1+\mu_\Psi}{2} R_0^2 + F(x^0) - F^\star}{\rho\epsilon}\right)$$

Proof. (i) For convenience, let $\Delta_k = F(x^k) - F^\star$ for all k. Define the truncated sequence $\{\Delta_k^\epsilon\}$ as follows:

$$\Delta_k^\epsilon = \begin{cases} \Delta_k & \text{if } \Delta_k \geq \epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

Using (13) and the same argument as used in the proof of [11, Theorem 1], one can have

$$\mathbf{E}_{i_k}[\Delta_{k+1}^\epsilon] \;\leq\; \left(1 - \frac{\epsilon}{2nc}\right)\Delta_k^\epsilon, \qquad \forall k \geq 0.$$

Taking expectation with respect to ξ_{k-1} on both sides of the above relation, we obtain that

$$\mathbf{E}_{\xi_k}[\Delta_{k+1}^\epsilon] \;\leq\; \left(1 - \frac{\epsilon}{2nc}\right)\mathbf{E}_{\xi_{k-1}}[\Delta_k^\epsilon], \qquad \forall k \geq 0. \tag{20}$$

In addition, using (9) and the relation $\Delta_k^\epsilon \leq \Delta_k$, we have

$$\mathbf{E}_{\xi_{k-1}}[\Delta_k^\epsilon] \;\leq\; \frac{n}{n+k}\left(\frac{1}{2}R_0^2 + F(x^0) - F^\star\right), \qquad \forall k \geq 0. \tag{21}$$

For any $t > 0$, let

$$K_1 = \left\lceil \frac{n}{t\epsilon}\left(\frac{1}{2}R_0^2 + F(x^0) - F^\star\right)\right\rceil - n, \qquad K_2 = \left\lceil \frac{2nc}{\epsilon}\log\left(\frac{t}{\rho}\right)\right\rceil.$$

It follows from (21) that $\mathbf{E}_{\xi_{K_1-1}}[\Delta^\epsilon_{K_1}] \leq t\epsilon$, which together with (20) implies that

$$\mathbf{E}_{\xi_{K_1+K_2-1}}[\Delta^\epsilon_{K_1+K_2}] \leq \left(1-\frac{\epsilon}{2nc}\right)^{K_2}\mathbf{E}_{\xi_{K_1-1}}[\Delta^\epsilon_{K_1}] \leq \left(1-\frac{\epsilon}{2nc}\right)^{K_2} t\epsilon \leq \rho\epsilon.$$

Notice from (20) that $\{\mathbf{E}_{\xi_{k-1}}[\Delta^\epsilon_k]\}$ is decreasing. Hence, we have

$$\mathbf{E}_{\xi_{k-1}}[\Delta^\epsilon_k] \leq \rho\epsilon, \quad \forall k \geq K(t), \tag{22}$$

where

$$K(t) := \frac{n}{t\epsilon}\left(\frac{1}{2}R_0^2 + F(x^0) - F^\star\right) + \frac{2nc}{\epsilon}\log\left(\frac{t}{\rho}\right) + 2 - n.$$

It is not hard to verify that

$$t^* := \frac{\frac{1}{2}R_0^2 + F(x^0) - F^\star}{2c} = \arg\min_{t>0} K(t).$$

Also, one can observe from (19) that $K \geq K(t^*)$, which together with (22) implies that

$$\mathbf{E}_{\xi_{k-1}}[\Delta^\epsilon_k] \leq \rho\epsilon, \quad \forall k \geq K.$$

Using this relation and Markov inequality, we obtain that

$$\mathbf{P}(F(x^k) - F^\star > \epsilon) = \mathbf{P}(\Delta_k > \epsilon) = \mathbf{P}(\Delta^\epsilon_k > \epsilon) \leq \frac{\mathbf{E}_{\xi_{k-1}}[\Delta^\epsilon_k]}{\epsilon} \leq \rho, \quad \forall k \geq K,$$

which immediately implies statement (i) holds.

(ii) Using the Markov inequality, the inequality (10) and the definition of $\tilde{K}$, we obtain that for any $k \geq \tilde{K}$,

$$\begin{aligned}
\mathbf{P}(F(x^k) - F^\star > \epsilon) &\leq \frac{\mathbf{E}_{\xi_{k-1}}[F(x^k) - F^\star]}{\epsilon} \\
&\leq \frac{1}{\epsilon}\left(1 - \frac{2(\mu_f + \mu_\Psi)}{n(1+\mu_f+2\mu_\Psi)}\right)^{\tilde{K}}\left(\frac{1+\mu_\Psi}{2}R_0^2 + F(x^0) - F^\star\right) \\
&\leq \frac{1}{\epsilon}\exp\left(-\frac{2(\mu_f+\mu_\Psi)\tilde{K}}{n(1+\mu_f+2\mu_\Psi)}\right)\left(\frac{1+\mu_\Psi}{2}R_0^2 + F(x^0) - F^\star\right) \\
&\leq \rho
\end{aligned}$$

and hence statement (ii) holds. □

We make the following remarks in comparing our results in Theorem 2 with those in [11].

- For any $0 < \epsilon < F(x^0) - F^\star$ and $\rho \in (0,1)$, Richtárik and Takáč [11, Theorem 5] showed that (18) holds for all $k \geq \bar{K}$, where

$$\bar{K} = \frac{2nc}{\epsilon}\left(1 + \log\frac{1}{\rho}\right) + 2 - \frac{2nc}{F(x^0) - F^\star}$$

and c is given in (14). Using the definitions of c and R_0 and the fact $R_0 \leq \bar{R}_0$, one can observe that

$$\tau := \frac{R_0^2 + 2[F(x^0) - F^\star]}{4c} \;\leq\; \frac{3}{4}.$$

By the definitions of K and $\bar{K}$, we have that for sufficiently small $\epsilon > 0$,

$$K - \bar{K} \;\approx\; \frac{2nc\log\tau}{\epsilon} \;\leq\; -\frac{2nc\log(4/3)}{\epsilon}.$$

In addition, by the definitions of R_0 and $\bar{R}_0$, one can see that R_0 can be much smaller than $\bar{R}_0$ and thus τ can be very small. It follows from the above relation that K can be substantially smaller than $\bar{K}$.

- For a *special* case of (1) where at least one of f and Ψ is strongly convex, i.e., $\mu_f + \mu_\Psi > 0$, Richtárik and Takáč [11, Theorem 8] showed that (18) holds for all $k \geq \hat{K}$, where

$$\hat{K} := \frac{n(1+\mu_\Psi)}{\mu_f + \mu_\Psi}\log\left(\frac{F(x^0) - F^\star}{\rho\epsilon}\right).$$

We then see that when ρ or ϵ is sufficiently small,

$$\frac{\tilde{K}}{\hat{K}} \;\approx\; \frac{1 + \mu_f + 2\mu_\Psi}{2(1+\mu_\Psi)} \;\leq\; 1$$

due to $0 \leq \mu_f \leq 1$. When $\mu_f < 1$, we have $\tilde{K} \leq \tilde{\tau}\hat{K}$ for some $\tilde{\tau} \in (0,1)$ and thus our complexity bound is tighter when ρ or ϵ is sufficiently small.

As discussed in [11, Section 2], the number of iterations required by the RBCD method for obtaining an ϵ-optimal solution with high probability can also be estimated by using a *multiple-run* strategy, each run with an independently generated random sequence $\{i_0, i_1, \ldots\}$. We next derive such an iteration complexity.

Theorem 3. *Let $0 < \epsilon < F(x^0) - F^\star$ and $\rho \in (0,1)$ be arbitrarily chosen, and let $r = \lceil \log(1/\rho) \rceil$. Suppose that we run the RBCD method starting with x^0 for r times independently, each time for the same number of iterations k. Let $x_{(j)}^k$ denote the output by the RBCD at the kth iteration of the jth run. Then there holds:*

$$\mathbf{P}\left(\min_{1\leq j\leq r} F(x_{(j)}^k) - F^\star \;\leq\; \epsilon\right) \;\geq\; 1 - \rho$$

for any $k \geq \underline{K}$, where

$$\underline{K} := \left\lceil \frac{en}{\epsilon}\left(\frac{1}{2}R_0^2 + F(x^0) - F^\star\right)\right\rceil - n.$$

Proof. Let $\xi_{k-1}^{(j)} = \left\{i_0^{(j)}, i_1^{(j)}, \ldots, i_{k-1}^{(j)}\right\}$ denote the random sequence used in the jth run. Using Markov inequality, (9) and the definition of $\underline{K}$, we obtain that for any $k \geq \underline{K}$,

$$\mathbf{P}\left(F(x_{(j)}^k) - F^\star \ > \ \epsilon\right) \ \leq \ \frac{\mathbf{E}_{\xi_{k-1}^{(j)}}[F(x_{(j)}^k) - F^\star]}{\epsilon} \ \leq \ \frac{n}{(n+k)\epsilon}\left(\frac{1}{2}R_0^2 + F(x^0) - F^\star\right) \ \leq \ \frac{1}{e}.$$

This together with the definition of r implies that

$$\mathbf{P}\left(\min_{1\leq j\leq r} F(x_{(j)}^k) - F^\star \ > \ \epsilon\right) \ = \ \Pi_{j=1}^r \mathbf{P}\left(F(x_{(j)}^k) - F^\star \ > \ \epsilon\right) \ \leq \ \frac{1}{e^r} \ \leq \ \rho,$$

and hence the conclusion holds. □

Remark. From Theorem 3, one can see that the total number of iterations by RBCD with a multiple-run strategy for obtaining an ϵ-optimal solution is at most

$$K^{\mathrm{M}} := \left(\left\lceil \frac{2en}{\epsilon}\left(R_0^2 + 2(F(x^0) - F^\star)\right)\right\rceil - n\right)\left\lceil \log\frac{1}{\rho}\right\rceil.$$

It was implicitly established in [11] that an ϵ-optimal solution can be found by RBCD with a multiple-run strategy in at most

$$\bar{K}^{\mathrm{M}} := \left\lceil \frac{2enc}{\epsilon} - \frac{2nc}{F(x^0) - F^\star}\right\rceil \left\lceil \log\frac{1}{\rho}\right\rceil$$

iterations. When ρ or ϵ is sufficiently small, we have

$$\frac{K^{\mathrm{M}}}{\bar{K}^{\mathrm{M}}} \ \approx \ \frac{R_0^2 + 2(F(x^0) - F^\star)}{c}.$$

Recall that $\bar{R}_0$ can be much larger than R_0, which together with (15) implies that c can be much larger than $R_0^2 + 2(F(x^0) - F^\star)$. It follows from the above relation that when ρ or ϵ is sufficiently small, K^{M} can be substantially smaller than $\bar{K}^{\mathrm{M}}$.

4 Accelerated randomized coordinate descent

In this section, we restrict ourselves to the unconstrained smooth minimization problem

$$\min_{x\in\Re^N} \ f(x), \tag{23}$$

where f is convex in $\Re^N$ with convexity parameter $\mu = \mu_f \geq 0$ with respect to the norm $\|\cdot\|_L$ and satisfies Assumption 1. It then follows from (2) that $\mu \leq 1$. Our aim is to analyze the convergence rate of the following accelerated randomized coordinate descent (ARCD) method.

Algorithm: ARCD(x^0)

Set $v^0 = x^0$, choose $\gamma_0 > 0$ arbitrarily, and repeat for $k = 0, 1, 2, \ldots$

1. Compute $\alpha_k \in (0, n]$ from the equation

$$\alpha_k^2 = \left(1 - \tfrac{\alpha_k}{n}\right)\gamma_k + \tfrac{\alpha_k}{n}\mu$$

and set

$$\gamma_{k+1} = \left(1 - \tfrac{\alpha_k}{n}\right)\gamma_k + \tfrac{\alpha_k}{n}\mu.$$

2. Compute y^k as

$$y^k = \tfrac{1}{\frac{\alpha_k}{n}\gamma_k + \gamma_{k+1}}\left(\tfrac{\alpha_k}{n}\gamma_k v^k + \gamma_{k+1}x^k\right).$$

3. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random, and update

$$x^{k+1} = y^k - \tfrac{1}{L_{i_k}} U_{i_k}\nabla_{i_k} f(y^k).$$

4. Set

$$v^{k+1} = \tfrac{1}{\gamma_{k+1}}\left(\left(1 - \tfrac{\alpha_k}{n}\right)\gamma_k v^k + \tfrac{\alpha_k}{n}\mu y^k - \tfrac{\alpha_k}{L_{i_k}} U_{i_k}\nabla_{i_k} f(y^k)\right).$$

Remark. For the above algorithm, claim that $\gamma_k > 0$ and α_k is well-defined for all k. Indeed, let $\gamma > 0$ be arbitrarily given and define

$$h(\alpha) := \alpha^2 - \left(1 - \frac{\alpha}{n}\right)\gamma - \frac{\alpha}{n}\mu, \qquad \forall \alpha \geq 0.$$

We observe that

$$h(0) = -\gamma \ < \ 0, \qquad h(n) = n^2 - \mu \ \geq \ 0,$$

where the last inequality is due to $\mu \leq 1$. Therefore, by continuity of h, there exists some $\alpha^* \in (0, n]$ such that $h(\alpha^*) = 0$. Moreover, if $\mu = 0$, we have $0 < \alpha^* < n$. Using these observations and the definitions of α_k and γ_k, it is not hard to see by induction that $\gamma_k > 0$ and α_k is well-defined for all k.

The above description of the ARCD method comes directly from the derivation using *randomized estimate sequence* we develop in Section 4.1, and is very convenient for the purpose of our convergence analysis. For implementation in practice, one can simplify the notations and use an equivalent algorithm described below. In the simplified description, it is also clear that the ARCD method is equivalent to the method (5.1) in [8, Section 5], with the following correspondences between the symbols used.

This paper	α_k	α_{k-1}	θ_k	β_k	μ
[8, (5.1)]	$1/\gamma_k$	b_k/a_k	α_k	β_k	σ

Algorithm: ARCD(x^0)

Set $v^0 = x^0$, choose $\alpha_{-1} \in (0, n]$, and repeat for $k = 0, 1, 2, \ldots$

1. Compute $\alpha_k \in (0, n]$ from the equation

$$\alpha_k^2 = \left(1 - \tfrac{\alpha_k}{n}\right) \alpha_{k-1}^2 + \tfrac{\alpha_k}{n}\mu,$$

and set

$$\theta_k = \tfrac{n\alpha_k - \mu}{n^2 - \mu}, \qquad \beta_k = 1 - \tfrac{\mu}{n\alpha_k}.$$

2. Compute y^k as

$$y^k \;=\; \theta_k v^k + (1 - \theta_k) x^k.$$

3. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random, and update

$$x^{k+1} = y^k - \tfrac{1}{L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k).$$

4. Set

$$v^{k+1} = \beta_k v^k + (1 - \beta_k) y^k - \tfrac{1}{\alpha_k L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k).$$

At each iteration k, the ARCD method generates y^k, x^{k+1} and v^{k+1}. One can observe that x^{k+1} and v^{k+1} depend on the realization of the random variable

$$\xi_k = \{i_0, i_1, \ldots, i_k\}$$

while y^k depends on the realization of ξ_{k-1}.

We now state a sharper expected-value type of convergence rate for the ARCD method than the one given in [8]. Its proof relies on a new technique called *randomized estimate sequence* that will be developed in Subsection 4.1. Therefore, we postpone the proof to Subsection 4.2.

Theorem 4. *Let $f^\star$ be the optimal value of problem (23), R_0 be defined in (8), and $\{x^k\}$ be the sequence generated by the ARCD method. Then, for any $k \geq 0$, there holds:*

$$\mathbf{E}_{\xi_{k-1}}[f(x^k)] - f^\star \;\leq\; \lambda_k \left(f(x^0) - f^\star + \frac{\gamma_0 R_0^2}{2} \right),$$

where $\lambda_0 = 1$ and $\lambda_k = \prod_{i=0}^{k-1} \left(1 - \tfrac{\alpha_i}{n}\right)$. In particular, if $\gamma_0 \geq \mu$, then

$$\lambda_k \;\leq\; \min\left\{ \left(1 - \frac{\sqrt{\mu}}{n}\right)^k, \;\; \left(\frac{n}{n + k\frac{\sqrt{\gamma_0}}{2}}\right)^2 \right\}.$$

Remark. We note that for $n = 1$, the ARCD method reduces to a deterministic accelerated full gradient method described in [6, (2.2.8)]; Our iteration complexity result above also becomes the same as the one given there.

Nesterov [8, Theorem 6] established the following convergence rate for the above ARCD method:

$$\mathbf{E}_{\xi_{k-1}}[f(x^k)]-f^\star \;\leq\; \begin{cases} \overbrace{\mu\left[2R_0^2+\frac{1}{n^2}(f(x^0)-f^\star)\right]\cdot\left[\left(1+\frac{\sqrt{\mu}}{2n}\right)^{k+1}-\left(1-\frac{\sqrt{\mu}}{2n}\right)^{k+1}\right]^{-2}}^{a_\mu} & \text{if } \mu>0, \\ \underbrace{\left(\frac{n}{k+1}\right)^2\cdot\left[2R_0^2+\frac{1}{n^2}(f(x^0)-f^\star)\right]}_{a_0} & \text{otherwise.} \end{cases}$$

In view of Theorem 4, our convergence rate is given by

$$\mathbf{E}_{\xi_{k-1}}[f(x^k)]-f^\star \;\leq\; \underbrace{\min\left\{\left(1-\frac{\sqrt{\mu}}{n}\right)^k,\;\left(\frac{n}{n+k\frac{\sqrt{\gamma_0}}{2}}\right)^2\right\}\left(f(x^0)-f^\star+\frac{\gamma_0 R_0^2}{2}\right)}_{b_\mu}$$

We now compare the above two rates by considering two cases: $\mu>0$ and $\mu=0$.

- Case (1): $\mu>0$. We can observe that for sufficiently large k,

$$a_\mu \;=\; O\left(\left(1+\frac{\sqrt{\mu}}{2n}\right)^{-2k}\right), \qquad b_\mu \;=\; O\left(\left(1-\frac{\sqrt{\mu}}{n}\right)^{k}\right).$$

  It is easy to verify that

$$\left(1+\frac{\sqrt{\mu}}{2n}\right)^{-2} \;>\; 1-\frac{\sqrt{\mu}}{n}$$

  and hence $a_\mu \gg b_\mu$ when k is sufficiently large, which implies that our rate is much tighter.

- Case (2): $\mu=0$. For sufficiently k, we have

$$\begin{aligned} a_0 &\approx (2n^2R_0^2+f(x^0)-f^\star)/k^2, \\ b_0 &\approx \left(2n^2R_0^2+\tfrac{4n^2}{\gamma_0}(f(x^0)-f^\star)\right)/k^2. \end{aligned}$$

  Therefore, when $\gamma_0>4n^2$, we obtain $b_0<a_0$ for sufficiently large k, which again implies that our rate is sharper.

4.1 Randomized estimate sequence

In [6], Nesterov introduced a powerful framework of *estimate sequence* for the development and analysis of accelerated full gradient methods. Here we extend it to a randomized block-coordinate descent setup, and use it to analyze the convergence rate of the ARCD method subsequently.

Definition 1. *Let $\phi_0(x)$ be a deterministic function and $\phi_k(x)$ be a random function depending on ξ_{k-1} for all $k \geq 1$, and $\lambda_k \geq 0$ for all $k \geq 0$. The sequence $\{(\phi_k(x), \lambda_k)\}_{k=0}^{\infty}$ is called a randomized estimate sequence of function $f(x)$ if*

$$\lambda_k \to 0 \tag{24}$$

and for any $x \in \Re^N$ and all $k \geq 0$ we have

$$\mathbf{E}_{\xi_{k-1}}[\phi_k(x)] \;\leq\; (1-\lambda_k) f(x) + \lambda_k \phi_0(x), \tag{25}$$

where $\mathbf{E}_{\xi_{-1}}[\phi_0(x)] \stackrel{\text{def}}{=} \phi_0(x)$.

Here we assume $\{\lambda_k\}_{k\geq 0}$ is a deterministic sequence that is independent of ξ_k.

Lemma 4. *Let $x^\star$ be an optimal solution to (23) and $f^\star$ be the optimal value. Suppose that $\{(\phi_k(x), \lambda_k)\}_{k=0}^{\infty}$ is a randomized estimate sequence of function $f(x)$. Assume that $\{x^k\}$ is a sequence such that for each $k \geq 0$,*

$$\mathbf{E}_{\xi_{k-1}}[f(x^k)] \;\leq\; \min_x \mathbf{E}_{\xi_{k-1}}[\phi_k(x)], \tag{26}$$

where $\mathbf{E}_{\xi_{-1}}[f(x^0)] \stackrel{\text{def}}{=} f(x^0)$. Then we have

$$\mathbf{E}_{\xi_{k-1}}[f(x^k)] - f^\star \;\leq\; \lambda_k \left(\phi_0(x^\star) - f^\star\right) \;\to\; 0.$$

Proof. Since $\{(\phi_k(x), \lambda_k)\}_{k=0}^{\infty}$ is a randomized estimate sequence of $f(x)$, it follows from (25) and (26) that

$$\begin{aligned}
\mathbf{E}_{\xi_{k-1}}[f(x^k)] &\leq \min_x \mathbf{E}_{\xi_{k-1}}[\phi_k(x)] \\
&\leq \min_x \{(1-\lambda_k) f(x) + \lambda_k \phi_0(x)\} \\
&\leq (1-\lambda_k) f(x^\star) + \lambda_k \phi_0(x^\star) \\
&= f^\star + \lambda_k(\phi_0(x^\star) - f^\star),
\end{aligned}$$

which together with (24) implies that the conclusion holds. □

As we will see next, our construction of the randomized estimate sequence satisfies a stronger condition, i.e.,

$$\mathbf{E}_{\xi_{k-1}} f(x^k) \;\leq\; \mathbf{E}_{\xi_{k-1}}[\min_x \phi_k(x)].$$

This implies that the assumption in Lemma 4, namely, (26) holds due to

$$\mathbf{E}_{\xi_{k-1}}[\min_x \phi_k(x)] \;\leq\; \min_x \mathbf{E}_{\xi_{k-1}}[\phi_k(x)].$$

Lemma 5. *Assume that f satisfies Assumption 1 with convexity parameter $\mu \geq 0$. In addition, suppose that*

- *$\phi_0(x)$ is an arbitrary deterministic function on $\Re^N$;*
- *$\{y^k\}_{k=1}^{\infty}$ is a sequence in $\Re^N$ such that y^k depends on ξ_{k-1};*
- *$\{\alpha_k\}_{k=1}^{\infty}$ is independent of ξ_k and satisfies $\alpha_k \in (0, n)$ for all $k \geq 0$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$.*

Then the pair of sequences $\{\phi_k(x)\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$ constructed by setting $\lambda_0 = 1$ and

$$\lambda_{k+1} = \left(1 - \frac{\alpha_k}{n}\right) \lambda_k, \tag{27}$$

$$\phi_{k+1}(x) = \left(1 - \frac{\alpha_k}{n}\right) \phi_k(x) + \alpha_k \left(\frac{1}{n} f(y^k) + \langle \nabla_{i_k} f(y^k), x_{i_k} - y^k_{i_k} \rangle + \frac{\mu}{2n} \|x - y^k\|_L^2\right) \tag{28}$$

is a randomized estimate sequence of $f(x)$.

Proof. It follows from (27) and $\lambda_0 = 1$ that $\lambda_k = \prod_{i=0}^{k-1}(1 - \alpha_i/n)$ for $k \geq 1$. Then we have

$$\log \lambda_k = \sum_{i=0}^{k-1} \log\left(1 - \frac{\alpha_i}{n}\right) \leq -\frac{1}{n} \sum_{i=0}^{k-1} \alpha_i \to -\infty$$

due to $\sum_{i=0}^{\infty} \alpha_i = \infty$. Hence, $\lambda_k \to 0$. We next prove by induction that (25) holds for all $k \geq 0$. Indeed, for $k = 0$, we know that $\lambda_0 = 1$ and hence

$$\mathbf{E}_{\xi_{-1}}[\phi_0(x)] = \phi_0(x) = (1 - \lambda_0) f(x) + \lambda_0 \phi_0(x),$$

that is, (25) holds for $k = 0$. Now suppose it holds for some $k \geq 0$. Using (28), we obtain that

$$\begin{aligned}
\mathbf{E}_{\xi_k}[\phi_{k+1}(x)] &= \mathbf{E}_{\xi_{k-1}}\left[\mathbf{E}_{i_k}[\phi_{k+1}(x)]\right] \\
&= \mathbf{E}_{\xi_{k-1}}\Bigg[\left(1 - \frac{\alpha_k}{n}\right)\phi_k(x) + \alpha_k\Bigg(\frac{1}{n} f(y^k) + \mathbf{E}_{i_k}\left[\langle \nabla_{i_k} f(y^k), x_{i_k} - y^k_{i_k}\rangle\right] \\
&\qquad\qquad + \frac{\mu}{2n}\|x - y^k\|_L^2\Bigg)\Bigg] \\
&= \mathbf{E}_{\xi_{k-1}}\left[\left(1 - \frac{\alpha_k}{n}\right)\phi_k(x) + \frac{\alpha_k}{n}\left(f(y^k) + \langle \nabla f(y^k), x - y^k\rangle + \frac{\mu}{2}\|x - y^k\|_L^2\right)\right] \\
&\leq \mathbf{E}_{\xi_{k-1}}\left[\left(1 - \frac{\alpha_k}{n}\right)\phi_k(x) + \frac{\alpha_k}{n} f(x)\right],
\end{aligned}$$

where the last inequality is due to convexity of f. Using the induction hypothesis, we have

$$\begin{aligned}
\mathbf{E}_{\xi_k}[\phi_{k+1}(x)] &\leq \left(1 - \frac{\alpha_k}{n}\right)\left((1 - \lambda_k) f(x) + \lambda_k \phi_0(x)\right) + \frac{\alpha_k}{n} f(x) \\
&= \left(1 - \left(1 - \frac{\alpha_k}{n}\right)\lambda_k\right) f(x) + \left(1 - \frac{\alpha_k}{n}\right)\lambda_k \phi_0(x) \\
&= (1 - \lambda_{k+1}) f(x) + \lambda_{k+1} \phi_0(x)
\end{aligned}$$

and hence (25) also holds for $k + 1$. This completes the proof. □

Lemma 6. *Let $\phi_0(x) = \phi_0^\star + \frac{\gamma_0}{2}\|x - v^0\|_L^2$. Then the randomized estimate sequence constructed in Lemma 5 preserves the canonical form of the functions, i.e., for all $k \geq 0$,*

$$\phi_k(x) = \phi_k^\star + \frac{\gamma_k}{2}\|x - v^k\|_L^2, \tag{29}$$

where the sequences $\{\gamma_k\}$, $\{v^k\}$ and $\{\phi_k^\star\}$ are defined as follows:

$$\begin{aligned}
\gamma_{k+1} &= \left(1 - \frac{\alpha_k}{n}\right)\gamma_k + \frac{\alpha_k}{n}\mu, && (30)\\
v^{k+1} &= \frac{1}{\gamma_{k+1}}\left(\left(1 - \frac{\alpha_k}{n}\right)\gamma_k v^k + \frac{\alpha_k}{n}\mu y^k - \frac{\alpha_k}{L_{i_k}}U_{i_k}\nabla_{i_k} f(y^k)\right) && (31)\\
\phi_{k+1}^\star &= \left(1 - \frac{\alpha_k}{n}\right)\phi_k^\star + \frac{\alpha_k}{n}f(y^k) - \frac{\alpha_k^2}{2\gamma_{k+1}L_{i_k}}\|\nabla_{i_k} f(y^k)\|^2 \\
&\quad + \frac{\alpha_k\left(1 - \frac{\alpha_k}{n}\right)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2n}\|y^k - v^k\|_L^2 + \langle \nabla_{i_k} f(y^k), v_{i_k}^k - y_{i_k}^k\rangle\right) && (32)
\end{aligned}$$

Proof. First we observe that $\phi_k(x)$ is a convex quadratic function due to (28) and the definition of $\phi_0(x)$. We now prove by induction that for ϕ_k is given by (29) all $k \geq 0$. Clearly, (29) holds for $k = 0$. Suppose now that it holds for some $k \geq 0$. It follows that the Hessian of $\phi_k(x)$ is a block-diagonal matrix given by

$$\nabla^2\phi_k(x) \;=\; \gamma_k \operatorname{diag}\left(L_1 I_{N_1}, \ldots, L_n I_{N_n}\right).$$

Using this relation, (28) and (30), we have

$$\begin{aligned}
\nabla^2\phi_{k+1}(x) &= \left(1 - \frac{\alpha_k}{n}\right)\nabla^2\phi_k(x) + \frac{\alpha_k}{n}\mu \operatorname{diag}\left(L_1 I_{N_1}, \ldots, L_n I_{N_n}\right) \\
&= \gamma_{k+1} \operatorname{diag}\left(L_1 I_{N_1}, \ldots, L_n I_{N_n}\right). && (33)
\end{aligned}$$

Using the induction hypothesis by substituting (29) into (28), we can write $\phi_{k+1}(x)$ as

$$\begin{aligned}
\phi_{k+1}(x) &= \left(1 - \frac{\alpha_k}{n}\right)\left(\phi_k^\star + \frac{\gamma_k}{2}\|x - v^k\|_L^2\right) \\
&\quad + \alpha_k\left(\frac{1}{n}f(y^k) + \langle \nabla_{i_k} f(y^k), x_{i_k} - y_{i_k}^k\rangle + \frac{\mu}{2n}\|x - y^k\|_L^2\right), && (34)
\end{aligned}$$

which together with (31) implies

$$\nabla\phi_{k+1}(v^{k+1}) = \left(1 - \frac{\alpha_k}{n}\right)\gamma_k\sum_{i=1}^n U_i L_i(v_i^{k+1} - v_i^k) + \alpha_k U_{i_k}\nabla_{i_k} f(y^k) + \frac{\alpha_k}{n}\mu\sum_{i=1}^n U_i L_i(v_i^{k+1} - y_i^k) = 0. \tag{35}$$

Letting $x = y^k$ in (34), one has

$$\phi_{k+1}(y^k) = \left(1 - \frac{\alpha_k}{n}\right)\left(\phi_k^\star + \frac{\gamma_k}{2}\|y^k - v^k\|_L^2\right) + \frac{\alpha_k}{n}f(y^k).$$

In view of (31), we have

$$v^{k+1} - y^k \;=\; \frac{1}{\gamma_{k+1}}\left(\left(1-\frac{\alpha_k}{n}\right)\gamma_k(v^k-y^k) - \frac{\alpha_k}{L_{i_k}}U_{i_k}\nabla_{i_k}f(y^k)\right),$$

and hence

$$\begin{aligned}\frac{\gamma_{k+1}}{2}\|y^k - v^{k+1}\|_L^2 \;=\; \frac{1}{2\gamma_{k+1}}\Bigg(&\left(1-\frac{\alpha_k}{n}\right)^2\gamma_k^2\|y^k-v^k\|_L^2 + \frac{\alpha_k^2}{L_{i_k}}\|\nabla_{i_k}f(y^k)\|^2 \\ &-2\alpha_k\left(1-\frac{\alpha_k}{n}\right)\gamma_k\left\langle\nabla_{i_k}f(y^k), v_{i_k}^k - y_{i_k}^k\right\rangle\Bigg).\end{aligned}$$

In addition, using (30) we obtain that

$$\left(1-\frac{\alpha_k}{n}\right)\frac{\gamma_k}{2} - \frac{1}{2\gamma_{k+1}}\left(1-\frac{\alpha_k}{n}\right)^2\gamma_k^2 \;=\; \frac{1}{2\gamma_{k+1}}\left(1-\frac{\alpha_k}{n}\right)\gamma_k\frac{\alpha_k}{n}\mu.$$

By virtue of the above relations and (32), it is not hard to conclude that

$$\phi_{k+1}(y^k) = \phi_{k+1}^\star + \frac{\gamma_{k+1}}{2}\|y^k - v^{k+1}\|_L^2,$$

which, together with (33), (35) and the fact that ϕ_{k+1} is quadratic, implies that

$$\phi_{k+1}(x) = \phi_{k+1}^\star + \frac{\gamma_{k+1}}{2}\|x - v^{k+1}\|_L^2.$$

Therefore, the conclusion holds. □

4.2 Proof of Theorem 4

Let $\phi_0(x) = f(v^0) + \gamma_0\|x - v^0\|_L^2/2$, $\{y^k\}$ and $\{\alpha_k\}$ be generated in the ARCD method. In addition, let $\{(\phi_k(x), \lambda_k\}$ be the randomized estimate sequence of $f(x)$ generated as in Lemma 5 by using such $\{y^k\}$ and $\{\alpha_k\}$.

First we prove by induction that for all $k \geq 0$,

$$\mathbf{E}_{\xi_{k-1}}[f(x^k)] \;\leq\; \mathbf{E}_{\xi_{k-1}}\left[\left\{\phi_k^\star = \min_x \phi_k(x)\right\}\right]. \tag{36}$$

For $k = 0$, using $v^0 = x^0$, the definition of $\phi_0(x)$ and $\mathbf{E}_{\xi_{-1}}[f(x^0)] = f(x^0)$, we have

$$\mathbf{E}_{\xi_{-1}}[f(x^0)] \;=\; f(x^0) \;=\; f(v^0) \;=\; \phi_0^\star,$$

and hence (36) holds for $k = 0$. Now suppose it holds for some $k \geq 0$. It follows from (32) that

$$\begin{aligned}\mathbf{E}_{\xi_k}[\phi_{k+1}^\star] \;&=\; \mathbf{E}_{\xi_{k-1}}\left[\mathbf{E}_{i_k}[\phi_{k+1}^\star]\right] \\ &=\; \mathbf{E}_{\xi_{k-1}}\Bigg[\left(1-\frac{\alpha_k}{n}\right)\phi_k^\star + \frac{\alpha_k}{n}f(y^k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\mathbf{E}_{i_k}\left[\frac{1}{L_{i_k}}\|\nabla_{i_k}f(y^k)\|^2\right] \\ &\qquad + \frac{\alpha_k\left(1-\frac{\alpha_k}{n}\right)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2n}\|y^k-v^k\|_L^2 + \mathbf{E}_{i_k}\left[\langle\nabla_{i_k}f(y^k), v_{i_k}^k - y_{i_k}^k\rangle\right]\right)\Bigg].\end{aligned} \tag{37}$$

Let

$$d_i(y^k) \;=\; -\frac{1}{L_i}\nabla_i f(y^k), \qquad i=1,\ldots,n,$$

and $d(y^k)=\sum_{i=1}^n U_i d_i(y^k)$. Then we have

$$\mathbf{E}_{i_k}\left[\frac{1}{L_{i_k}}\|\nabla_{i_k} f(y^k)\|^2\right] \;=\; \frac{1}{n}\|d(y^k)\|_L^2.$$

Moreover,

$$\mathbf{E}_{i_k}\left[\langle \nabla_{i_k} f(y^k), v^k_{i_k}-y^k_{i_k}\rangle\right] \;=\; \frac{1}{n}\langle \nabla f(y^k), v^k-y^k\rangle.$$

Using these two equalities and dropping the term $\|y^k-v^k\|_L^2$ in (37), we arrive at

$$\begin{aligned}\mathbf{E}_{\xi_k}[\phi^\star_{k+1}] \;\geq\; \mathbf{E}_{\xi_{k-1}}\Bigg[&\left(1-\frac{\alpha_k}{n}\right)\phi^\star_k+\frac{\alpha_k}{n}f(y^k)-\frac{\alpha_k^2}{2n\gamma_{k+1}}\|d(y^k)\|_L^2\\ &+\frac{\alpha_k}{n}\left(1-\frac{\alpha_k}{n}\right)\frac{\gamma_k}{\gamma_{k+1}}\langle\nabla f(y^k),v^k-y^k\rangle\Bigg].\end{aligned}$$

By the induction hypothesis and the convexity of f, we obtain that

$$\mathbf{E}_{\xi_{k-1}}[\phi^\star_k] \;\geq\; \mathbf{E}_{\xi_{k-1}}[f(x^k)] \;\geq\; \mathbf{E}_{\xi_{k-1}}[f(y^k)+\langle\nabla f(y^k),x^k-y^k\rangle].$$

Combining the above two inequalities gives

$$\begin{aligned}\mathbf{E}_{\xi_k}[\phi^\star_{k+1}] \;\geq\; \mathbf{E}_{\xi_{k-1}}\Bigg[&f(y^k)-\frac{\alpha_k^2}{2n\gamma_{k+1}}\|d(y^k)\|_L^2\\ &+\left(1-\frac{\alpha_k}{n}\right)\left\langle\nabla f(y^k),\frac{\alpha_k\gamma_k}{n\gamma_{k+1}}(v^k-y^k)+(x^k-y^k)\right\rangle\Bigg].\end{aligned}$$

Recall that

$$y^k \;=\; \frac{1}{\frac{\alpha_k}{n}\gamma_k+\gamma_{k+1}}\left(\frac{\alpha_k}{n}\gamma_k v^k+\gamma_{k+1}x^k\right).$$

This relation together with the above inequality yields

$$\mathbf{E}_{\xi_k}[\phi^\star_{k+1}] \;\geq\; \mathbf{E}_{\xi_{k-1}}\left[f(y^k)-\frac{\alpha_k^2}{2n\gamma_{k+1}}\|d(y^k)\|_L^2\right].$$

Also, we observe that $\alpha_k^2=\gamma_{k+1}$. Substituting it into the above inequality gives

$$\mathbf{E}_{\xi_k}[\phi^\star_{k+1}] \;\geq\; \mathbf{E}_{\xi_{k-1}}\left[f(y^k)-\frac{1}{2n}\|d(y^k)\|_L^2\right].$$

In addition, notice that

$$x^{k+1} \;=\; y^k-\frac{1}{L_{i_k}}U_{i_k}\nabla_{i_k}f(y^k) \;=\; y^k+U_{i_k}d_{i_k}(y^k),$$

which together with Corollary 1 yields

$$\mathbf{E}_{\xi_k}[\phi_{k+1}^\star] \;\geq\; \mathbf{E}_{\xi_{k-1}}\left[\mathbf{E}_{i_k} f(x^{k+1})\right] \;=\; \mathbf{E}_{\xi_k}[f(x^{k+1})].$$

Therefore, (36) holds for all $k+1$. Further, by Lemma 4, we have

$$\mathbf{E}_{\xi_{k-1}}[f(x^k)] - f^\star \;\leq\; \lambda_k \left(f(x^0) - f^\star + \frac{\gamma_0}{2}\|x^0 - x^\star\|_L^2\right).$$

Finally, we estimate the decay of λ_k, using the same arguments in the proof of [6, Lemma 2.2.4]. Here we assume $\gamma_0 \geq \mu$ (it suffices to set $\gamma_0 = 1$ because $\mu \leq 1$). Indeed, if $\gamma_k \geq \mu$, then

$$\gamma_{k+1} \;=\; \left(1 - \frac{\alpha_k}{n}\right)\gamma_k + \frac{\alpha_k}{n}\mu \;\geq\; \mu.$$

So we have $\gamma_k \geq \mu$ for all $k \geq 0$. Since $\alpha_k^2 = \gamma_{k+1}$, we have $\alpha_k \geq \sqrt{\mu}$ for all $k \geq 0$. Therefore,

$$\lambda_k \;=\; \prod_{i=0}^{k-1}(1 - \frac{\alpha_i}{n}) \;\leq\; \left(1 - \frac{\sqrt{\mu}}{n}\right)^k.$$

In addition, we have $\gamma_k \geq \gamma_0\lambda_k$. To see this, we note $\gamma_0 = \gamma_0\lambda_0$ and use induction

$$\gamma_{k+1} \;\geq\; \left(1 - \frac{\alpha_k}{n}\right)\gamma_k \;\geq\; \left(1 - \frac{\alpha_k}{n}\right)\gamma_0\lambda_k \;=\; \gamma_0\lambda_{k+1}.$$

This implies

$$\alpha_k \;=\; \sqrt{\gamma_{k+1}} \;\geq\; \sqrt{\gamma_0\lambda_{k+1}}. \tag{38}$$

Since $\{\lambda_k\}$ is a decreasing sequence, we have

$$\begin{aligned}\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}} \;&=\; \frac{\sqrt{\lambda_k} - \sqrt{\lambda_{k+1}}}{\sqrt{\lambda_k}\sqrt{\lambda_{k+1}}} \;=\; \frac{\lambda_k - \lambda_{k+1}}{\sqrt{\lambda_k}\sqrt{\lambda_{k+1}}(\sqrt{\lambda_k} + \sqrt{\lambda_{k+1}})} \\ &\geq\; \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k\sqrt{\lambda_{k+1}}} \;=\; \frac{\lambda_k - \left(1 - \frac{\alpha_k}{n}\right)\lambda_k}{2\lambda_k\sqrt{\lambda_{k+1}}} \;=\; \frac{\frac{\alpha_k}{n}}{2\sqrt{\lambda_{k+1}}}.\end{aligned}$$

Combining with (38) gives

$$\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}} \;\geq\; \frac{\sqrt{\gamma_0}}{2n}.$$

By further noting $\lambda_0 = 1$, we obtain

$$\frac{1}{\sqrt{\lambda_k}} \;\geq\; 1 + \frac{k}{n}\frac{\sqrt{\gamma_0}}{2}.$$

Therefore

$$\lambda_k \;\leq\; \left(\frac{n}{n + k\frac{\sqrt{\gamma_0}}{2}}\right)^2.$$

This completes the proof for Theorem 4.

References

[1] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. Coordinate descent method for large-scale l_2-loss linear support vector machines. *Journal of Machine Learning Research*, 9:1369–1398, 2008.

[2] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In ICML 2008, pages 408–415, 2008.

[3] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.

[4] Y. Li and S. Osher. Coordinate descent optimization for l_1 minimization with application to compressed sensing; a greedy algorithm. *Inverse Problems and Imaging*, 3:487–503, 2009.

[5] Z. Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 2002.

[6] Y. Nesterov. Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Boston, 2004.

[7] Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Paper 1007/76, Catholic University of Louvain, Belgium, 2007.

[8] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2): 341–362, 2012.

[9] Z. Qin, K. Scheinberg, and D. Goldfarb. Effiient block-coordinate descent algorithms for the group lasso. To appear in *Mathematical Programming Computation*, 2010.

[10] P. Richtárik and M. Takáč. Efficient serial and parallel coordinate descent method for huge-scale truss topology design. *Operations Research Proceedings*, 27–32, 2012.

[11] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. To appear in *Mathematical Programming*, 2011.

[12] P. Richtárik and M. Takáč. Parallel coordinate descent methods for big data optimization. Technical report, November 2012.

[13] A. Saha and A. Tewari. On the non-asymptotic convergence of cyclic coordinate descent methods. *SIAM Journal on Optimization*, 23(1):576–601, 2013.

[14] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l_1 regularized loss minimization. In Proceedings of the 26th International Conference on Machine Learning, 2009.

[15] S. Shalev-Shwartz and T. Zhang. Proximal stochastic dual coordinate ascent. Technical report, 2012.

[16] R. Tappenden, P. Richtárik and J. Gondzio. Inexact coordinate descent: complexity and preconditioning. Technical report, April 2013.

[17] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:475–494, 2001.

[18] P. Tseng and S. Yun. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *Journal of Optimization Theory and Applications*, 140:513–535, 2009.

[19] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117:387–423, 2009.

[20] Z. Wen, D. Goldfarb, and K. Scheinberg. Block coordinate descent methods for semidefinite programming. In Miguel F. Anjos and Jean B. Lasserre, editors, Handbook on Semidefinite, Cone and Polynomial Optimization: Theory, Algorithms, Software and Applications. Springer, Volume 166: 533–564, 2012.

[21] S. J. Wright. Accelerated block-coordinate relaxation for regularized optimization. *SIAM Journal on Optimization*, 22:159-186, 2012.

[22] T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

[23] S. Yun and K.-C. Toh. A coordinate gradient descent method for l_1-regularized convex minimization. *Computational Optimization and Applications*, 48:273–307, 2011.