Randomized Block Coordinate Non-Monotone Gradient Method for a Class of Nonlinear Programming

Zhaosong Lu* Lin Xiao †

June 25, 2013

Abstract

In this paper we propose a randomized block coordinate non-monotone gradient (RBCNMG) method for minimizing the sum of a smooth (possibly nonconvex) function and a block-separable (possibly nonconvex nonsmooth) function. At each iteration, this method randomly picks a block according to any prescribed probability distribution and typically solves several associated proximal subproblems that usually have a closed-form solution, until a certain sufficient reduction on the objective function is achieved. We show that the sequence of expected objective values generated by this method converges to the expected value of the limit of the objective value sequence produced by a random single run. Moreover, the solution sequence generated by this method is arbitrarily close to an approximate stationary point with high probability. When the problem under consideration is convex, we further establish that the sequence of expected objective values generated by this method converges to the optimal value of the problem. In addition, the objective value sequence is arbitrarily close to the optimal value with high probability. We also conduct some preliminary experiments to test the performance of our RBCNMG method on the ℓ_1-regularized least-squares problem. The computational results demonstrate that our method substantially outperform the randomized block coordinate descent method proposed in [16].

Key words: Randomized block coordinate, composite minimization, non-monotone gradient method.

1 Introduction

Nowadays first-order (namely, gradient-type) methods are the prevalent tools for solving large-scale problems arising in science and engineering. As the size of problems becomes huge, it is,

*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).

however, greatly challenging to these methods because gradient evaluation can be prohibitively expensive. Due to this reason, block coordinate descent (BCD) methods and their variants have been studied for solving various large-scale problems (see, for example, [3, 6, 28, 8, 22, 23, 25, 26, 14, 29, 18, 7, 15, 19, 17, 20]). Recently, Nesterov [13] proposed a randomized BCD (RBCD) method, which is promising for solving a class of huge-scale convex optimization problems, provided that their partial gradients can be efficiently updated. The iteration complexity for finding an approximate solution is analyzed in [13]. More recently, Richtárik and Takáč [16] extended Nesterov's RBCD method [13] to solve a more general class of convex optimization problems in the form of

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

where f is convex differentiable in $\Re^N$, and Ψ is a block separable convex function. 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, $\{x_i : i = 1, \ldots, n\}$ 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 a current iterate x^k, the RBCD method [16] picks $i \in \{1, \ldots, n\}$ uniformly, and it solves a block-wise proximal subproblem in the form of:

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

and sets $x_i^{k+1} = x^k + d_i(x^k)$ and $x_j^{k+1} = x_j^k$ for all $j \neq i$, where $\nabla_i f$ is the *partial gradient* of f with respect to x_i and $L_i > 0$ is the Lipschitz constant of $\nabla_i f$ with respect to the norm $\|\cdot\|$. The iteration complexity for finding an approximate solution with high probability is initially established in [16] and it has recently been improved by Lu and Xiao [9]. One can observe that for $n = 1$, the RBCD method becomes a classical proximal (full) gradient method with a "constant $1/L$" stepsize. It has been noticed that in practice this method converges much slower than the same type of methods but with a variable stepsize (e.g., spectral-type stepsize [1, 2, 5, 27, 10]). In addition, for the RBCD method [16], the random block is chosen uniformly at each iteration, which means each block is treated equally. Since the Lipschitz constants of the block gradients $\nabla_i f$ may differ much each other, it looks more plausible to choose the random block non-uniformly. We shall also mention that the above RBCD method is so far developed only for convex optimization problems. A natural question is whether this method can be extended to solve nonconvex optimization problems.

To address the aforementioned questions, we consider a class of (possibly nonconvex) nonlinear programming problems in the form of (1) that satisfy the following assumption:

Assumption 1 *f is differentiable (but possibly nonconvex) in $\Re^N$. Each $\Psi_i : \Re^{N_i} \to \Re \cup \{+\infty\}$ is a closed (but possibly nonconvex nonsmooth) function for $i = 1, \ldots, n$.*

For the above partition of $x \in \Re^N$, 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

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

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 $\|\cdot\|$.

Throughout this paper we assume that the set of optimal solutions of problem (1), denoted by X^*, is nonempty and the optimal value of (1) is denoted by $F^\star$. We also make the following assumption, which is imposed in [13, 16] as well.

Assumption 2 *The function F is bounded below in $\Re^N$ and uniformly continuous in Ω, and moreover, the gradient of function f is coordinate-wise Lipschitz continuous with constants $L_i > 0$ in $\Re^N$, that is,*

$$\|\nabla_i f(x + U_i h_i) - \nabla_i f(x)\| \;\le\; L_i \|h_i\|, \qquad \forall h_i \in \Re^{n_i},\ i = 1, \ldots, n, \quad \forall x \in \Omega,$$

where

$$\Omega \stackrel{\text{def}}{=} \{x \in \Re^N : F(x) \le F(x^0)\}$$

for some $x^0 \in \Re^n$.

In this paper we propose a randomized block coordinate non-monotone gradient (RBCNMG) method for solving problem (1) that satisfies Assumptions (1) and (2). At each iteration, this method randomly picks a block according to any prescribed probability distribution and typically solves several associated proximal subproblems in the form of (2) with $L\|d\|^2$ replaced by $d^T H d$ for some positive definite matrix H (see (4) for details) until a certain sufficient reduction on the objective function is achieved. Here, H is an approximation of the restricted Hessian of f with respect to the selected block, which can be, for example, estimated by the spectral method (e.g., see [1, 2, 5, 27, 10]). The resulting method is generally non-monotone. We show that the sequence of expected objective values generated by this method converges to the expected value of the limit of the objective value sequence produced by a random single run. Moreover, the solution sequence generated by this method is arbitrarily close to an approximate stationary point with high probability. When the problem under consideration is convex, we further establish that the sequence of expected objective values generated by this method converges to the optimal value of the problem. In addition, the objective value sequence is arbitrarily close to the optimal value with high probability. We also conduct some preliminary experiments to test the performance of our RBCNMG method on the ℓ_1-regularized least-squares problem. The computational results demonstrate that our method substantially outperform the randomized block coordinate descent (RBCD) method proposed in [16].

This paper is organized as follows. In Section 2 we propose a RBCNMG method for solving problem (1) and analyze its convergence. In Section 3 we conduct numerical experiments to compare RBCNMG method with the RBCD method [16] for solving ℓ_1-regularized least-squares problem.

Before ending this section we introduce some notations that are used throughout this paper and also state a known fact. The space of symmetric matrices is denoted by $\mathcal{S}^n$. If $X \in \mathcal{S}^n$ is positive semidefinite (resp. definite), we write $X \succeq 0$ (resp. $X \succ 0$). Also, we write $X \succeq Y$ or $Y \preceq X$ to mean $X - Y \succeq 0$. The cone of positive symmetric definite matrices in $\mathcal{S}^n$ is denoted by $\mathcal{S}^n_{++}$. Finally, it immediately follows from Assumption 2 that

$$f(x + U_i h_i) \leq f(x) + \nabla_i f(x)^T h_i + \frac{L_i}{2}\|h_i\|^2, \quad \forall h_i \in \Re^{n_i}, \ i = 1, \ldots, n, \quad \forall x \in \Omega. \tag{3}$$

2 Randomized block coordinate non-monotone gradient method

Given a point $x \in \Re^N$ and a matrix $H_i \in \mathcal{S}^{n_i}_{++}$, we denote by $d_{H_i}(x)$ a solution of the following problem:

$$d_{H_i}(x) := \arg\min_d \left\{ \nabla_i f(x)^T d + \frac{1}{2} d^T H_i d + \Psi_i(x + d) \right\}, \quad i = 1, \ldots, n. \tag{4}$$

Randomized block coordinate non-monotone gradient (RBCNMG) method:

Let x^0 be given as in Assumption 2. Choose parameters $\eta > 1$, $\sigma > 0$, $0 < \underline{\lambda} \leq \bar{\lambda}$, $0 < \underline{\theta} < \bar{\theta}$, integer $M \geq 0$, and $0 < p_i < 1$ for $i = 1, \ldots, n$ such that $\sum_{i=1}^n p_i = 1$. Set $k = 0$.

1) Set $d^k = 0$. Pick $i_k = i \in \{1, \ldots, n\}$ with probability p_i.

1) Choose $\theta_k^0 \in [\underline{\theta}, \bar{\theta}]$ and $H_{i_k} \in \mathcal{S}^{n_{i_k}}$ such that $\underline{\lambda} I \preceq H_{i_k} \preceq \bar{\lambda} I$.

2) For $j = 0, 1, \ldots$

   2a) Let $\theta_k = \theta_k^0 \eta^j$. Solve (4) with $x = x^k$, $i = i_k$, $H_i = \theta_k H_{i_k}$ to obtain $d_{i_k}^k = d_{H_i}(x)$.

   2b) If d^k satisfies

$$F(x^k + d^k) \ \leq \ \max_{[k-M]^+ \leq i \leq k} F(x^i) - \frac{\sigma}{2}\|d^k\|^2, \tag{5}$$

   go to step 3).

3) Set $x^{k+1} = x^k + d^k$ and $k \leftarrow k + 1$.

end

We first show that the inner iterations of the above RBCNMG method must terminate within a finite number of loops.

Lemma 2.1 *Let $\{\theta_k\}$ be the sequence generated by the above RBCNMG method. Then, there holds*

$$\underline{\theta} \ \leq \ \theta_k \ \leq \ \eta(\max_i L_i + \sigma)/\underline{\lambda}. \tag{6}$$

Proof. It is clear that $\theta_k \geq \underline{\theta}$. We now show that $\theta_k \ \leq 2\eta \max_i L_i/(1-\sigma)$. For any $\theta \geq 0$, let $d \in \Re^N$ be such that $d_{[i_k]} = d_{\theta H_{i_k}}(x^k)$ and $d_i = 0$ for $i \neq i_k$. Using (3), (4), and the relation $H_{i_k} \succeq \underline{\lambda} I$, we have

$$\begin{aligned} F(x^k+d) &= f(x^k+d) + \Psi(x^k+d) \ \leq \ f(x^k) + \nabla_{i_k} f(x^k)^T d_{i_k} + \tfrac{L_{i_k}}{2}\|d_{i_k}\|^2 + \Psi(x^k+d) \\ &= F(x^k) + \underbrace{\nabla_{i_k} f(x^k)^T d + \frac{\theta}{2} d^T H_{i_k} d + \Psi_{i_k}(x^k_{i_k} + d_{i_k}) - \Psi_{i_k}(x^k_{i_k})}_{\leq 0} + \tfrac{L_{i_k}}{2}\|d\|^2 - \tfrac{\theta}{2} d^T H_{i_k} d \\ &\leq F(x^k) + \tfrac{L_{i_k} - \theta\underline{\lambda}}{2}\|d\|^2. \end{aligned}$$

It then follows that whenever $\theta \geq (L_{i_k} + \sigma)/\underline{\lambda}$,

$$F(x^k+d) \ \leq \ F(x^k) - \frac{\sigma}{2}\|d\|^2 \ \leq \ \max_{[k-M]^+ \leq i \leq k} F(x^i) - \frac{\sigma}{2}\|d\|^2. \tag{7}$$

Suppose for contradiction that $\theta_k > \eta(\max_i L_i + \sigma)/\underline{\lambda}$. Then, $\theta := \theta_k/\eta \geq (L_{i_k} + \sigma)/\underline{\lambda}$ and hence (7) holds for such θ. It means that (5) still holds when replacing θ_k by θ_k/η, which contradicts with our specific choice of θ_k. ∎

After k iterations, the RBCNMG method generates a random output $(x^k, F(x^k))$, which depends on the observed realization of random vector

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

We define

$$\ell(k) = \arg\max_i \{F(x^i) : i = [k-M]^+, \ldots, k\}, \quad \forall k \geq 0. \tag{8}$$

We next show that the expected value $\mathbf{E}_{\xi_k}[\|d^k\|] \to 0$ and $\mathbf{E}_{\xi_{k-1}}[F(x^k)]$ converges as $k \to \infty$. Before proceeding, we establish a technical lemma that will be used subsequently.

Lemma 2.2 *Let $\{y^k\}$ and $\{z^k\}$ be two random sequences in Ω generated from the random vectors $\{\xi_{k-1}\}$. Suppose that*

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|y^k - z^k\|] = 0.$$

Then there holds:

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[|F(y^k) - F(z^k)|] = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(y^k) - F(z^k)] = 0.$$

Proof. From Assumption 2, we know that F is uniformly continuous in Ω. For any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that $|F(x) - F(y)| \le \epsilon/2$, provided that $x, y \in \Omega$ satisfying $\|x-y\| \le \delta_\epsilon$. Using this relation and the fact that $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\Delta^k\|] = 0$, where $\Delta^k := y^k - z^k$, we obtain that for sufficiently large k,

$$
\begin{aligned}
\mathbf{E}_{\xi_{k-1}}[|F(y^k) - F(z^k)|] \;&=\; \mathbf{E}_{\xi_{k-1}}\left[|F(y^k) - F(z^k)| \;\middle|\; \|\Delta^k\| > \delta_\epsilon\right] \mathbf{P}(\|\Delta^k\| > \delta_\epsilon) \\
&\quad + \mathbf{E}_{\xi_{k-1}}\left[|F(y^k) - F(z^k)| | \|\Delta^k\| \le \delta_\epsilon\right] \mathbf{P}(\|\Delta^k\| \le \delta_\epsilon) \\
&\le\; [F(x^0) - \min_{x\in\Omega} F(x)] \tfrac{\mathbf{E}_{\xi_{k-1}}[\|\Delta^k\|]}{\delta_\epsilon} + \tfrac{\epsilon}{2} \;\le\; \epsilon.
\end{aligned} \tag{9}
$$

Due to the arbitrarity of ϵ, we see that the first statement of this lemma holds. The first statement together with the following well-known inequality

$$
|\mathbf{E}_{\xi_{k-1}}[F(y^k) - F(z^k)]| \;\le\; \mathbf{E}_{\xi_{k-1}}\left[|F(y^k) - F(z^k)|\right]
$$

then implies that the second statement also holds. ■

We next show that the sequence of expected objective values generated by the RBCNMG method converges to the expected value of the limit of the objective value sequence produced by a random single run.

Theorem 2.3 *Let $\{x^k\}$ and $\{d^k\}$ be the sequences generated by the above RBCNMG method. Then,* $\lim_{k\to\infty} \mathbf{E}_{\xi_k}[\|d^k\|] = 0$ *and*

$$
\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \;=\; \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}], \tag{10}
$$

where $\xi_\infty := \{i_1, i_2, \cdots\}$ *and*

$$
F^*_{\xi_\infty} \;=\; \lim_{k\to\infty} F(x^k).
$$

Proof. By the definition of x^{k+1} and (5), we see that $F(x^{k+1}) \le F(x^{\ell(k)})$, which together with (8) implies that $F(x^{\ell(k+1)}) \le F(x^{\ell(k)})$. It then follows that

$$
\mathbf{E}_{\xi_k}[F(x^{\ell(k+1)})] \;\le\; \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})], \quad \forall k \ge 1.
$$

Hence, $\{\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})]\}$ is non-increasing. Since F is bounded below, so is $\{\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})]\}$. Therefore, there exists $F^* \in \Re$ such that

$$
\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] = F^*. \tag{11}
$$

We first prove by induction that the following limits hold for all $j \ge 1$:

$$
\begin{aligned}
&\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-j}\|] = 0, \\
&\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)-j})] = F^*.
\end{aligned} \tag{12}
$$

Indeed, using (5) and (8), we have

$$F(x^{k+1}) \;\leq\; F(x^{\ell(k)}) - \frac{\sigma}{2}\|d^k\|^2, \quad \forall k \geq 0. \tag{13}$$

Replacing k by $\ell(k)-1$ in (13), we obtain that

$$F(x^{\ell(k)}) \;\leq\; F(x^{\ell(\ell(k)-1)}) - \frac{\sigma}{2}\|d^{\ell(k)-1}\|^2, \quad \forall k \geq M+1.$$

In view of this relation, $\ell(k) \geq k-M$, and monotonicity of $\{F(x^{\ell(k)})\}$, one can have

$$F(x^{\ell(k)}) \;\leq\; F(x^{\ell(k-M-1)}) - \frac{\sigma}{2}\|d^{\ell(k)-1}\|^2, \quad \forall k \geq M+1.$$

Then we have

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \;\leq\; \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k-M-1)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|^2], \quad \forall k \geq M+1.$$

Notice that

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k-M-1)})] \;=\; \mathbf{E}_{\xi_{k-M-2}}[F(x^{\ell(k-M-1)})], \quad \forall k \geq M+2.$$

It follows from the above two relations that

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \;\leq\; \mathbf{E}_{\xi_{k-M-2}}[F(x^{\ell(k-M-1)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|^2], \quad \forall k \geq M+2.$$

This together with (11) implies that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|^2] = 0.$$

We thus have

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|] = 0. \tag{14}$$

One can also observe that $F(x^k) \leq F(x^0)$ and hence $\{x^k\} \subset \Omega$. Using this fact, (11), (14) and Lemma 2.2, we obtain that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)-1})] \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \;=\; F^*.$$

Therefore, (12) holds for $j=1$. Suppose now that it holds for some j. We need to show that it also holds for $j+1$. From (13) with k replaced by $\ell(k)-j-1$, we have

$$F(x^{\ell(k)-j}) \;\leq\; F(x^{\ell(\ell(k)-j-1)}) - \frac{\sigma}{2}\|d^{\ell(k)-j-1}\|^2, \quad \forall k \geq M+j+1.$$

By this relation, $\ell(k) \geq k-M$, and monotonicity of $\{F(x^{\ell(k)})\}$, one can have

$$F(x^{\ell(k)-j}) \;\leq\; F(x^{\ell(k-M-j-1)}) - \frac{\sigma}{2}\|d^{\ell(k)-j-1}\|^2, \quad \forall k \geq M+j+1.$$

Then we obtain that

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)-j})] \ \le \ \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k-M-j-1)})] - \frac{\sigma}{2}\|d^{\ell(k)-j-1}\|^2, \quad \forall k \ge M+j+1.$$

Notice that

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k-M-j-1)})] \ = \ \mathbf{E}_{\xi_{k-M-j-2}}[F(x^{\ell(k-M-j-1)})], \quad \forall k \ge M+j+2.$$

It follows from the above two relations that

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)-j})] \ \le \ \mathbf{E}_{\xi_{k-M-j-2}}[F(x^{\ell(k-M-j-1)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-j-1}\|^2], \quad \forall k \ge M+j+2.$$

Using this relation, the induction hypothesis, (11), and a similar argument as above, we can obtain that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-j-1}\|] = 0.$$

This relation together with Lemma 2.2 and the induction hypothesis yields

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)-j-1})] \ = \ \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)-j})] \ = \ F^*.$$

Therefore, (12) holds for $j+1$ and the proof of (12) is completed.

For all $k \ge M+1$, we define

$$\tilde{d}^{\ell(k)-j} = \begin{cases} d^{\ell(k)-j} & \text{if } j \le \ell(k)-(k-M-1), \\ 0 & \text{otherwise}, \end{cases} \qquad j = 1,\ldots,M+1.$$

It is not hard to observe that

$$\|\tilde{d}^{\ell(k)-j}\| \ \le \ \|d^{\ell(k)-j}\|,$$

$$x^{\ell(k)} = x^{k-M-1} + \sum_{j=1}^{M+1} \tilde{d}^{\ell(k)-j}.$$

It follows from (12) that $\lim\limits_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\tilde{d}^{\ell(k)-j}\|] = 0$ for $j = 1,\ldots,M+1$, and hence

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}\left[\left\|\sum_{j=1}^{M+1} \tilde{d}^{\ell(k)-j}\right\|\right] = 0.$$

This together with (12) and Lemma 2.2 implies that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{k-M-1})] \ = \ \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \ = \ F^*.$$

Notice that $\mathbf{E}_{\xi_{k-M-2}}[F(x^{k-M-1})] = \mathbf{E}_{\xi_{k-1}}[F(x^{k-M-1})]$. Combining these relations, we have

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-M-2}}[F(x^{k-M-1})] \ = \ F^*,$$

which is equivalent to

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] \;=\; F^*.$$

In addition, it follows from (13) that

$$\mathbf{E}_{\xi_k}[F(x^{k+1})] \;\leq\; \mathbf{E}_{\xi_k}[F(x^{\ell(k)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_k}[\|d^k\|^2], \quad \forall k \geq 0. \tag{15}$$

Notice that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_k}[F(x^{\ell(k)})] \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \;\;=\; F^* \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_k}[F(x^{k+1})]. \tag{16}$$

Using (15) and (16), we conclude that $\lim_{k\to\infty} \mathbf{E}_{\xi_k}[\|d^k\|] = 0$.

Finally, we claim that $\mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}] = F^*$. Indeed, by a similar argument as above, one can show that the limit of $\{F(x^k)\}$ exists and is finite. Hence, $F^*_{\xi_\infty}$ is well-defined for all ξ_∞ and moreover the random sequence $\{F(x^k)\}$ converges to F^*_∞ with probability one as $k \to \infty$. It then implies that $\{F(x^k)\}$ converges to F^*_∞ in probability, that is, for any $\epsilon > 0$,

$$\lim_{k\to\infty} \mathbf{P}(\xi_\infty : |F(x^k) - F^*_\infty| \geq \epsilon) \;=\; 0.$$

Notice that F is bounded below in $\Re^N$ and $F(x^k) \leq F(x^0)$. Hence, there exists some $c > 0$ such that $|F(x^k)| \leq c$ and $|F^*_\infty| \leq c$ for all ξ_∞. It then follows that for any $\epsilon > 0$ and $k \geq 0$,

$$\begin{aligned} \mathbf{E}_{\xi_\infty}[|F(x^k) - F^*_\infty|] \;&=\; \mathbf{E}_{\xi_\infty}\left[|F(x^k) - F^*_\infty| \;\middle|\; |F(x^k) - F^*_\infty| \geq \epsilon\right] \mathbf{P}(\xi_\infty : |F(x^k) - F^*_\infty| \geq \epsilon) \\ &\quad + \mathbf{E}_{\xi_\infty}\left[|F(x^k) - F^*_\infty| \;\middle|\; |F(x^k) - F^*_\infty| < \epsilon\right] \mathbf{P}(\xi_\infty : |F(x^k) - F^*_\infty| < \epsilon) \\ &\leq\; 2\mathbf{P}(\xi_\infty : |F(x^k) - F^*_\infty| \geq \epsilon)c + \epsilon. \end{aligned}$$

Letting $k \to \infty$ and $\epsilon \to 0$ on both sides of the above inequality, we have

$$\lim_{k\to\infty} \mathbf{E}_{\xi_\infty}[|F(x^k) - F^*_\infty|] \;=\; 0,$$

which implies that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_\infty}[F(x^k) - F^*_\infty] \;=\; 0,$$

As shown above, $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)]$ exists and so does $\lim_{k\to\infty} \mathbf{E}_{\xi_\infty}[F(x^k)]$ due to

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_\infty}[F(x^k)].$$

Combining the above two relations, we conclude that $\mathbf{E}_{\xi_\infty}[F^*_\infty]$ is well-defined and moreover,

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] \;=\; \mathbf{E}_{\xi_\infty}[F^*_\infty].$$

This completes our proof. ∎

We next show that when k is sufficiently large, x^k is arbitrarily close to an approximate stationary point of (1) with high probability.

Theorem 2.4 *Let $\bar{d}^{k,i}$ denote the vector d^k obtained in Step (2) of the RBCNMG method when $i_k = i$ is chosen for $i = 1, \ldots, n$, and $\bar{d}^k = \sum_{i=1}^n \bar{d}^{k,i}$. Then there hold:*

(i)

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|] = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\mathrm{dist}(-\nabla f(x^k + \bar{d}^k), \partial\Psi(x^k + \bar{d}^k))] \; = \; 0. \quad (17)$$

(ii) For any $\epsilon > 0$ and $\rho \in (0,1)$, there exists K such that for all $k \geq K$,

$$\mathbf{P}\left(\max\left\{\|x^k - \bar{x}^k\|, |F(x^k) - F(\bar{x}^k)|, \mathrm{dist}(-\nabla f(\bar{x}^k), \partial\Psi(\bar{x}^k))\right\} \leq \epsilon \text{ for some } \bar{x}^k\right) \; \geq \; 1-\rho.$$

Proof. (i) We observe from (5) that

$$F(x^k + \bar{d}^{k,i}) \; \leq \; F(x^{\ell(k)}) - \frac{\sigma}{2}\|\bar{d}^{k,i}\|^2.$$

It follows that

$$\begin{aligned} \mathbf{E}_{i_k}[F(x^k + \bar{d}^{k,i})] \; &= \; \textstyle\sum_{i=1}^n p_i F(x^k + \bar{d}^{k,i}) \; \leq \; F(x^{\ell(k)}) - \frac{\sigma}{2}\sum_{i=1}^n p_i\|\bar{d}^{k,i}\|^2, \\ &\leq \; F(x^{\ell(k)}) - \tfrac{1}{2}\sigma(\min_i p_i)\textstyle\sum_{i=1}^n \|\bar{d}^{k,i}\|^2. \end{aligned}$$

Hence, we obtain that

$$\mathbf{E}_{\xi_k}[F(x^{k+1})] \; \leq \; \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] - \frac{1}{2}\sigma(\min_i p_i)\sum_{i=1}^n \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^{k,i}\|^2],$$

which together with (10) implies that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^{k,i}\|] = 0, \quad i = 1, \ldots, n.$$

It immediately follows from the above equalities that the first relation of (17) holds.

Let $\bar{\theta}_{k,i}$ denote the value of θ_k obtained in Step (2) of the RBCNMG method when $i_k = i$ is chosen for $i = 1, \ldots, n$. From Lemma 2.1, we see that $\bar{\theta}_{k,i}$ satisfies (6) for all k and i. By the definition of $\bar{d}^{k,i}$ and (4), we observe that

$$\bar{d}^k = \arg\min_d \left\{\nabla f(x^k)^T d + \frac{1}{2}\sum_{i=1}^n d_i^T(\bar{\theta}_{k,i} H_i)d_i + \Psi(x^k + d)\right\}$$

for some $\underline{\lambda} I \preceq H_i \preceq \bar{\lambda} I$, $i = 1, \ldots, n$. By the first-order optimality condition (see, for example, Proposition 2.3.2 of [4]), one can have

$$0 \in \nabla f(x^k) + H\bar{d}^k + \partial\Psi(x^k + \bar{d}^k), \quad (18)$$

where H is a block diagonal matrix consisting of $\bar{\theta}_{k,1}H_1$, ..., $\bar{\theta}_{k,n}H_n$ as its diagonal blocks. It follows from (6) and the fact $\underline{\lambda}I \preceq H_i \preceq \bar{\lambda}I$ for all i that

$$\underline{\lambda}I \ \preceq \ H \ \preceq \ (\bar{\lambda}\eta(\max_i L_i + \sigma)/\underline{\lambda})I,$$

and hence,

$$\|H\bar{d}^k\| \leq (\bar{\lambda}\eta(\max_i L_i + \sigma)/\underline{\lambda})\|\bar{d}^k\|. \tag{19}$$

By Lemma 2 of Nesterov [13] and Assumption 2, we see that

$$\|\nabla f(x^k + \bar{d}^k) - \nabla f(x^k)\| \ \leq \ (\sum_{i=1}^n L_i)\|\bar{d}^k\|.$$

Using this relation along with (18) and (19), we obtain that

$$\mathrm{dist}(-\nabla f(x^k + \bar{d}^k), \partial\Psi(x^k + \bar{d}^k)) \ \leq \ \left(\bar{\lambda}\eta(\max_i L_i + \sigma)/\underline{\lambda} + \sum_{i=1}^n L_i\right)\|\bar{d}^k\|,$$

which together with the first relation of (17) implies that the second statement of (17) also holds.

(ii) Let $\bar{x}^k = x^k + \bar{d}^k$, where $\bar{d}^k$ is defined above. By statement (i), we know that $\mathbf{E}_{\xi_{k-1}}[x^k - \bar{x}^k] \to 0$, which together with Lemma 2.2 implies that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[|F(x^k) - F(\bar{x}^k)|] \ = \ 0.$$

Using this relation and statement (i), we then have

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}\left[\max\left\{\|x^k - \bar{x}^k\|, |F(x^k) - F(\bar{x}^k)|, \mathrm{dist}(-\nabla f(\bar{x}^k), \partial\Psi(\bar{x}^k))\right\}\right] \ = \ 0.$$

Statement (ii) immediately follows from this relation and the Markov inequality. ∎

We finally show that when f and Ψ are convex functions, $F(x^k)$ can be arbitrarily close to the optimal value of (1) with high probability for sufficiently large k.

Theorem 2.5 *Let $\{x^k\}$ be generated by the RBCNMG method. Suppose that f and Ψ are convex functions. Assume that there exists a subsequence K such that $\{\mathrm{dist}(x^k, X^*)\}_K$ is bounded almost surely (a.s.), where X^* is the set of optimal solutions of (1). Then there hold:*

(i)

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \min_{x\in\Re^N} F(x).$$

(ii) For any $\epsilon > 0$ and $\rho \in (0,1)$, there exists K such that for all $k \geq K$,

$$\mathbf{P}\left(F(x^k) - \min_{x\in\Re^N} F(x) \ \leq \ \epsilon\right) \ \geq \ 1 - \rho.$$

Proof. (i) Let $\bar{d}^k$ be defined in the premise of Theorem 2.4. It follows from Theorem 2.4 that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|] = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|s^k\|] \;=\; 0 \tag{20}$$

for some $s^k \in \partial F(x^k + \bar{d}^k)$. Using Lemma 2.2 and Theorem 2.3, we have

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k + \bar{d}^k)] = \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = F^* \tag{21}$$

for some $F^* \in \Re$. Let x^k_* be the projection of x^k onto X^*. By the convexity of F, we then we have

$$F(x^k + \bar{d}^k) \;\leq\; F(x^k_*) + (s^k)^T(x^k + \bar{d}^k - x^k_*). \tag{22}$$

One can observe that

$$\begin{aligned}
|\mathbf{E}_{\xi_{k-1}}[(s^k)^T(x^k + \bar{d}^k - x^k_*)]| \;&\leq\; \mathbf{E}_{\xi_{k-1}}[|(s^k)^T(x^k + \bar{d}^k - x^k_*)|] \\
&\leq\; \mathbf{E}_{\xi_{k-1}}[\|s^k\|\|(x^k + \bar{d}^k - x^k_*)\|] \\
&\leq\; \sqrt{\mathbf{E}_{\xi_{k-1}}[\|s^k\|^2]}\sqrt{\mathbf{E}_{\xi_{k-1}}[\|(x^k + \bar{d}^k - x^k_*)\|^2]} \\
&\leq\; \sqrt{\mathbf{E}_{\xi_{k-1}}[\|s^k\|^2]}\sqrt{2\mathbf{E}_{\xi_{k-1}}[(\mathrm{dist}(x^k, X^*))^2 + \|\bar{d}^k\|^2]},
\end{aligned}$$

which together with (20) and the assumption that $\{\mathrm{dist}(x^k, X^*)\}_K$ is bounded a.s. implies that

$$\lim_{k\in K\to\infty} \mathbf{E}_{\xi_{k-1}}[(s^k)^T(x^k + \bar{d}^k - x^k_*)] = 0.$$

Using this relation, (21) and (22), we obtain that

$$\begin{aligned}
\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] \;&=\; \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k + \bar{d}^k)] = \lim_{k\in K\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k + \bar{d}^k)] \\
&\leq\; \lim_{k\in K\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k_*)] \;=\; \min_{x\in\Re^N} F(x),
\end{aligned}$$

and hence, $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \min_{x\in\Re^N} F(x)$.

(ii) Statement (ii) immediately follows from statement (i), the Markov inequality, and the fact that $F(x^k) \;\geq\; \min_{x\in\Re^N} F(x)$. ∎

3 Computational results

In this section we study the numerical behavior of the RBCNMG method on the ℓ_1-regularized least-squares problem:

$$F^* = \min_{x\in\Re^N} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1,$$

where $A \in \Re^{m\times N}$, $b \in \Re^m$, and $\lambda > 0$ is a regularization parameter. We generated a random instance with $m = 2000$ and $N = 1000$ following the procedure described in [12, Section 6].

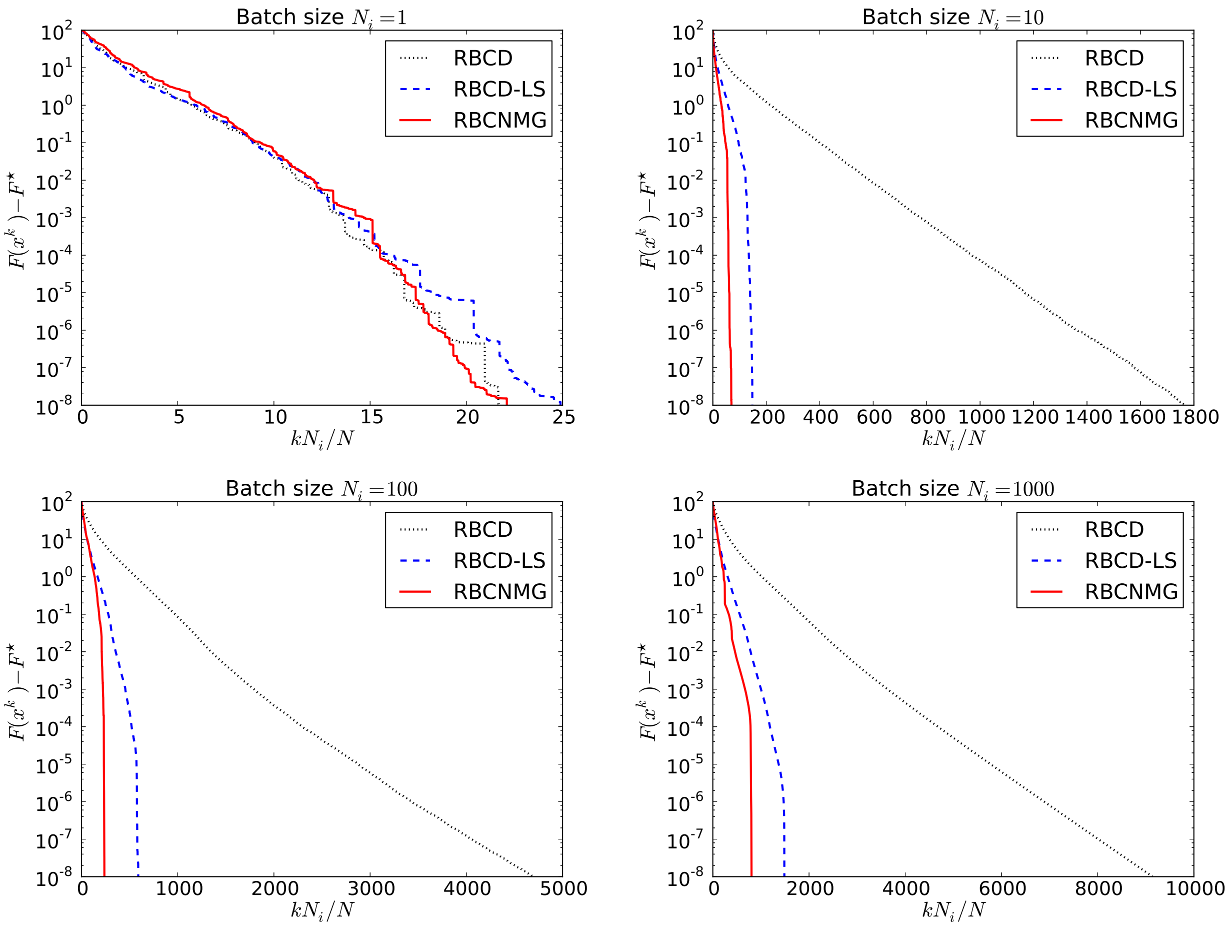

Figure 1: Comparison of different methods when the block-coordinates are chosen uniformly at random ($p_i \propto L_i^\alpha$ with $\alpha = 0$).

The advantage of this procedure is that an optimal solution $x^\star$ is generated together with A and b, and hence the optimal value $F^\star$ is known. We compare the RBCNMG method with two other methods:

- The RBCD method with constant step sizes $1/L_i$.
- The RBCD method with a block-coordinate-wise adaptive line search scheme that is similar to the one used in [12].

We choose the same initial point $x^0 = 0$ for all three methods and terminate the methods once they reach $F(x^k) - F^\star \leq 10^{-8}$.

Figure 1 shows the behavior of different algorithms when the block-coordinates are chosen uniformly at random. We used four different blocksizes, i.e., $N_i = 1, 10, 100, 1000$ respectively for all blocks $i = 1, \ldots, N/N_i$. We note that for the case of $N_i = 1000 = N$ all three methods

	$N_i = 1$	$N_i = 10$	$N_i = 100$	$N_i = 1000$
RBCD	21.7/2.1	1763.3/24.2	4700.8/12.4	9144.0/21.5
RBCD-LS	24.9/3.5	147.9/3.5	590.2/2.8	1488.0/7.0
RBCNMG	22.1/3.8	69.7/1.8	238.4/1.2	806.0/4.5

Table 1: $\alpha = 0$: number of N coordinate passes (kN_i/N) and running time (seconds).

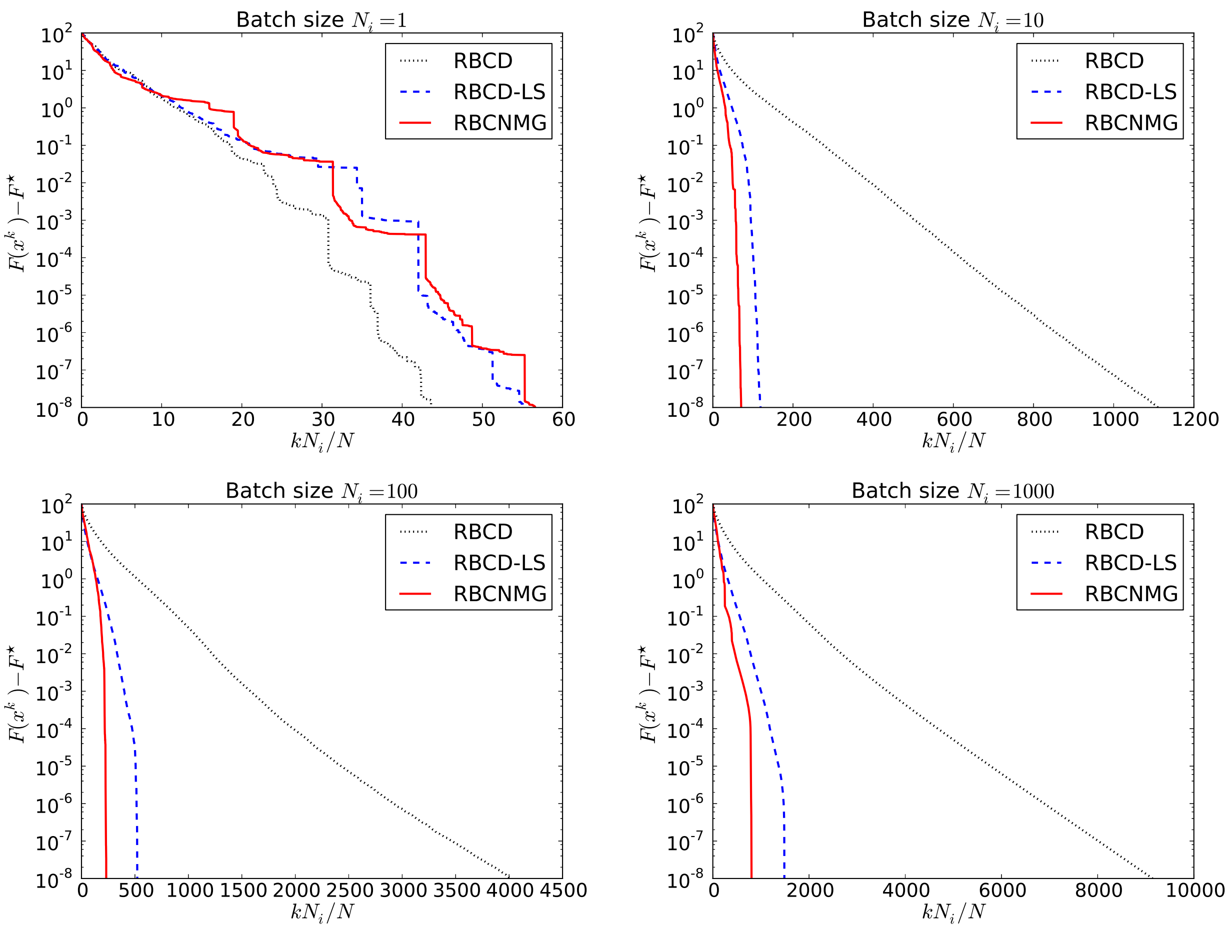

Figure 2: Comparison of different methods when the block-coordinates are chosen with probability $p_i \propto L_i^\alpha$ with $\alpha = 0.5$.

become a deterministic full gradient method. Table 1 gives the number of N coordinate passes, i.e., kN_i/N, and time (in seconds) used by different methods.

In addition, Figures 2 and 3 show the behavior of different algorithms when the block-coordinates are chosen according to the probability $p_i \propto L_i^\alpha$ with $\alpha = 0.5$ and $\alpha = 1$, respectively, while Tables 2 and 3 present the number of N coordinate passes and time (in seconds) used by different methods to reach $F(x^k) - F^\star \leq 10^{-8}$ under these probability distributions.

Here we give some observations and discussions:

	$N_i = 1$	$N_i = 10$	$N_i = 100$	$N_i = 1000$
RBCD	43.6/4.2	1110.3/14.5	4018.6/11.1	9144.0/21.5
RBCD-LS	55.0/8.2	119.2/2.9	522.1/2.3	1488.0/7.0
RBCNMG	56.6/9.6	71.0/1.6	231.5/1.1	806.0/4.5

Table 2: $\alpha = 0.5$: number of N coordinate passes (kN_i/N) and running time (seconds).

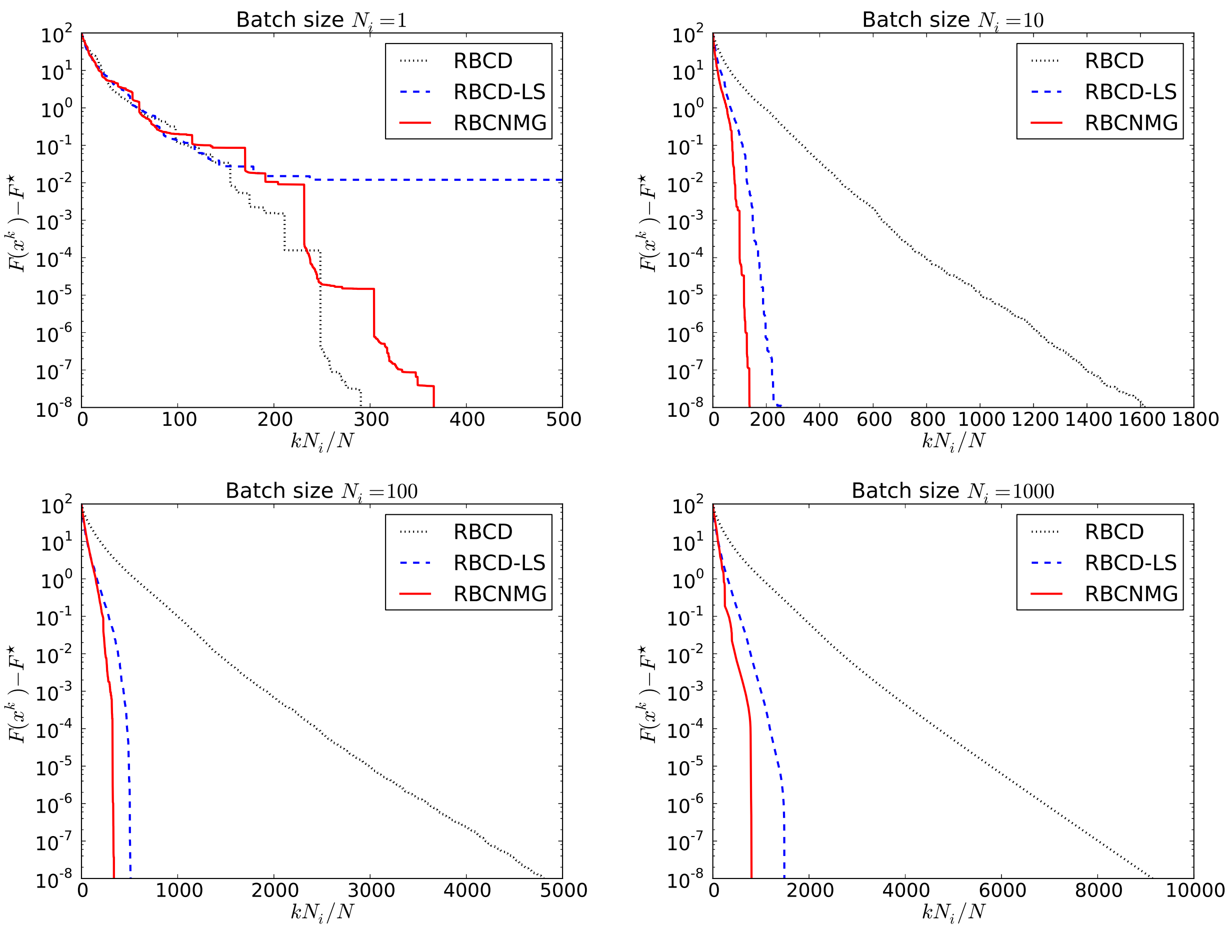

Figure 3: Comparison of different methods when the block-coordinates are chosen with probability $p_i \propto L_i^\alpha$ with $\alpha = 1$.

- When the blocksize (batch size) $N_i = 1$, these three methods behave similarly. The reason is that in this case, along each coordinate the function f is one-dimension quadratic, different line search methods have roughly the same estimate of the partial Lipschitz constant, and uses roughly the same stepsize.

- When $N_i = N$, the behavior of the full gradient methods are the same for $\alpha = 0, 0.5, 1$. That is, the last subplot in the three figures are identical.

- Increasing the blocksize tends to reduce the computation time initially, but eventually slows down again when the blocksize is too big.
- The total number of N coordinate passes (kN_i/N) is not a good indicator of computation time.
- The methods RBCD and RBCD-LS generally perform better when the block-coordinates are chosen with probability $p_i \propto L_i^\alpha$ with $\alpha = 0.5$ than the other two probability distributions. Nevertheless, the RBCNMG method seems to perform best when the block-coordinates are chosen uniformly at random.
- Our RBCNMG method outperforms the other two methods when the blocksize $N_i > 1$. Moreover, it is substantially superior to the full gradient method when the blocksizes are appropriately chosen.

References

[1] J. Barzilai and J. M. Borwein. Two point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148, 1988.

[2] E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. Optimiz*, 4:1196–1211, 2000.

[3] 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.

[4] F. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, Philadephia, PA, USA, 1990.

[5] Y. H. Dai and H. C. Zhang. Adaptive two-pint stepsize gradient algorithm. *Numerical Algorithms*, 27:377–385, 2001.

[6] 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.

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

	$N_i = 1$	$N_i = 10$	$N_i = 100$	$N_i = 1000$
RBCD	290.4/28.3	1622.7/21.8	4812.3/12.6	9144.0/21.5
RBCD-LS	500.0/90.7	278.1/6.0	511.6/2.7	1488.0/7.0
RBCNMG	366.2/61.6	138.4/3.4	337.2/1.7	806.0/4.5

Table 3: $\alpha = 1$: number of N coordinate passes (kN_i/N) and running time (seconds).

[8] 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.

[9] Z. Lu and L. Xiao. On the complexity analysis of randomized block coordinate descent methods. Submitted, May 2013.

[10] Z. Lu and Y. Zhang. An augmented Lagrangian approach for sparse principal component analysis. *Mathematical Programming* 135, pp. 149–193 (2012).

[11] 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.

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

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

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

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

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

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

[18] 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.

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

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

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

[22] 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.

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

[24] E. Van Den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comp.*, 31(2):890-912, 2008.

[25] 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.

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

[27] S. J. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE T. Image Process.*, 57:2479–2493, 2009.

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

[29] 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.