A Randomized Nonmonotone Block Proximal Gradient Method for a Class of Structured Nonlinear Programming

Zhaosong Lu* Lin Xiao †

June 8, 2014

Abstract

In this paper we propose a randomized nonmonotone block proximal gradient (RNBPG) 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 solves typically several associated proximal subproblems that usually have a closed-form solution, until a certain progress on objective value is achieved. In contrast to the usual randomized block coordinate descent method [22, 19], our method enjoys a nonmonotone flavor and uses a variable stepsize that can partially utilize the local curvature information of the smooth component of objective function. We show that the expected objective values generated by the method converge to the expected limit of the objective values obtained by a random single run of the method. Moreover, any accumulation point of the solution sequence of the method is a stationary point of the problem *almost surely* and the method is capable of finding an approximate stationary point with high probability. We also establish a sublinear rate of convergence for the method in terms of the minimal expected squared norm of certain proximal gradients over the iterations. When the problem under consideration is convex, we show that the expected objective values generated by RNBPG converge to the optimal value of the problem. Under some assumptions, we further establish a sublinear and linear rate of convergence on the expected objective values generated by a monotone version of RNBPG. Finally, we conduct some preliminary experiments to test the performance of RNBPG 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 [22].

Key words: Randomized block proximal gradient method, structured minimization

*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).

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, 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, [4, 11, 34, 13, 28, 29, 31, 32, 20, 35, 24, 12, 21, 25, 23, 26]). Recently, Nesterov [18] proposed a randomized BCD (RBCD) method, which is promising for solving a class of huge-scale convex optimization problems, provided the involved partial gradients can be efficiently updated. The iteration complexity for finding an approximate optimal solution is analyzed in [18]. More recently, Richtárik and Takáč [22] extended Nesterov's RBCD method [18] to solve a more general class of convex optimization problems in the form of

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

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 [22] picks $i \in \{1, \ldots, n\}$ uniformly, solves a block-wise proximal subproblem in the form of

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

and sets $x_i^{k+1} = x_i^k + d_i(x^k)$ 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 $\|\cdot\|$ (see Assumption 1 for details). The iteration complexity of finding an approximate optimal solution with high probability is established in [22] and has recently been improved by Lu and Xiao [14]. Very recently, Patrascu and Necoara [19] extended this method to solve problem (1) in which F is nonconvex, and they studied convergence of the method under the assumption that the block is chosen uniformly at each iteration.

One can observe that for $n = 1$, the RBCD method [22, 19] becomes a classical proximal (full) gradient method with a constant stepsize $1/L$. It is known that the latter method tends to be practically much slower than the same type of methods but with a variable stepsize, for example, spectral-type stepsize [1, 3, 6, 33, 15]) that utilizes partial local curvature information of the smooth component f. The variable stepsize strategy shall also be applicable to the RBCD method and improve its practical performance dramatically. In addition, the RBCD method is a monotone method, that is, the objective values generated by the method are monotonically decreasing. As mentioned in the literature (see, for example, [7, 8, 36]), nonmontone

methods often produce solutions of better quality than the monotone counterparts for nonconvex optimization problems. These motivate us to propose a randomized nonmonotone block proximal gradient method with a variable stepsize for solving a class of (possibly nonconvex) structured nonlinear programming problems in the form of (1) satisfying Assumption 1 below.

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 $F^\star$. For simplicity of presentation, we associate $\Re^N$ with the standard Euclidean norm, denoted by $\|\cdot\|$. We also make the following assumption.

Assumption 1 *f is differentiable (but possibly nonconvex) in $\Re^N$. Each Ψ_i is a (possibly nonconvex nonsmooth) function from $\Re^{N_i}$ to $\Re \cup \{+\infty\}$ for $i = 1, \ldots, n$. The gradient of function f is coordinate-wise Lipschitz continuous with constants $L_i > 0$ in $\Re^N$, that is,*

$$\|\nabla_i f(x+h) - \nabla_i f(x)\| \;\leq\; L_i\|h\| \qquad \forall h \in \mathcal{S}_i,\; i = 1, \ldots, n;\; \forall x \in \Re^N,$$

where

$$\mathcal{S}_i = \left\{(h_1, \ldots, h_n) \in \Re^{N_1} \times \cdots \times \Re^{N_n} : h_j = 0 \quad \forall j \neq i\right\}.$$

In this paper we propose a randomized nonmonotone block proximal gradient (RNBPG) method for solving problem (1) that satisfies the above assumptions. At each iteration, this method randomly picks a block according to any prescribed (not necessarily uniform) probability distribution and solves typically several associated proximal subproblems in the form of (2) with L_i replaced by some θ, which can be, for example, estimated by the spectral method (e.g., see [1, 3, 6, 33, 15]), until a certain progress on the objective value is achieved. In contrast to the usual RBCD method [22, 19], our method enjoys a nonmonotone flavor and uses a variable stepsize that can partially utilize the local curvature information of the smooth component f. For arbitrary probability distribution[1], We show that the expected objective values generated by the method converge to the expected limit of the objective values obtained by a random single run of the method. Moreover, any accumulation point of the solution sequence of the method is a stationary point of the problem *almost surely* and the method is capable of finding an approximate stationary point with high probability. We also establish a sublinear rate of convergence for the method in terms of the minimal expected squared norm of certain proximal gradients over the iterations. When the problem under consideration is convex, we show that the expected objective values generated by RNBPG converge to the optimal value of the problem. Under some assumptions, we further establish a sublinear and linear rate of convergence on the expected objective values generated by a monotone version of RNBPG. Finally, we conduct some preliminary experiments to test the performance of RNBPG 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 [22].

This paper is organized as follows. In Section 2 we propose a RNBPG method for solving structured nonlinear programming problem (1) and analyze its convergence. In Section 3 we

[1]The convergence analysis of the RBCD method conducted in [22, 19] is only for uniform probability distribution.

analyze the convergence of RNBPG for solving structured convex problem. In Section 4 we conduct numerical experiments to compare RNBPG method with the RBCD method [22] 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 some known facts. The domain of the function F is denoted by $\mathrm{dom}(F)$. t^+ stands for $\max\{0, t\}$ for any real number t. Given a closed set S and a point x, $\mathrm{dist}(x, S)$ denotes the distance between x and S. For symmetric matrices X and Y, $X \preceq Y$ means that $Y - X$ is positive semidefinite. Given a positive definite matrix Θ and a vector x, $\|x\|_\Theta = \sqrt{x^T \Theta x}$. In addition, $\|\cdot\|$ denotes the Euclidean norm. Finally, it immediately follows from Assumption 1 that

$$f(x+h) \le f(x) + \nabla f(x)^T h + \frac{L_i}{2}\|h\|^2 \qquad \forall h \in \mathcal{S}_i,\ i = 1, \ldots, n;\ \forall x \in \Re^N. \tag{3}$$

By Lemma 2 of Nesterov [18] and Assumption 1, we also know that ∇f is Lipschitz continuous with constant $L_f := \sum_i L_i$, that is,

$$\|\nabla f(x) - \nabla f(y)\| \le L_f \|x - y\| \qquad x, y \in \Re^N. \tag{4}$$

2 Randomized nonmonotone block proximal gradient method

In this section we propose a RNBPG method for solving structured nonlinear programming problem (1) and analyze its convergence.

We start by presenting a RNBPG method as follows. At each iteration, this method randomly picks a block according to any prescribed (not necessarily uniform) probability distribution and solves typically several associated proximal subproblems in the form of (2) with L_i replaced by some θ_k until a certain progress on objective value is achieved.

Randomized nonmonotone block proximal gradient (RNBPG) method

Choose $x^0 \in \mathrm{dom}(F)$, $\eta > 1$, $\sigma > 0$, $0 < \underline{\theta} \le \bar{\theta}$, integer $M \ge 0$, 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. Choose $\theta_k^0 \in [\underline{\theta}, \bar{\theta}]$.

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

   2a) Let $\theta_k = \theta_k^0 \eta^j$. Compute

$$(d^k)_{i_k} = \arg\min_s \left\{ \nabla_{i_k} f(x^k)^T s + \frac{\theta_k}{2}\|s\|^2 + \Psi_{i_k}(x_{i_k}^k + s) \right\}.$$

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 $x^{k+1} = x^k + d^k$, $k \leftarrow k+1$ and go to step 1).

end

Remark 2.1 *The above method becomes a monotone method if* $M = 0$. ∎

Before studying convergence of RNBPG, we introduce some notations and state some facts that will be used subsequently.

Let $\bar{d}^{k,i}$ denote the vector d^k obtained in Step (2) of RNBPG if i_k is chosen to be i. Define

$$\bar{d}^k = \sum_{i=1}^n \bar{d}^{k,i}, \qquad \bar{x}^k = x^k + \bar{d}^k. \tag{6}$$

One can observe that $(\bar{d}^{k,i})_t = 0$ for $t \neq i$ and there exist $\theta^0_{k,i} \in [\underline{\theta}, \bar{\theta}]$ and the smallest nonnegative integer j such that $\theta_{k,i} = \theta^0_{k,i}\eta^j$ and

$$F(x^k + \bar{d}^{k,i}) \leq F(x^{\ell(k)}) - \frac{\sigma}{2}\|\bar{d}^{k,i}\|^2, \tag{7}$$

where

$$(\bar{d}^{k,i})_i = \arg\min_s \left\{ \nabla_i f(x^k)^T s + \frac{\theta_{k,i}}{2}\|s\|^2 + \Psi_i(x^k_i + s) \right\}, \tag{8}$$

$$\ell(k) = \arg\max_i \{F(x^i) : i = [k-M]^+, \ldots, k\} \qquad \forall k \geq 0. \tag{9}$$

Let Θ_k denote the block diagonal matrix $(\theta_{k,1}I_1, \ldots, \theta_{k,n}I_n)$, where I_i is the $N_i \times N_i$ identity matrix. By the definition of $\bar{d}^k$ and (8), we observe that

$$\bar{d}^k = \arg\min_d \left\{ \nabla f(x^k)^T d + \frac{1}{2} d^T \Theta_k d + \Psi(x^k + d) \right\}. \tag{10}$$

After k iterations, RNBPG 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 $\mathbf{E}_{\xi_{-1}}[F(x^0)] = F(x^0)$. Also, define

$$\Omega(x^0) = \{x \in \Re^N : F(x) \leq F(x^0)\}, \tag{11}$$

$$L_{\max} = \max_i L_i, \qquad p_{\min} = \min_i p_i, \tag{12}$$

$$c = \max\left\{\bar{\theta}, \eta(L_{\max} + \sigma)\right\}. \tag{13}$$

The following lemma establishes some relations between the expectations of $\|d^k\|$ and $\|\bar{d}^k\|$.

Lemma 2.2 *Let d^k be generated by RNBPG and $\bar{d}^k$ defined in (6). There hold*

$$\mathbf{E}_{\xi_k}[\|d^k\|^2] \geq p_{\min}\ \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|^2], \tag{14}$$

$$\mathbf{E}_{\xi_k}[\|d^k\|] \geq p_{\min}\ \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|]. \tag{15}$$

Proof. By (12) and the definitions of d^k and $\bar{d}^k$, we can observe that

$$\mathbf{E}_{i_k}[\|d^k\|^2] = \sum_i p_i\|\bar{d}^{k,i}\|^2 \geq (\min_i p_i)\sum_i \|\bar{d}^{k,i}\|^2 = p_{\min}\|\bar{d}^k\|^2,$$

$$\mathbf{E}_{i_k}[\|d^k\|] = \sum_i p_i\|\bar{d}^{k,i}\| \geq (\min_i p_i)\sum_i \|\bar{d}^{k,i}\| \geq p_{\min}\sqrt{\sum_i \|\bar{d}^{k,i}\|^2} \geq p_{\min}\|\bar{d}^k\|.$$

The conclusion of this lemma follows by taking expectation with respect to ξ_{k-1} on both sides of the above inequalities. ∎

We next show that the inner loops of the above RNBPG method must terminate finitely. As a byproduct, we provide a uniform upper bound on Θ_k.

Lemma 2.3 *Let $\{\theta_k\}$ be the sequence generated by RNBPG, Θ_k defined above, and c defined in (13). There hold*

(i) $\underline{\theta} \leq \theta_k \leq c \qquad \forall k.$

(ii) $\underline{\theta} I \preceq \Theta_k \preceq cI \qquad \forall k.$

Proof. (i) It is clear that $\theta_k \geq \underline{\theta}$. We now show $\theta_k \leq c$ by dividing the proof into two cases.

Case (i) $\theta_k = \theta_k^0$. Since $\theta_k^0 \leq \bar{\theta}$, it follows that $\theta_k \leq \bar{\theta}$ and the conclusion holds.

Case (ii) $\theta_k = \theta_k^0\eta^j$ for some integer $j > 0$. Suppose for contradiction that $\theta_k > c$. By (12) and (13), we then have

$$\tilde{\theta}_k := \theta_k/\eta > c/\eta \geq L_{\max} + \sigma \geq L_{i_k} + \sigma. \tag{16}$$

Let $d \in \Re^N$ such that $d_i = 0$ for $i \neq i_k$ and

$$d_{i_k} = \arg\min_s \left\{ \nabla_{i_k} f(x^k)^T s + \frac{\tilde{\theta}_k}{2}\|s\|^2 + \Psi_{i_k}(x_{i_k}^k + s) \right\}. \tag{17}$$

It follows that

$$\nabla_{i_k} f(x^k)^T d_{i_k} + \frac{\tilde{\theta}_k}{2}\|d_{i_k}\|^2 + \Psi_{i_k}(x_{i_k}^k + d_{i_k}) - \Psi_{i_k}(x_{i_k}^k) \leq 0.$$

Also, by (9) and the definitions of θ_k and $\tilde{\theta}_k$, one knows that

$$F(x^k + d) > F(x^{\ell(k)}) - \frac{\sigma}{2}\|d\|^2. \tag{18}$$

On the other hand, using (3), (9), (16), (17) and the definition of d, we have

$$\begin{aligned} F(x^k+d) &= f(x^k+d)+\Psi(x^k+d) \;\le\; f(x^k)+\nabla_{i_k} f(x^k)^T d_{i_k}+\tfrac{L_{i_k}}{2}\|d_{i_k}\|^2+\Psi(x^k+d) \\ &= F(x^k)+\underbrace{\nabla_{i_k} f(x^k)^T d_{i_k}+\frac{\tilde\theta_k}{2}\|d_{i_k}\|^2+\Psi_{i_k}(x^k_{i_k}+d_{i_k})-\Psi_{i_k}(x^k_{i_k})}_{\le 0}+\tfrac{L_{i_k}-\tilde\theta_k}{2}\|d_{i_k}\|^2 \\ &\le F(x^k)+\tfrac{L_{i_k}-\tilde\theta_k}{2}\|d_{i_k}\|^2 \;\le\; F(x^{\ell(k)})-\tfrac{\sigma}{2}\|d\|^2, \end{aligned}$$

which is a contradiction to (18). Hence, $\theta_k \le c$ and the conclusion holds.

(ii) Let $\theta^{k,i}$ be defined above. It follows from statement (i) that $\underline{\theta} \le \theta_{k,i} \le c$, which together with the definition of Θ_k implies that statement (ii) holds. ∎

The next result provides some bound on the norm of a proximal gradient, which will be used in the subsequent analysis on convergence rate of RNBPG.

Lemma 2.4 *Let $\{x^k\}$ be generated by RNBPG, $\bar d^k$ and c defined in (10) and (13), respectively, and*

$$\hat g^k = \arg\min_d \left\{ \nabla f(x^k)^T d+\frac{1}{2}\|d\|^2+\Psi(x^k+d)\right\}. \tag{19}$$

Assume that Ψ is convex. There holds

$$\|\hat g^k\| \le \frac{c}{2}\left[1+\frac{1}{\underline{\theta}}+\sqrt{1-\frac{2}{c}+\frac{1}{\underline{\theta}^2}}\,\right]\|\bar d^k\|. \tag{20}$$

Proof. The conclusion of this lemma follows from (10), (19), Lemma 2.3 (ii), and [15, Lemma 3.5] with $H=\Theta_k$, $\tilde H = I$, $Q=\Theta_k^{-1}$, $d=\bar d^k$ and $\tilde d = \hat g^k$. ∎

The following lemma studies uniform continuity of the expectation of F with respect to random sequences.

Lemma 2.5 *Suppose that F is uniform continuous in some $S \subseteq \mathrm{dom}(F)$. Let y^k and z^k be two random vectors in S generated from ξ_{k-1}. Assume that there exists $C>0$ such that $|F(y^k)-F(z^k)| \le C$ for all k, and moreover,*

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

Then there hold

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

Proof. Since F is uniformly continuous in S, it follows that given any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that $|F(x) - F(y)| < \epsilon/2$ for all $x, y \in S$ satisfying $\|x - y\| < \delta_\epsilon$. Using these relations, the Markov inequality, and the assumption that $|F(y^k) - F(z^k)| \le C$ for all k and $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\Delta^k\|] = 0$, where $\Delta^k = y^k - z^k$, we obtain that for sufficiently large k,

$$\begin{aligned}\mathbf{E}_{\xi_{k-1}}[|F(y^k) - F(z^k)|] &= \mathbf{E}_{\xi_{k-1}}\left[|F(y^k) - F(z^k)| \;\middle|\; \|\Delta^k\| \ge \delta_\epsilon\right] \mathbf{P}(\|\Delta^k\| \ge \delta_\epsilon) \\ &\quad + \mathbf{E}_{\xi_{k-1}}\left[|F(y^k) - F(z^k)| \;\middle|\; \|\Delta^k\| < \delta_\epsilon\right] \mathbf{P}(\|\Delta^k\| < \delta_\epsilon) \\ &\le \frac{C\mathbf{E}_{\xi_{k-1}}[\|\Delta^k\|]}{\delta_\epsilon} + \frac{\epsilon}{2} \;\le\; \epsilon.\end{aligned}$$

Due to the arbitrarity of ϵ, we see that the first statement of this lemma holds. The second statement immediately follows from the first statement and the well-known inequality

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

■

We are ready to establish the first main result, that is, the expected objective values generated by the RNBPG method converge to the expected limit of the objective values obtained by a random single run of the method.

Theorem 2.6 *Let $\{x^k\}$ and $\{d^k\}$ be the sequences generated by the RNBPG method. Assume that F is uniform continuous in $\Omega(x^0)$, where $\Omega(x^0)$ is defined in (11). Then the following statements hold:*

(i) $\lim_{k\to\infty} \|d^k\| = 0$ *and* $\lim_{k\to\infty} F(x^k) = F^*_{\xi_\infty}$ *for some* $F^*_{\xi_\infty} \in \Re$, *where* $\xi_\infty = \{i_1, i_2, \cdots\}$.

(ii) $\lim_{k\to\infty} \mathbf{E}_{\xi_k}[\|d^k\|] = 0$ *and*

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

Proof. By (5) and (9), we have

$$F(x^{k+1}) \;\le\; F(x^{\ell(k)}) - \frac{\sigma}{2}\|d^k\|^2 \qquad \forall k \ge 0. \tag{22}$$

Hence, $F(x^{k+1}) \le F(x^{\ell(k)})$, which together with (9) implies that $F(x^{\ell(k+1)}) \le F(x^{\ell(k)})$. It then follows that

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

Hence, $\{F(x^{\ell(k)})\}$ and $\{\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})]\}$ are non-increasing. Since F is bounded below, so are $\{F(x^{\ell(k)})\}$ and $\{\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})]\}$. It follows that there exist some $F^*_{\xi_\infty}$, $\tilde{F}^* \in \Re$ such that

$$\lim_{k\to\infty} F(x^{\ell(k)}) = F^*_{\xi_\infty}, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] = \tilde{F}^*. \tag{23}$$

We first show by induction that the following relations hold for all $j \geq 1$:

$$\lim_{k\to\infty} \|d^{\ell(k)-j}\| = 0, \qquad \lim_{k\to\infty} F(x^{\ell(k)-j}) = F^*_{\xi_\infty}. \tag{24}$$

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-j}\|] = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)-j})] = \tilde{F}^*. \tag{25}$$

Indeed, replacing k by $\ell(k)-1$ in (22), we obtain that

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

which together with $\ell(k) \geq k-M$ and monotonicity of $\{F(x^{\ell(k)})\}$ yields

$$F(x^{\ell(k)}) \;\leq\; F(x^{\ell(k-M-1)}) - \frac{\sigma}{2}\|d^{\ell(k)-1}\|^2 \qquad \forall k \geq M+1. \tag{26}$$

Then we have

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \;\leq\; \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k-M-1)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|^2] \qquad \forall k \geq M+1. \tag{27}$$

Notice that

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

It follows from this relation and (27) that

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \;\leq\; \mathbf{E}_{\xi_{k-M-2}}[F(x^{\ell(k-M-1)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|^2] \qquad \forall k \geq M+1. \tag{28}$$

In view of (23), (26), (28), and $(\mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|])^2 \leq \mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|^2]$, one can have

$$\lim_{k\to\infty} \|d^{\ell(k)-1}\| = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|] = 0. \tag{29}$$

One can also observe that $F(x^k) \leq F(x^0)$ and hence $\{x^k\} \subset \Omega(x^0)$. Using this fact, (23), (29), Lemma 2.5, and uniform continuity of F over $\Omega(x^0)$, we obtain that

$$\lim_{k\to\infty} F(x^{\ell(k)-1}) \;=\; \lim_{k\to\infty} F(x^{\ell(k)}) \;=\; F^*_{\xi_\infty},$$

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

Therefore, (24) and (25) hold for $j=1$. Suppose now that they hold for some $j \geq 1$. We need to show that they also hold for $j+1$. Replacing k by $\ell(k)-j-1$ in (22) gives

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

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

$$F(x^{\ell(k)-j}) \;\leq\; F(x^{\ell(k-M-j-1)}) - \frac{\sigma}{2}\|d^{\ell(k)-j-1}\|^2 \qquad \forall k \geq M+j+1. \tag{30}$$

Then we obtain that

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

Notice that

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

It follows from these two relations that

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)-j})] \le \mathbf{E}_{\xi_{k-M-j-2}}[F(x^{\ell(k-M-j-1)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-j-1}\|^2], \quad \forall k \ge M+j+1. \tag{31}$$

Using (23), (30), (31), the induction hypothesis, and a similar argument as above, we can obtain that

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

These relations, together with Lemma 2.5, uniform continuity of F over $\Omega(x^0)$ and the induction hypothesis, yield

$$\lim_{k\to\infty} F(x^{\ell(k)-j-1}) \;=\; \lim_{k\to\infty} F(x^{\ell(k)-j}) \;=\; F^*_{\xi_\infty},$$

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

Hence, (24) and (25) hold for $j+1$, and the proof of (24) and (25) is completed.

For all $k \ge 2M+1$, 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

$$\|\tilde{d}^{\ell(k)-j}\| \;\le\; \|d^{\ell(k)-j}\|, \tag{32}$$

$$x^{\ell(k)} \;=\; x^{k-M-1} + \sum_{j=1}^{M+1} \tilde{d}^{\ell(k)-j}. \tag{33}$$

It follows from (24), (25) and (32) that $\lim_{k\to\infty} \|\tilde{d}^{\ell(k)-j}\| = 0$ and $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\tilde{d}^{\ell(k)-j}\|] = 0$ for $j = 1, \ldots, M+1$. Hence,

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

These, together with (24), (25), (33), Lemma 2.5 and uniform continuity of F over $\Omega(x^0)$, imply that

$$\lim_{k\to\infty} F(x^{k-M-1}) \;=\; \lim_{k\to\infty} F(x^{\ell(k)}) = F^*_{\xi_\infty}, \tag{34}$$

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{k-M-1})] \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] = \tilde{F}^*. \tag{35}$$

It follows from (34) that $\lim_{k\to\infty} F(x^k) = F^*_{\xi_\infty}$. Using this, (22) and (23), one can see that $\lim_{k\to\infty} \|d^k\| = 0$. Hence, statement (i) holds. Notice that $\mathbf{E}_{\xi_{k-M-2}}[F(x^{k-M-1})] = \mathbf{E}_{\xi_{k-1}}[F(x^{k-M-1})]$. Combining this relation with (35), we have

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

which is equivalent to

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

In addition, it follows from (22) that

$$\mathbf{E}_{\xi_k}[F(x^{k+1})] \;\le\; \mathbf{E}_{\xi_k}[F(x^{\ell(k)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_k}[\|d^k\|^2] \qquad \forall k \ge 0. \tag{36}$$

Notice that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_k}[F(x^{\ell(k)})] \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \;=\; \tilde{F}^* \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_k}[F(x^{k+1})]. \tag{37}$$

Using (36) and (37), we conclude that $\lim_{k\to\infty} \mathbf{E}_{\xi_k}[\|d^k\|] = 0$.

Finally, we claim that $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}]$. Indeed, we know that $\{x^k\} \subset \Omega(x^0)$. Hence, $F^* \le F(x^k) \le F(x^0)$, where $F^* = \min_x F(x)$. It follows that

$$|F(x^k)| \le \max\{|F(x^0)|, |F^*|\} \qquad \forall k.$$

Using this relation and dominated convergence theorem (see, for example, [2, Theorem 5.4]), we have

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

which, together with $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \lim_{k\to\infty} \mathbf{E}_{\xi_\infty}[F(x^k)]$, implies that $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}]$. Hence, statement (ii) holds.

■

The following result shows that when k is sufficiently large, x^k is an approximate stationary point of (1) with high probability.

Theorem 2.7 *Let $\{x^k\}$ be generated by RNBPG, and $\bar{d}^k$ and $\bar{x}^k$ defined in (6). Assume that F is uniformly continuous and Ψ is locally Lipschitz continuous in $\Omega(x^0)$, where $\Omega(x^0)$ is defined in (11). Then there hold*

(i)

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|] = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\mathrm{dist}(-\nabla f(\bar{x}^k), \partial\Psi(\bar{x}^k)] \;=\; 0, \tag{38}$$

where $\partial\Psi$ denotes the Clarke subdifferential of Ψ.

(ii) Any accumulation point of $\{x^k\}$ is a stationary point of problem (1) almost surely.

(iii) Suppose further that F is uniformly continuous in

$$\mathcal{S} = \left\{ x : F(x) \leq F(x^0) + \max\left\{ \frac{n}{\sigma} |L_f - \underline{\theta}|, 1 \right\} (F(x^0) - F^*) \right\}. \tag{39}$$

Then $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[|F(x^k) - F(\bar{x}^k)|] = 0$. Moreover, for any $\epsilon > 0$ and $\rho \in (0,1)$, there exists K such that for all $k \geq K$,

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

Proof. (i) We know from Theorem 2.6 (ii) that $\lim_{k\to\infty} \mathbf{E}_{\xi_k}[\|d^k\|] = 0$, which together with (15) implies $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|] = 0$. Notice that $\bar{d}^k$ is an optimal solution of problem (10). By the first-order optimality condition (see, for example, Proposition 2.3.2 of [5]) of (10) and $\bar{x}^k = x^k + \bar{d}^k$, one can have

$$0 \in \nabla f(x^k) + \Theta_k \bar{d}^k + \partial\Psi(\bar{x}^k). \tag{40}$$

In addition, it follows from (4) that

$$\|\nabla f(\bar{x}^k) - \nabla f(x^k)\| \;\leq\; L_f \|\bar{d}^k\|.$$

Using this relation along with Lemma 2.3 (ii) and (40), we obtain that

$$\mathrm{dist}(-\nabla f(\bar{x}^k), \partial\Psi(\bar{x}^k)) \;\leq\; (c + L_f)\, \|\bar{d}^k\|,$$

which together with the first relation of (38) implies that the second relation of (38) also holds.

(ii) Let x^* be an accumulation point of $\{x^k\}$. There exists a subsequence $\mathcal{K}$ such that $\lim_{k\in\mathcal{K}\to\infty} x^k = x^*$. Since $\mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|] \to 0$, it follows that $\{\bar{d}^k\}_{k\in\mathcal{K}} \to 0$ almost surely. This together with the second relation of (38) and outer semi-continuity of $\partial\Psi$ yields

$$\mathrm{dist}(-\nabla f(x^*), \partial\Psi(x^*)) = \lim_{k\in\mathcal{K}\to\infty} \mathrm{dist}(-\nabla f(\bar{x}^k), \partial\Psi(\bar{x}^k)) = 0$$

almost surely. Hence, x^* is a stationary point of problem (1) almost surely.

(iii) Recall that $\bar{x}^k = x^k + \bar{d}^k$. It follows from (4) that

$$f(\bar{x}^k) \leq f(x^k) + \nabla f(x^k)^T \bar{d}^k + \frac{1}{2} L_f \|\bar{d}^k\|^2.$$

Using this relation and Lemma 2.3 (ii), we have

$$\begin{aligned} F(\bar{x}^k) \;&\leq\; f(x^k) + \nabla f(x^k)^T \bar{d}^k + \frac{1}{2} L_f \|\bar{d}^k\|^2 + \Psi(x^k + \bar{d}^k) \\ &\leq\; f(x^k) + \nabla f(x^k)^T \bar{d}^k + \frac{1}{2} (\bar{d}^k)^T \Theta_k \bar{d}^k + \Psi(x^k + \bar{d}^k) + \frac{1}{2} (L_f - \underline{\theta}) \|\bar{d}^k\|^2. \end{aligned} \tag{41}$$

In view of (10), one has

$$\nabla f(x^k)^T \bar{d}^k + \frac{1}{2} (\bar{d}^k)^T \Theta_k \bar{d}^k + \Psi(x^k + \bar{d}^k) \leq \Psi(x^k),$$

which together with (41) yields

$$F(\bar{x}^k) \leq F(x^k) + \frac{1}{2}(L_f - \underline{\theta})\|\bar{d}^k\|^2.$$

Using this relation and the fact that $F(\bar{x}^k) \geq F^*$ and $F(x^k) \leq F(x^0)$, one can obtain that

$$|F(\bar{x}^k) - F(x^k)| \leq \max\left\{\frac{1}{2}|L_f - \underline{\theta}|\,\|\bar{d}^k\|^2, F(x^0) - F^*\right\} \qquad \forall k. \tag{42}$$

In addition, since $F^{l(k)} \leq F(x^0)$ and $F(\bar{x}^k) \geq F^*$, it follows from (7) that $\|\bar{d}^{k,i}\|^2 \leq 2(F(x^0) - F^*)/\sigma$. Hence, one has

$$\|\bar{d}^k\|^2 = \sum_{i=1}^n \|\bar{d}^{k,i}\|^2 \leq 2n(F(x^0) - F^*)/\sigma \qquad \forall k.$$

This inequality together with (42) yields

$$|F(\bar{x}^k) - F(x^k)| \leq \max\left\{\frac{n}{\sigma}|L_f - \underline{\theta}|, 1\right\}(F(x^0) - F^*) \qquad \forall k,$$

and hence $\{|F(\bar{x}^k) - F(x^k)|\}$ is bounded. Also, this inequality together with $F(x^k) \leq F(x^0)$ and the definition of $\mathcal{S}$ implies that $\bar{x}^k$, $x^k \in \mathcal{S}$ for all k. In addition, by statement (i), we know $\mathbf{E}_{\xi_{k-1}}[\|x^k - \bar{x}^k\|] \to 0$. In view of these facts and invoking Lemma 2.5, one has

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[|F(x^k) - F(\bar{x}^k)|] \;=\; 0. \tag{43}$$

Observe that

$$\begin{aligned} 0 \;&\leq\; \max\left\{\|x^k - \bar{x}^k\|, |F(x^k) - F(\bar{x}^k)|, \mathrm{dist}(-\nabla f(\bar{x}^k), \partial\Psi(\bar{x}^k))\right\} \\ &\leq\; \|x^k - \bar{x}^k\| + |F(x^k) - F(\bar{x}^k)| + \mathrm{dist}(-\nabla f(\bar{x}^k), \partial\Psi(\bar{x}^k)). \end{aligned}$$

Using these inequalities, (43) and statement (i), we see that

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

The rest of statement (iii) follows from this relation and the Markov inequality. ■

Before ending this section we establish a sublinear rate of convergence of RNBPG in terms of the minimal expected squared norm of certain proximal gradients over the iterations.

Theorem 2.8 *Let $\bar{g}^k = -\Theta_k\bar{d}^k$, $p_{\min}$, $\hat{g}^k$ and c be defined in (12), (19) and (13), respectively, and F^* the optimal value of (1). The following statements hold*

(i)

$$\min_{1\leq t\leq k} \mathbf{E}_{\xi_{t-1}}[\|\bar{g}^t\|^2] \leq \frac{2c^2(F(x^0) - F^*)}{\sigma p_{\min}} \cdot \frac{1}{\lfloor (k+1)/(M+1)\rfloor} \qquad \forall k \geq M.$$

(ii) Assume further that Ψ is convex. Then

$$\min_{1\le t\le k}\mathbf{E}_{\xi_{t-1}}[\|\hat{g}^t\|^2]\le\frac{c^2(F(x^0)-F^*)}{2\sigma p_{\min}}\left[1+\frac{1}{\underline{\theta}}+\sqrt{1-\frac{2}{c}+\frac{1}{\underline{\theta}^2}}\right]^2\cdot\frac{1}{\lfloor(k+1)/(M+1)\rfloor}\qquad\forall k\ge M.$$

Proof. (i) Using $\bar{g}^k=-\Theta_k\bar{d}^k$, Lemma 2.3 (ii), and (14), one can observe that

$$\mathbf{E}_{\xi_k}[\|d^k\|^2]\ge p_{\min}\mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|^2]=p_{\min}\mathbf{E}_{\xi_{k-1}}[\|\Theta_k^{-1}\bar{g}^k\|^2]\ge\frac{p_{\min}}{c^2}\mathbf{E}_{\xi_{k-1}}[\|\bar{g}^k\|^2].\tag{44}$$

Let $j(t)=\ell((M+1)t)-1$ and $\bar{j}(t)=(M+1)t-1$ for all $t\ge0$. One can see from (28) that

$$\mathbf{E}_{\xi_{\bar{j}(t)}}[F(x^{j(t)+1})]\ \le\ \mathbf{E}_{\xi_{\bar{j}(t-1)}}[F(x^{j(t-1)+1})]-\frac{\sigma}{2}\mathbf{E}_{\xi_{\bar{j}(t)}}[\|d^{j(t)}\|^2]\qquad\forall t\ge1.$$

Summing up the above inequality over $t=1,\ldots,s$, we have

$$\mathbf{E}_{\xi_{\bar{j}(s)}}[F(x^{j(s)+1})]\le F(x^0)-\frac{\sigma}{2}\sum_{t=1}^{s}\mathbf{E}_{\xi_{\bar{j}(t)}}[\|d^{j(t)}\|^2]\le F(x^0)-\frac{\sigma s}{2}\min_{1\le t\le s}\mathbf{E}_{\xi_{\bar{j}(t)}}[\|d^{j(t)}\|^2],$$

which together with $\mathbf{E}_{\xi_{\bar{j}(s)}}[F(x^{j(s)+1})]\ge F^*$ implies that

$$\min_{1\le t\le s}\mathbf{E}_{\xi_{\bar{j}(t)}}[\|d^{j(t)}\|^2]\le\frac{2(F(x^0)-F^*)}{\sigma s}.\tag{45}$$

Given any $k\ge M$, let $s_k=\lfloor(k+1)/(M+1)\rfloor$. Observe that

$$\bar{j}(s_k)=(M+1)s_k-1\le k.$$

Using this relation and (45), we have

$$\min_{1\le t\le k}\mathbf{E}_{\xi_t}[\|d^t\|^2]\le\min_{1\le\tilde{t}\le s_k}\mathbf{E}_{\xi_{\bar{j}(\tilde{t})}}[\|d^{j(\tilde{t})}\|^2]\le\frac{2(F(x^0)-F^*)}{\sigma\lfloor(k+1)/(M+1)\rfloor}\qquad\forall k\ge M,$$

which together with (44) implies that statement (i) holds.

(ii) It follows from (14) and (45) that

$$\min_{1\le t\le s}\mathbf{E}_{\xi_{\bar{j}(t)-1}}[\|\bar{d}^{j(t)}\|^2]\le\frac{2(F(x^0)-F^*)}{\sigma s p_{\min}}.$$

Using this relation and a similar argument as above, one has

$$\min_{1\le t\le k}\mathbf{E}_{\xi_{t-1}}[\|\bar{d}^t\|^2]\le\min_{1\le\tilde{t}\le s_k}\mathbf{E}_{\xi_{\bar{j}(\tilde{t})-1}}[\|\bar{d}^{j(\tilde{t})}\|^2]\le\frac{2(F(x^0)-F^*)}{\sigma p_{\min}\lfloor(k+1)/(M+1)\rfloor}\qquad\forall k\ge M.$$

Statement (ii) immediately follows from this inequality and (20). ∎

3 Convergence of RNBPG for structured convex problems

In this section we study convergence of RNBPG for solving structured convex problem (1). To this end, we assume throughout this section that f and Ψ are both convex functions.

The following result shows that $F(x^k)$ can be arbitrarily close to the optimal value F^* of (1) with high probability for sufficiently large k.

Theorem 3.1 *Let $\{x^k\}$ be generated by the RNBPG method, and let F^* and X^* the optimal value and the set of optimal solutions of (1), respectively. Suppose that f and Ψ are convex functions and F is uniformly continuous in $\mathcal{S}$, where $\mathcal{S}$ is defined in (39). Assume that there exists a subsequence $\mathcal{K}$ such that $\{\mathbf{E}_{\xi_{k-1}}[\mathrm{dist}(x^k, X^*)]\}_{\mathcal{K}}$ is bounded. Then there hold:*

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

(ii) For any $\epsilon > 0$ and $\rho \in (0,1)$, there exists K such that for all $k \geq K$,
$$\mathbf{P}\left(F(x^k) - F^* \ \leq \ \epsilon\right) \ \geq \ 1-\rho.$$

Proof. (i) Let $\bar{d}^k$ be defined in (6). Using the assumption that F is uniformly continuous in $\mathcal{S}$ and Theorem 2.7, one has

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|] = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|s^k\|] \ = \ 0, \tag{46}$$

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k+\bar{d}^k)] = \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \tilde{F}^* \tag{47}$$

for some $s^k \in \partial F(x^k+\bar{d}^k)$ and $\tilde{F}^* \in \Re$. Let x_*^k be the projection of x^k onto X^*. By the convexity of F, we have

$$F(x^k+\bar{d}^k) \ \leq \ F(x_*^k) + (s^k)^T(x^k+\bar{d}^k-x_*^k). \tag{48}$$

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 (46) and the assumption that $\{\mathbf{E}_{\xi_{k-1}}[\mathrm{dist}(x^k, X^*)]\}_{\mathcal{K}}$ is bounded, implies that

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

Using this relation, (47) and (48), we obtain that

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

which together with $\tilde{F}^* \ge F^*$ yields $\tilde{F}^* = F^*$. Statement (i) follows from this relation and (47).

(ii) Statement (ii) immediately follows from statement (i), the Markov inequality, and the fact $F(x^k) \;\ge\; F^*$. ∎

In the rest of this section we study the rate of convergence of a monotone version of RNBPG, i.e., $M = 0$, or equivalently, (5) is replaced by

$$F(x^k + d^k) \;\le\; F(x^k) - \frac{\sigma}{2}\|d^k\|^2. \tag{49}$$

The following lemma will be subsequently used to establish a sublinear rate of convergence of RNBPG with $M = 0$.

Lemma 3.2 *Suppose that a nonnegtive sequence $\{\Delta_k\}$ satisfies*

$$\Delta_k \;\le\; \Delta_{k-1} - \alpha\Delta_k^2 \qquad \forall k \ge 1 \tag{50}$$

for some $\alpha > 0$. Then

$$\Delta_k \le \frac{\max\{2/\alpha, \Delta_0\}}{k+1} \qquad \forall k \ge 0.$$

Proof. We divide the proof into two cases.

Case (i): Suppose $\Delta_k > 0$ for all $k \ge 0$. Let $\bar{\Delta}_k = 1/\Delta_k$. It follows from (50) that

$$\bar{\Delta}_k^2 - \bar{\Delta}_{k-1}\bar{\Delta}_k - \alpha\bar{\Delta}_{k-1} \ge 0 \qquad \forall k \ge 1,$$

which together with $\bar{\Delta}_k > 0$ implies that

$$\bar{\Delta}_k \ge \frac{\bar{\Delta}_{k-1} + \sqrt{\bar{\Delta}_{k-1}^2 + 4\alpha\bar{\Delta}_{k-1}}}{2}. \tag{51}$$

We next show by induction that

$$\bar{\Delta}_k \ge \beta(k+1) \qquad \forall k \ge 0, \tag{52}$$

where $\beta = \min\left\{\alpha/2, \bar{\Delta}_0\right\}$. By the definition of β, one can see that (52) holds for $k = 0$. Suppose it holds for some $k \ge 0$. We now need to show (52) also holds for $k+1$. Indeed, since $\beta \le \alpha/2$, we have

$$\alpha(k+1) \ge \alpha\,(k/2+1) = \alpha(k+2)/2 \ge \beta(k+2).$$

which yields

$$4\alpha\beta(k+1) \geq \beta^2(4k+8) = [2\beta(k+2) - \beta(k+1)]^2 - \beta^2(k+1)^2.$$

It follows that

$$\sqrt{\beta^2(k+1)^2 + 4\alpha\beta(k+1)} \geq 2\beta(k+2) - \beta(k+1),$$

which is equivalent to

$$\beta(k+1) + \sqrt{\beta^2(k+1)^2 + 4\alpha\beta(k+1)} \geq 2\beta(k+2).$$

Using this inequality, (51) and the induction hypothesis $\bar{\Delta}_k \geq \beta(k+1)$, we obtain that

$$\bar{\Delta}_{k+1} \;\geq\; \frac{\bar{\Delta}_k + \sqrt{\bar{\Delta}_k^2 + 4\alpha\bar{\Delta}_k}}{2} \;\geq\; \frac{\beta(k+1) + \sqrt{\beta^2(k+1)^2 + 4\alpha\beta(k+1)}}{2} \;\geq\; \beta(k+2),$$

namely, (52) holds for $k+1$. Hence, the induction is completed and (52) holds for all $k \geq 0$. The conclusion of this lemma follows from (52) and the definitions of $\bar{\Delta}_k$ and β.

Case (ii) Suppose there exists some $\tilde{k}$ such that $\Delta_{\tilde{k}} = 0$. Let K be the smallest of such integers. Since $\Delta_k \geq 0$, it follows from (50) that $\Delta_k = 0$ for all $k \geq K$ and $\Delta_k > 0$ for every $0 \leq k < K$. Clearly, the conclusion of this lemma holds for $k \geq K$. And it also holds for $0 \leq k < K$ due to a similar argument as for Case (i). ∎

We next establish a sublinear rate of convergence on the expected objective values for the RNBPG method with $M = 0$ when applied to problem (1), where f and ψ are assumed to be convex. Before proceeding, we define the following quantities

$$r = \max_x \left\{ \mathrm{dist}(x, X^*) : x \in \Omega(x^0) \right\}, \tag{53}$$

$$q = \max_x \left\{ \|\nabla f(x)\| : x \in \Omega(x^0) \right\}, \tag{54}$$

where X^* denotes the set of optimal solutions of (1) and $\Omega(x^0)$ is defined in (11).

Theorem 3.3 *Let c, r, q be defined in (13), (53), (54), respectively. Assume that r and q are finite. Suppose that Ψ is L_Ψ-Lipschitz continuous in* $\mathrm{dom}(\Psi)$*, namely,*

$$|\Psi(x) - \Psi(y)| \leq L_\Psi \|x - y\| \qquad x, y \in \mathrm{dom}(\Psi) \tag{55}$$

for some $L_\phi > 0$. Let $\{x^k\}$ be generated by RNBPG with $M = 0$. Then

$$\mathbf{E}_{\xi_{k-1}}[F(x^k)] - F^* \;\leq\; \frac{\max\{2/\alpha, F(x^0) - F^*\}}{k+1} \qquad \forall k \geq 0,$$

where

$$\alpha = \frac{\sigma p_{\min}^2}{2(L_\Psi + q + cr)^2}. \tag{56}$$

Proof. Let $\bar{x}^k$ be defined in (6). For each x^k, let $x_*^k \in X^*$ such that $\|x^k - x_*^k\| = \mathrm{dist}(x^k, X^*)$. Due to $x^k \in \Omega(x^0)$ and (53), we know that $\|x^k - x_*^k\| \le r$. By the definition of $\bar{x}^k$ and (10), one can observe that

$$[\nabla f(x^k) + \Theta_k(\bar{x}^k - x^k)]^T(\bar{x}^k - x_*^k) + \Psi(\bar{x}^k) - \Psi(x_*^k) \le 0. \tag{57}$$

Using this inequality, (54), and (55), we have

$$\begin{aligned}
F(x^k) - F^* &= f(x^k) - f(x_*^k) + \Psi(x^k) - \Psi(\bar{x}^k) + \Psi(\bar{x}^k) - \Psi(x_*^k) \\
&\le \nabla f(x^k)^T(x^k - x_*^k) + L_\Psi\|x^k - \bar{x}^k\| + \Psi(\bar{x}^k) - \Psi(x_*^k) \\
&= \nabla f(x^k)^T(x^k - \bar{x}^k) + \nabla f(x^k)^T(\bar{x}^k - x_*^k) + L_\Psi\|x^k - \bar{x}^k\| + \Psi(\bar{x}^k) - \Psi(x_*^k) \\
&\le (L_\Psi + q)\|x^k - \bar{x}^k\| + (x^k - \bar{x}^k)^T\Theta_k(\bar{x}^k - x_*^k) \\
&\quad + \underbrace{[\nabla f(x^k) + \Theta_k(\bar{x}^k - x^k)]^T(\bar{x}^k - x_*^k) + \Psi(\bar{x}^k) - \Psi(x_*^k)}_{\le 0} \\
&\le (L_\Psi + q)\|x^k - \bar{x}^k\| + (x^k - \bar{x}^k)^T\Theta_k(\bar{x}^k - x_*^k) \\
&\le (L_\Psi + q)\|x^k - \bar{x}^k\| + \underbrace{(x^k - \bar{x}^k)^T\Theta_k(\bar{x}^k - x^k)}_{\le 0} + (x^k - \bar{x}^k)^T\Theta_k(x^k - x_*^k) \\
&\le (L_\Psi + q)\|x^k - \bar{x}^k\| + (x^k - \bar{x}^k)^T\Theta_k(x^k - x_*^k) \\
&\le (L_\Psi + q)\|x^k - \bar{x}^k\| + \|\Theta_k\|\|x^k - \bar{x}^k\|\|x^k - x_*^k\| \\
&\le (L_\Psi + q + cr)\|x^k - \bar{x}^k\| = (L_\Psi + q + cr)\|\bar{d}^k\|,
\end{aligned}$$

where the first inequality follows from convexity of f and (55), the second inequality is due to (54), the third inequality follows from (57), and the last inequality is due to $\|x^k - x_*^k\| \le r$. The preceding inequality, (15) and the fact $F(x^{k+1}) \le F(x^k)$ yield

$$\mathbf{E}_{\xi_k}[F(x^{k+1}] - F^* \le \mathbf{E}_{\xi_{k-1}}[F(x^k)] - F^* \le (L_\Psi + q + cr)\mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|] \le \frac{L_\Psi + q + cr}{p_{\min}}\mathbf{E}_{\xi_{k-1}}[\|d^k\|].$$

In addition, using $\left(\mathbf{E}_{\xi_{k-1}}[\|d^k\|]\right)^2 \le \mathbf{E}_{\xi_{k-1}}[\|d^k\|^2]$ and (49), one has

$$\mathbf{E}_{\xi_k}[F(x^{k+1})] \;\le\; \mathbf{E}_{\xi_{k-1}}[F(x^k)] - \frac{\sigma}{2}\mathbf{E}_{\xi_{k-1}}[\|d^k\|^2] \;\le\; \mathbf{E}_{\xi_{k-1}}[F(x^k)] - \frac{\sigma}{2}\left(\mathbf{E}_{\xi_{k-1}}[\|d^k\|]\right)^2.$$

Let $\Delta_k = \mathbf{E}_{\xi_{k-1}}[F(x^k)] - F^*$. Combining the preceding two inequalities, we obtain that

$$\Delta_{k+1} \;\le\; \Delta_k - \alpha\Delta_{k+1}^2 \qquad \forall k \ge 0,$$

where α is defined in (56). Notice that $\Delta_0 = F(x^0) - F^*$. Using this relation, the definition of Δ_k, and Lemma 3.2, one can see that the conclusion of this theorem holds. ∎

The next result shows that under an error bound assumption the RNBPG method with $M = 0$ is globally linearly convergent in terms of the expected objective values.

Theorem 3.4 *Let $\{x^k\}$ be generated by RNBPG with $M = 0$. Suppose that there exists $\tau > 0$ such that*

$$\mathrm{dist}(x^k, X^*) \le \tau \|\hat{g}^k\| \qquad \forall k \ge 0, \tag{58}$$

where $\hat{g}^k$ is given in (19) and X^* denotes the set of optimal solutions of (1). Then there holds

$$\mathbf{E}_{\xi_k}[F(x^k)] - F^* \le \left[\frac{2\varpi + (1-p_{\min})\sigma}{2\varpi + \sigma}\right]^k (F(x^0) - F^*) \qquad \forall k \ge 0,$$

where

$$\varpi = \frac{(c+L_f)\tau^2 c^2}{8}\left[1 + \frac{1}{\underline{\theta}} + \sqrt{1 - \frac{2}{c} + \frac{1}{\underline{\theta}^2}}\,\right]^2 + \frac{L_{\max} - \underline{\theta}}{2}.$$

Proof. For each x^k, let $x_*^k \in X^*$ such that $\|x^k - x_*^k\| = \mathrm{dist}(x^k, X^*)$. Let $\bar{d}^k$ be defined in (6), and

$$\Phi(\bar{d}^k; x^k) = f(x^k) + \nabla f(x^k)^T \bar{d}^k + \frac{1}{2}\|\bar{d}^k\|^2_{\Theta_k} + \Psi(x^k + \bar{d}^k).$$

It follows from (4) that

$$f(x+h) \ge f(x) + \nabla f(x)^T h - \frac{1}{2} L_f \|h\|^2 \qquad \forall x, h \in \Re^N.$$

Using this inequality, (10) and Lemma 2.3 (ii), we have that

$$\begin{aligned}
\Phi(\bar{d}^k; x^k) &\le f(x^k) + \nabla f(x^k)^T (x_*^k - x^k) + \tfrac{1}{2}\|x_*^k - x^k\|^2_{\Theta_k} + \Psi(x_*^k) \\
&\le f(x_*^k) + \tfrac{1}{2} L_f \|x_*^k - x^k\|^2 + \tfrac{1}{2}\|x_*^k - x^k\|^2_{\Theta_k} + \Psi(x_*^k) \\
&\le F(x_*^k) + \tfrac{1}{2}\gamma \|x_*^k - x^k\|^2 = F^* + \tfrac{1}{2}\gamma [\mathrm{dist}(x^k, X^*)]^2.
\end{aligned}$$

where $\gamma = c + L_f$. Using this relation and (58), one can obtain that

$$\Phi(\bar{d}^k; x^k) \le F^* + \frac{1}{2}\gamma\tau^2 \|\hat{g}^k\|^2.$$

It follows from this inequality and (20) that

$$\Phi(\bar{d}^k; x^k) \le F^* + \frac{1}{8}\gamma\tau^2 c^2 \left[1 + \frac{1}{\underline{\theta}} + \sqrt{1 - \frac{2}{c} + \frac{1}{\underline{\theta}^2}}\,\right]^2 \|\bar{d}^k\|^2,$$

which along with (14) yields

$$\mathbf{E}_{\xi_{k-1}}[\Phi(\bar{d}^k; x^k)] \le F^* + \frac{\gamma\tau^2 c^2}{8 p_{\min}} \left[1 + \frac{1}{\underline{\theta}} + \sqrt{1 - \frac{2}{c} + \frac{1}{\underline{\theta}^2}}\,\right]^2 \mathbf{E}_{\xi_k}[\|d^k\|^2]. \tag{59}$$

In addition, by (3) and the definition of $\bar{d}^{k,i}$, we have

$$F(x^k + \bar{d}^{k,i}) \leq f(x^k) + \nabla f(x^k)^T \bar{d}^{k,i} + \frac{L_i}{2}\|\bar{d}^{k,i}\|^2 + \Psi(x^k + \bar{d}^{k,i}) \qquad \forall i. \tag{60}$$

It also follows from (8) that

$$\nabla f(x^k)^T \bar{d}^{k,i} + \frac{\theta_{k,i}}{2}\|\bar{d}^{k,i}\|^2 + \Psi(x^k + \bar{d}^{k,i}) - \Psi(x^k) \leq 0 \qquad \forall i. \tag{61}$$

Using these two inequalities, we can obtain that

$$\begin{aligned}
\mathbf{E}_{i_k}[F(x^{k+1})] &= \mathbf{E}_{i_k}[F(x^k + \bar{d}^{k,i_k})] \;=\; \textstyle\sum_{i=1}^n p_i F(x^k + \bar{d}^{k,i}) \\
&\leq \textstyle\sum_{i=1}^n p_i [f(x^k) + \nabla f(x^k)^T \bar{d}^{k,i} + \frac{L_i}{2}\|\bar{d}^{k,i}\|^2 + \Psi(x^k + \bar{d}^{k,i})] \\
&= F(x^k) + \textstyle\sum_{i=1}^n p_i [\nabla f(x^k)^T \bar{d}^{k,i} + \frac{L_i}{2}\|\bar{d}^{k,i}\|^2 + \Psi(x^k + \bar{d}^{k,i}) - \Psi(x^k)] \\
&= F(x^k) + \textstyle\sum_{i=1}^n p_i \underbrace{[\nabla f(x^k)^T \bar{d}^{k,i} + \frac{\theta_{k,i}}{2}\|\bar{d}^{k,i}\|^2 + \Psi(x^k + \bar{d}^{k,i}) - \Psi(x^k)]}_{\leq 0} \\
&\quad + \tfrac{1}{2}\textstyle\sum_{i=1}^n p_i (L_i - \theta_{k,i})\|\bar{d}^{k,i}\|^2 \\
&\leq F(x^k) + p_{\min} \textstyle\sum_{i=1}^n [\nabla f(x^k)^T \bar{d}^{k,i} + \frac{\theta_{k,i}}{2}\|\bar{d}^{k,i}\|^2 + \Psi(x^k + \bar{d}^{k,i}) - \Psi(x^k)] \\
&\quad + \tfrac{1}{2}\textstyle\sum_{i=1}^n p_i (L_i - \theta_{k,i})\|\bar{d}^{k,i}\|^2 \\
&= F(x^k) + p_{\min} [\nabla f(x^k)^T \bar{d}^k + \tfrac{1}{2}\|\bar{d}^k\|^2_{\Theta_k} + \Psi(x^k + \bar{d}^k) - \Psi(x^k)] \\
&\quad + \tfrac{1}{2}\textstyle\sum_{i=1}^n p_i (L_i - \theta_{k,i})\|\bar{d}^{k,i}\|^2 \\
&\leq (1 - p_{\min}) F(x^k) + p_{\min} \Phi(\bar{d}^k; x^k) + \tfrac{L_{\max} - \underline{\theta}}{2}\, \mathbf{E}_{i_k}[\|d^k\|^2 \;\big|\; \xi_{k-1}],
\end{aligned}$$

where the first inequality follows from (60) and the second inequality is due to (61). Taking expectation with respect to ξ_{k-1} on both sides of the above inequality gives

$$\mathbf{E}_{\xi_k}[F(x^{k+1})] \leq (1 - p_{\min})\mathbf{E}_{\xi_{k-1}}[F(x^k)] + p_{\min}\mathbf{E}_{\xi_{k-1}}[\Phi(\bar{d}^k; x^k)] + \frac{L_{\max} - \underline{\theta}}{2}\, \mathbf{E}_{\xi_k}[\|d^k\|^2].$$

Using this inequality and (59), we obtain that

$$\mathbf{E}_{\xi_k}[F(x^{k+1})] \leq (1 - p_{\min})\mathbf{E}_{\xi_{k-1}}[F(x^k)] + p_{\min} F^* + \varpi \mathbf{E}_{\xi_k}[\|d^k\|^2] \qquad \forall k \geq 0,$$

where ϖ is defined above. In addition, it follows from (49) that

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

Combining these two inequalities, we obtain that

$$\mathbf{E}_{\xi_k}[F(x^{k+1})] - F^* \leq \frac{2\varpi + (1 - p_{\min})\sigma}{2\varpi + \sigma}\left(\mathbf{E}_{\xi_{k-1}}[F(x^k)] - F^*\right) \qquad \forall k \geq 0,$$

and the conclusion of this theorem immediately follows. ∎

Remark 3.5 *The error bound condition (58) holds for a class of problems, especially when f is strongly convex. More discussion about this condition can be found, for example, in [9].* ■

4 Computational results

In this section we study the numerical behavior of the RNBPG 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 [17, Section 6]. 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 RNBPG 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 [17].

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

	$N_i = 1$	$N_i = 10$	$N_i = 100$	$N_i = 1000$
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
RNBPG	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 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 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.

	$N_i = 1$	$N_i = 10$	$N_i = 100$	$N_i = 1000$
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
RNBPG	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).

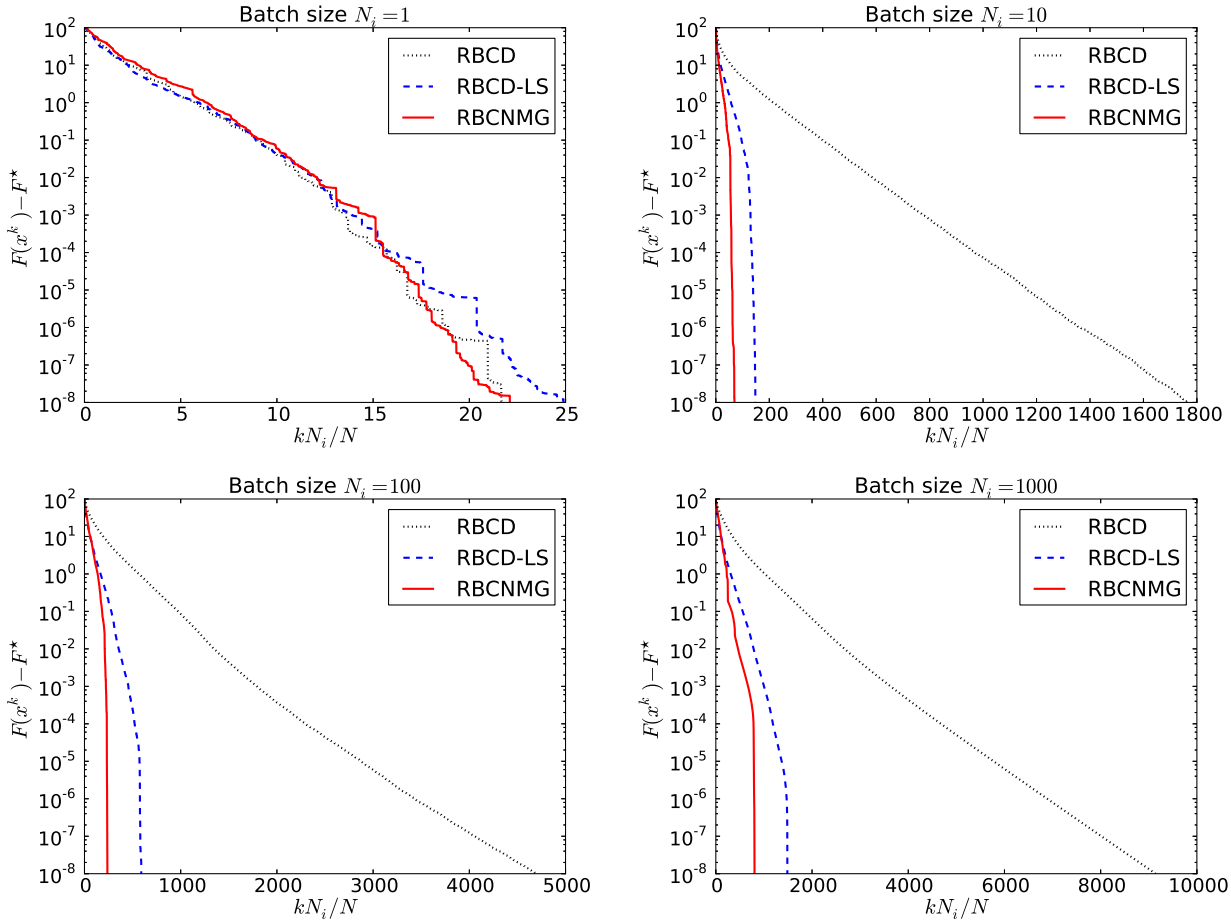

Figure 1: Comparison of different methods when the block-coordinates are chosen uniformly at random ($p_i \propto L_i^\alpha$ with $\alpha = 0$).

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:

- When the blocksize (batch size) $N_i = 1$, these three methods behave similarly. The rea-

	$N_i = 1$	$N_i = 10$	$N_i = 100$	$N_i = 1000$
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
RNBPG	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).

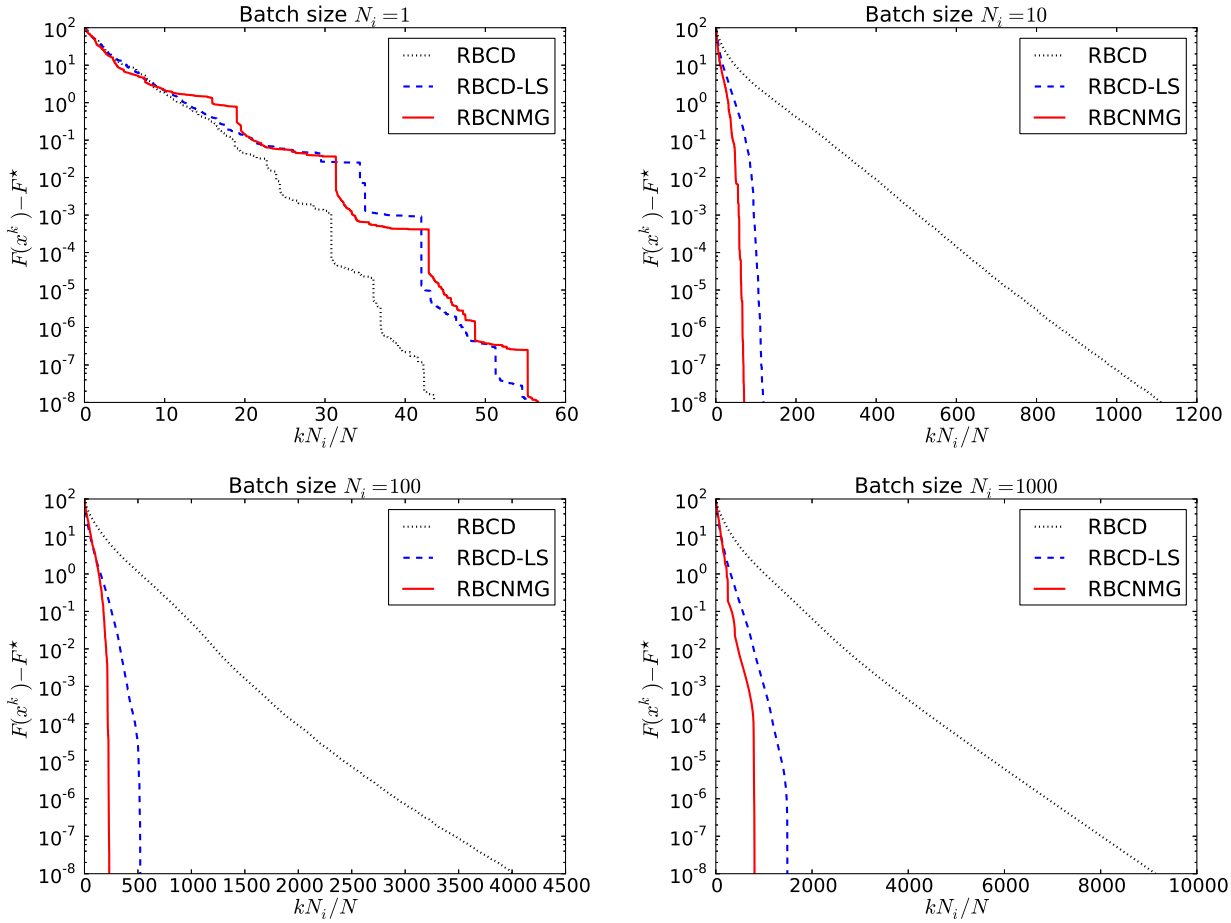

Figure 2: Comparison of different methods when the block-coordinates are chosen with probability $p_i \propto L_i^\alpha$ with $\alpha = 0.5$.

son 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

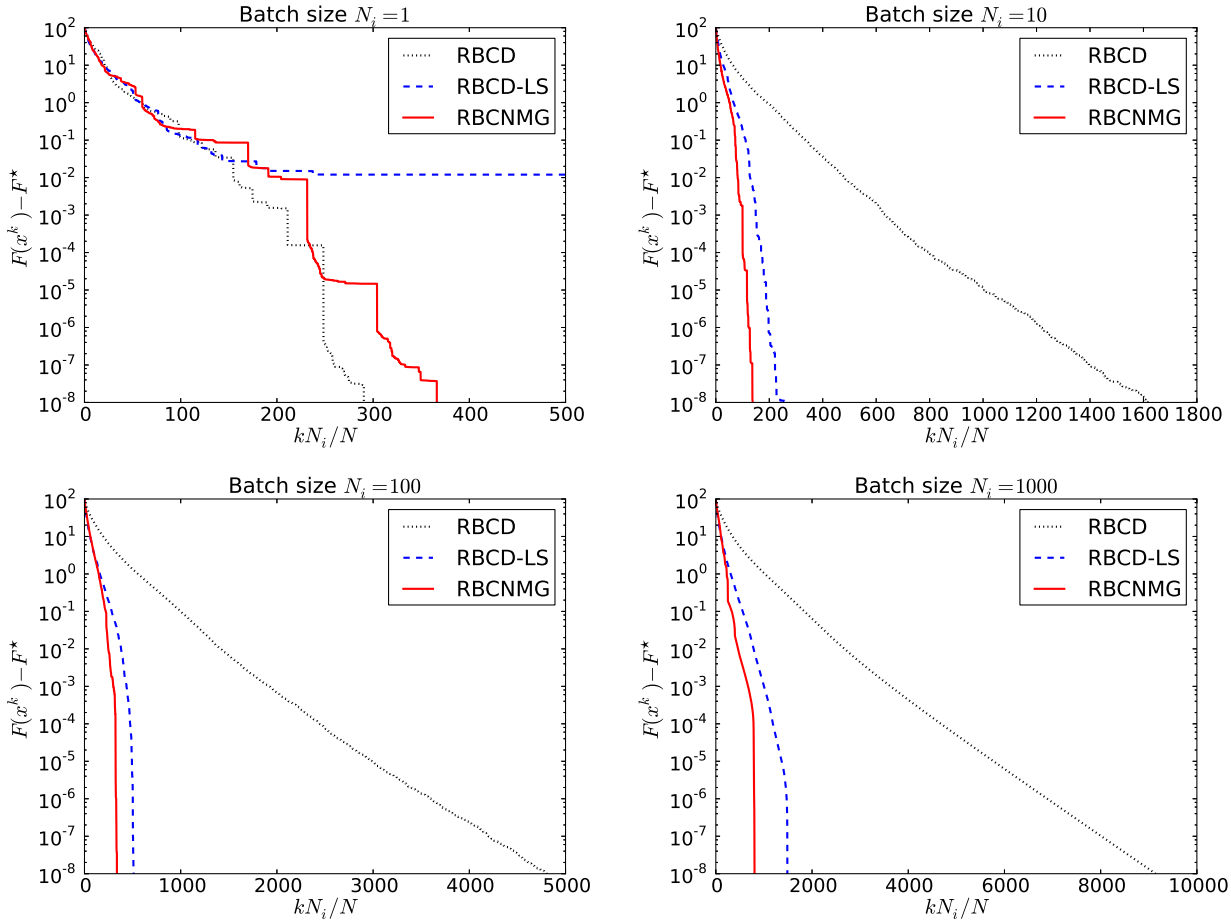

Figure 3: Comparison of different methods when the block-coordinates are chosen with probability $p_i \propto L_i^\alpha$ with $\alpha = 1$.

distributions. Nevertheless, the RNBPG method seems to perform best when the block-coordinates are chosen uniformly at random.

- Our RNBPG 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] P. Billingsley. Probability and Measure. 3rd edition, John Wiley & Sons, New York, 1995.

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

[4] 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.

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

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

[7] M. C. Ferris, S. Lucidi, and M. Roma. Nonmonotone curvilinear line search methods for unconstrained optimization. *Computational Optimization and Applications*, 6(2): 117–136, 1996.

[8] L. Grippo, F. Lampariello, and S. Lucidi. A Nonmonotone Line Search Technique for Newton's Method. *SIAM Journal on Numerical Analysis*, 23(4): 707–716, 1986.

[9] M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction methods. arXiv:1208.3922, 2012. Submitted.

[10] M. Hong, X. Wang, M. Razaviyayn and Z.-Q. Luo. Iteration complexity analysis of block coordinate descent methods. *arXiv:1310.6957*, 2013.

[11] 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.

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

[13] 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.

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

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

[16] 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.

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

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

[19] A. Patrascu and I. Necoara. Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization. arXiv:1305.4027v1. May 2013.

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

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

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

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

[24] 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.

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

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

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

[28] 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.

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

[30] 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.

[31] 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.

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

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

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

[35] 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.

[36] H. Zhang and W. Hager. A nonmonotone line search technique and its application to unconstrained optimization. *SIAM Journal on Optimzation* 14(4):1043–1056, 2004.