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

Zhaosong Lu* Lin Xiao †

March 9, 2015 (Revised: May 13, 2016; December 30, 2016)

Abstract

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 [24, 21], our method has a nonmonotone flavor and uses variable stepsizes that can partially utilize the local curvature information of the smooth component of objective function. We show that 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, a dual SVM problem in machine learning, the ℓ_0- regularized least-squares problem, and a regularized matrix completion model. The computational results demonstrate that our method substantially outperforms the randomized block coordinate *descent* method with fixed or variable stepsizes.

Key words: nonconvex composite optimization, randomized algorithms, block coordinate gradient method, nonmonotone line search.

AMS 2000 subject classification: 65K05, 90C06, 90C30

*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, 12, 37, 14, 31, 32, 34, 35, 22, 38, 26, 13, 23, 27, 25, 29]). Recently, Nesterov [19] 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 [19]. More recently, Richtárik and Takáč [24] extended Nesterov's RBCD method [19] 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 [24] 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 $\nabla_i f \in \Re^{N_i}$ 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 [24] and has recently been improved by Lu and Xiao [15]. Very recently, Patrascu and Necoara [21] 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 [24, 21] 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 variable stepsizes, for example, spectral-type stepsize [1, 3, 6, 36, 16]) 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, [8, 9, 39]), nonmonotone 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 variable stepsizes 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,*

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

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 an arbitrary 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, 36, 16]), until a certain progress on the objective value is achieved. In contrast to the usual RBCD method [24, 21], our method enjoys a nonmonotone flavor and uses variable stepsizes 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, a dual SVM problem in machine learning, the ℓ_0-regularized least-squares problem, and a regularized matrix completion model. The

[1]The convergence analysis of the RBCD method conducted in [24, 21] is only for uniform probability distribution.

computational results demonstrate that our method substantially outperforms the randomized block coordinate *descent* method with fixed or variable stepsizes.

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 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 with fixed or variable stepsizes.

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 [19] 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

$$(d^k)_{i_k} = \arg\min_s \left\{ \nabla_{i_k} f(x^k)^T s + \frac{\theta_k}{2}\|s\|^2 + \Psi_{i_k}(x_{i_k}^k + s) \right\}. \tag{5}$$

2b) If d^k satisfies

$$F(x^k + d^k) \ \leq \ \max_{[k-M]^+ \leq i \leq k} F(x^i) - \frac{\sigma}{2}\|d^k\|^2, \tag{6}$$

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

$$\bar{d}^k = \sum_{i=1}^n \bar{d}^{k,i}, \qquad \bar{x}^k = x^k + \bar{d}^k. \tag{7}$$

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

$$F(x^k + \bar{d}^{k,i}) \leq F(x^{\ell(k)}) - \frac{\sigma}{2}\|\bar{d}^{k,i}\|^2, \tag{8}$$

where

$$(\bar{d}^{k,i})_i = \arg\min_s \left\{ \nabla_i f(x^k)^T s + \frac{\theta_{k,i}}{2}\|s\|^2 + \Psi_i(x^k_i + s) \right\}, \tag{9}$$

$$\ell(k) = \arg\max_i \{F(x^i) : i = [k-M]^+, \ldots, k\} \qquad \forall k \geq 0. \tag{10}$$

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 (9), we observe that

$$\bar{d}^k = \arg\min_d \left\{ \nabla f(x^k)^T d + \frac{1}{2} d^T \Theta_k d + \Psi(x^k + d) \right\}. \tag{11}$$

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

$$\Omega(x^0) = \{x \in \Re^N : F(x) \leq F(x^0)\}, \tag{12}$$

$$L_{\max} = \max_i L_i, \qquad p_{\min} = \min_i p_i, \tag{13}$$

$$c = \max\left\{\bar{\theta}, \eta(L_{\max} + \sigma)\right\}. \tag{14}$$

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 (7). There hold*

$$\mathbf{E}_{\xi_k}[\|d^k\|^2] \geq p_{\min}\ \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|^2], \tag{15}$$

$$\mathbf{E}_{\xi_k}[\|d^k\|] \geq p_{\min}\ \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|]. \tag{16}$$

Proof. By (13) 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 (14). 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 (13) and (14), we then have

$$\tilde{\theta}_k := \theta_k/\eta > c/\eta \geq L_{\max} + \sigma \geq L_{i_k} + \sigma. \tag{17}$$

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

$$d_{i_k} = \arg\min_s \left\{ \nabla_{i_k} f(x^k)^T s + \frac{\tilde{\theta}_k}{2}\|s\|^2 + \Psi_{i_k}(x_{i_k}^k + s) \right\}. \tag{18}$$

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

$$F(x^k + d) > F(x^{\ell(k)}) - \frac{\sigma}{2}\|d\|^2. \tag{19}$$

On the other hand, using (3), (10), (17), (18) 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 (19). 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 (11) and (14), respectively, and*

$$\hat g^k = \arg\min_d \left\{ \nabla f(x^k)^T d+\frac{1}{2}\|d\|^2+\Psi(x^k+d)\right\}. \tag{20}$$

Assume that Ψ is convex. There holds

$$\|\hat g^k\| \le \frac{c}{2}\left[1+\frac{1}{\underline{\theta}}+\sqrt{1-\frac{2}{c}+\frac{1}{\underline{\theta}^2}}\,\right]\|\bar d^k\|. \tag{21}$$

Proof. The conclusion of this lemma follows from (11), (20), Lemma 2.3 (ii), and [16, Lemma 3.5] with $H=\Theta_k$, $\tilde H = I$, $Q=\Theta_k^{-1}$, $d=\bar d^k$ and $\tilde d = \hat g^k$. ∎

We note that by the definition in (14), we have $c \ge \overline{\theta} > \underline{\theta}$, which implies

$$1-\frac{2}{c}+\frac{1}{\underline{\theta}^2} = \left(1-\frac{1}{\underline{\theta}}\right)^2+\frac{2}{\underline{\theta}}-\frac{2}{c} > 0.$$

Therefore, the expression under the square root in (21) is always positive.

2.1 Convergence of expected objective value

In this subsection we show that the sequence of 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.

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 $|F(y^k) - F(z^k)| \leq C$ 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 $|F(y^k) - F(z^k)| \leq C$ 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,

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

Due to the arbitrarily 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

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

■

We next 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. One can see that when $n = 1$, the RNBPG method reduces to a (deterministic) full nonmonotone proximal gradient (NPG) method. The convergence result of the full NPG method, which can be viewed as a special case of the following result, has been established by Wright et al. [36]. A key relation used in their analysis is $k - M - 1 = \ell(k) - j$ for some $j = 1, 2, \ldots, M + 1$. Given that $\ell(k)$ is random when $n > 1$, such a relation clearly does not hold. Due to this, some of their analysis is no longer applicable to the RNBPG method. To overcome such a difficulty, we construct an auxiliary sequence $\{\tilde{d}^{\ell(k)-j}\}$ and establish its relation with the sequences $\{d^{\ell(k)-j}\}$ and $\{x^{\ell(k)} - x^{k-M-1}\}$, which enables us to prove the following main result.

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 (12). 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*

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] = \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}]. \tag{22}$$

Proof. By (6) and (10), we have

$$F(x^{k+1}) \;\le\; F(x^{\ell(k)}) - \frac{\sigma}{2}\|d^k\|^2 \qquad \forall k \ge 0. \tag{23}$$

Hence, $F(x^{k+1}) \le F(x^{\ell(k)})$, which together with (10) 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

$$\lim_{k\to\infty} F(x^{\ell(k)}) = F^*_{\xi_\infty}, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] = \tilde{F}^*. \tag{24}$$

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

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

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-j}\|] = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)-j})] = \tilde{F}^*. \tag{26}$$

Indeed, replacing k by $\ell(k)-1$ in (23), we obtain that

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

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

$$F(x^{\ell(k)}) \;\le\; F(x^{\ell(k-M-1)}) - \frac{\sigma}{2}\|d^{\ell(k)-1}\|^2 \qquad \forall k \ge M+1. \tag{27}$$

Then we have

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

Notice that

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

It follows from this relation and (28) that

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \;\le\; \mathbf{E}_{\xi_{k-M-2}}[F(x^{\ell(k-M-1)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|^2] \qquad \forall k \ge M+1. \tag{29}$$

In view of (24), (27), (29), 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

$$\lim_{k\to\infty} \|d^{\ell(k)-1}\| = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-1}\|] = 0. \tag{30}$$

One can also observe that $F(x^k) \leq F(x^0)$ and hence $\{x^k\} \subset \Omega(x^0)$. Using this fact, (24), (30), Lemma 2.5, and uniform continuity of F over $\Omega(x^0)$, we obtain that

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

Therefore, (25) and (26) 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 (23) 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

$$F(x^{\ell(k)-j}) \leq F(x^{\ell(k-M-j-1)}) - \frac{\sigma}{2}\|d^{\ell(k)-j-1}\|^2 \qquad \forall k \geq M + j + 1. \tag{31}$$

Then we obtain that

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

Notice that

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

It follows from these two relations that

$$\mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)-j})] \leq \mathbf{E}_{\xi_{k-M-j-2}}[F(x^{\ell(k-M-j-1)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_{k-1}}[\|d^{\ell(k)-j-1}\|^2], \quad \forall k \geq M+j+1. \tag{32}$$

Using (24), (31), (32), 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

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

Hence, (25) and (26) hold for $j+1$, and the proof of (25) and (26) is completed.

For all $k \geq 2M+1$, we define

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

It is not hard to observe that

$$\|\tilde{d}^{\ell(k)-j}\| \;\leq\; \|d^{\ell(k)-j}\|, \tag{33}$$

$$x^{\ell(k)} \;=\; x^{k-M-1} + \sum_{j=1}^{M+1} \tilde{d}^{\ell(k)-j}. \tag{34}$$

It follows from (25), (26) and (33) 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 (25), (26), (34), Lemma 2.5 and uniform continuity of F over $\Omega(x^0)$, imply that

$$\lim_{k\to\infty} F(x^{k-M-1}) \;=\; \lim_{k\to\infty} F(x^{\ell(k)}) = F^*_{\xi_\infty}, \tag{35}$$

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{k-M-1})] \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] = \tilde{F}^*. \tag{36}$$

It follows from (35) that $\lim_{k\to\infty} F(x^k) = F^*_{\xi_\infty}$. Using this, (23) and (24), 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 (36), 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 (23) that

$$\mathbf{E}_{\xi_k}[F(x^{k+1})] \;\leq\; \mathbf{E}_{\xi_k}[F(x^{\ell(k)})] - \frac{\sigma}{2}\mathbf{E}_{\xi_k}[\|d^k\|^2] \qquad \forall k \geq 0. \tag{37}$$

Notice that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_k}[F(x^{\ell(k)})] \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\ell(k)})] \;=\; \tilde{F}^* \;=\; \lim_{k\to\infty} \mathbf{E}_{\xi_k}[F(x^{k+1})]. \tag{38}$$

Using (37) and (38), 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.

■

2.2 Convergence to stationary points

In this subsection we show 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 (7). Assume that F is uniformly continuous and Ψ is locally Lipschitz continuous in $\Omega(x^0)$, where $\Omega(x^0)$ is defined in (12). Then there hold*

(i)

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|] = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\mathrm{dist}(-\nabla f(\bar{x}^k), \partial\Psi(\bar{x}^k)] \;=\; 0, \tag{39}$$

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

$$\mathcal{S} = \left\{x : F(x) \le F(x^0) + \max\left\{\frac{n}{\sigma}\left|L_f - \underline{\theta}\right|, 1\right\}(F(x^0) - F^*)\right\}. \tag{40}$$

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 $k \ge K$,

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

Proof. (i) We know from Theorem 2.6 (ii) that $\lim_{k\to\infty} \mathbf{E}_{\xi_k}[\|d^k\|] = 0$, which together with (16) 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 (11). By the first-order optimality condition (see, for example, Proposition 2.3.2 of [5]) of (11) and $\bar{x}^k = x^k + \bar{d}^k$, one can have

$$0 \in \nabla f(x^k) + \Theta_k \bar{d}^k + \partial\Psi(\bar{x}^k). \tag{41}$$

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 (41), 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 (39) implies that the second relation of (39) 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 (39) 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

$$\begin{aligned} F(\bar{x}^k) \;&\leq\; f(x^k) + \nabla f(x^k)^T \bar{d}^k + \frac{1}{2} L_f \|\bar{d}^k\|^2 + \Psi(x^k + \bar{d}^k) \\ &\leq\; f(x^k) + \nabla f(x^k)^T \bar{d}^k + \frac{1}{2} (\bar{d}^k)^T \Theta_k \bar{d}^k + \Psi(x^k + \bar{d}^k) + \frac{1}{2} \left(L_f - \underline{\theta}\right) \|\bar{d}^k\|^2. \end{aligned} \tag{42}$$

In view of (11), 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 (42) 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

$$|F(\bar{x}^k) - F(x^k)| \leq \max\left\{ \frac{1}{2} \left|L_f - \underline{\theta}\right| \|\bar{d}^k\|^2, F(x^0) - F^* \right\} \qquad \forall k. \tag{43}$$

In addition, since $F^{l(k)} \leq F(x^0)$ and $F(\bar{x}^k) \geq F^*$, it follows from (8) 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 (43) yields

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

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

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[|F(x^k) - F(\bar{x}^k)|] \;=\; 0. \tag{44}$$

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, (44) 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. ∎

2.3 Convergence rate analysis

In this subsection 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 (13), (20) and (14), 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

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

Proof. (i) Using $\bar{g}^k = -\Theta_k \bar{d}^k$, Lemma 2.3 (ii), and (15), one can observe that

$$\mathbf{E}_{\xi_k}[\|d^k\|^2] \geq p_{\min} \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|^2] = p_{\min} \mathbf{E}_{\xi_{k-1}}[\|\Theta_k^{-1}\bar{g}^k\|^2] \geq \frac{p_{\min}}{c^2} \mathbf{E}_{\xi_{k-1}}[\|\bar{g}^k\|^2]. \tag{45}$$

Let $j(t) = l((M+1)t) - 1$ and $\bar{j}(t) = (M+1)t - 1$ for all $t \geq 0$. One can see from (29) that

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

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

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

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

$$\min_{1\leq t\leq s}\mathbf{E}_{\xi_{\bar{j}(t)}}[\|d^{j(t)}\|^2] \leq \frac{2(F(x^0) - F^*)}{\sigma s}. \tag{46}$$

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

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

Using this relation and (46), we have

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

which together with (45) implies that statement (i) holds.

(ii) It follows from (15) and (46) that

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

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

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

Statement (ii) immediately follows from this inequality and (21). ∎

3 Convergence analysis 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 (40). 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(7). Using the assumption that F is uniformly continuous in $\mathcal{S}$ and Theorem 2.7, one has

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|\bar{d}^k\|] = 0, \qquad \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[\|s^k\|] \ = \ 0, \tag{47}$$

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k+\bar{d}^k)] = \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \tilde{F}^* \tag{48}$$

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

$$F(x^k+\bar{d}^k) \ \leq \ F(x_*^k) + (s^k)^T(x^k+\bar{d}^k-x_*^k). \tag{49}$$

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 (47) 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, (48) and (49), we obtain that

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

which together with $\tilde{F}^* \geq F^*$ yields $\tilde{F}^* = F^*$. Statement (i) follows from this relation and (48).

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

In the rest of this section we study the rate of convergence of a monotone version of RNBPG, i.e., $M = 0$, or equivalently, (6) is replaced by

$$F(x^k + d^k) \;\leq\; F(x^k) - \frac{\sigma}{2}\|d^k\|^2. \tag{50}$$

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 nonnegative sequence* $\{\Delta_k\}$ *satisfies*

$$\Delta_k \;\leq\; \Delta_{k-1} - \alpha\Delta_k^2 \qquad \forall k \geq 1 \tag{51}$$

for some $\alpha > 0$. *Then*

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

Proof. We divide the proof into two cases.
Case (i): Suppose $\Delta_k > 0$ for all $k \geq 0$. Let $\bar{\Delta}_k = 1/\Delta_k$. It follows from (51) that

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

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

$$\bar{\Delta}_k \geq \frac{\bar{\Delta}_{k-1} + \sqrt{\bar{\Delta}_{k-1}^2 + 4\alpha\bar{\Delta}_{k-1}}}{2}. \tag{52}$$

We next show by induction that

$$\bar{\Delta}_k \geq \beta(k+1) \qquad \forall k \geq 0, \tag{53}$$

where $\beta = \min\left\{\alpha/2, \bar{\Delta}_0\right\}$. By the definition of β, one can see that (53) holds for $k = 0$. Suppose it holds for some $k \geq 0$. We now need to show (53) also holds for $k+1$. Indeed, since $\beta \leq \alpha/2$, we have

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

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, (52) 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, (53) holds for $k+1$. Hence, the induction is completed and (53) holds for all $k \geq 0$. The conclusion of this lemma follows from (53) 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 (51) 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

$$r = \max_x \left\{ \mathrm{dist}(x, X^*) : x \in \Omega(x^0) \right\}, \tag{54}$$

$$q = \max_x \left\{ \|\nabla f(x)\| : x \in \Omega(x^0) \right\}, \tag{55}$$

where X^* denotes the set of optimal solutions of (1) and $\Omega(x^0)$ is defined in (12).

Theorem 3.3 *Let c, r, q be defined in (14), (54), (55), respectively. Assume that r and q are finite. Suppose that Ψ is L_Ψ-Lipschitz continuous in* $\mathrm{dom}(\Psi)$*, namely,*

$$|\Psi(x) - \Psi(y)| \leq L_\Psi \|x - y\| \qquad x, y \in \mathrm{dom}(\Psi) \tag{56}$$

for some $L_\Psi > 0$. 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

$$\alpha = \frac{\sigma p_{\min}^2}{2(L_\Psi + q + cr)^2}. \tag{57}$$

Proof. Let $\bar{x}^k$ be defined in (7). 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 (54), we know that $\|x^k - x_*^k\| \leq r$. By the definition of $\bar{x}^{k+1}$ and (11), one can observe that

$$[\nabla f(x^k) + \Theta_k(\bar{x}^{k+1} - x^k)]^T(\bar{x}^{k+1} - x_*^k) + \Psi(\bar{x}^{k+1}) - \Psi(x_*^k) \leq 0. \tag{58}$$

Using this inequality, (55), and (56), we have

$$
\begin{aligned}
&F(x^k) - F^* = f(x^k) - f(x_*^k) + \Psi(x^k) - \Psi(\bar{x}^{k+1}) + \Psi(\bar{x}^{k+1}) - \Psi(x_*^k) \\
&\le \nabla f(x^k)^T(x^k - x_*^k) + L_\Psi \|x^k - \bar{x}^{k+1}\| + \Psi(\bar{x}^{k+1}) - \Psi(x_*^k) \\
&= \nabla f(x^k)^T(x^k - \bar{x}^{k+1}) + \nabla f(x^k)^T(\bar{x}^{k+1} - x_*^k) + L_\Psi \|x^k - \bar{x}^{k+1}\| + \Psi(\bar{x}^{k+1}) - \Psi(x_*^k) \\
&\le (L_\Psi + q)\|x^k - \bar{x}^{k+1}\| + (x^k - \bar{x}^{k+1})^T \Theta_k (\bar{x}^{k+1} - x_*^k) \\
&\quad + \underbrace{[\nabla f(x^k) + \Theta_k(\bar{x}^{k+1} - x^k)]^T(\bar{x}^{k+1} - x_*^k) + \Psi(\bar{x}^{k+1}) - \Psi(x_*^k)}_{\le 0} \\
&\le (L_\Psi + q)\|x^k - \bar{x}^{k+1}\| + (x^k - \bar{x}^{k+1})^T \Theta_k (\bar{x}^{k+1} - x_*^k) \\
&\le (L_\Psi + q)\|x^k - \bar{x}^{k+1}\| + \underbrace{(x^k - \bar{x}^{k+1})^T \Theta_k (\bar{x}^{k+1} - x^k)}_{\le 0} + (x^k - \bar{x}^{k+1})^T \Theta_k (x^k - x_*^k) \\
&\le (L_\Psi + q)\|x^k - \bar{x}^{k+1}\| + (x^k - \bar{x}^{k+1})^T \Theta_k (x^k - x_*^k) \\
&\le (L_\Psi + q)\|x^k - \bar{x}^{k+1}\| + \|\Theta_k\| \|x^k - \bar{x}^{k+1}\| \|x^k - x_*^k\| \\
&\le (L_\Psi + q + cr)\|x^k - \bar{x}^{k+1}\| = (L_\Psi + q + cr)\|\bar{d}^k\|,
\end{aligned}
$$

where the first inequality follows from convexity of f and (56), the second inequality is due to (55), the third inequality follows from (58), and the last inequality is due to $\|x^k - x_*^k\| \le r$. The preceding inequality, (16) 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 (50), 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 (57). 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. Suppose that there exists $\tau > 0$ such that*

$$
\operatorname{dist}(x^k, X^*) \le \tau\|\hat{g}^k\| \qquad \forall k \ge 0, \tag{59}
$$

where $\hat{g}^k$ is given in (20) and X^* denotes the set of optimal solutions of (1). Then there holds

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

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 (7), 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

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

Using this inequality, (11) and Lemma 2.3 (ii), we have that

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

where $\gamma = c + L_f$. Using this relation and (59), one can obtain that

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

It follows from this inequality and (21) that

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

which along with (15) yields

$$\mathbf{E}_{\xi_{k-1}}[\Phi(\bar{d}^k; x^k)] \leq F^* + \frac{\gamma\tau^2 c^2}{8p_{\min}}\left[1 + \frac{1}{\underline{\theta}} + \sqrt{1 - \frac{2}{c} + \frac{1}{\underline{\theta}^2}}\ \right]^2 \mathbf{E}_{\xi_k}[\|d^k\|^2]. \tag{60}$$

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

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

It also follows from (9) that

$$\nabla f(x^k)^T \bar{d}^{k,i} + \frac{\theta_{k,i}}{2}\|\bar{d}^{k,i}\|^2 + \Psi(x^k + \bar{d}^{k,i}) - \Psi(x^k) \leq 0 \qquad \forall i. \tag{62}$$

Using these two inequalities, we can obtain that

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

where the first inequality follows from (61) and the second inequality is due to (62). 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 (60), 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 (50) 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 (59) holds for a class of problems, especially when f is strongly convex. More discussion about this condition can be found, for example, in [10].* ∎

4 Numerical experiments

In this section we illustrate the numerical behavior of the RNBPG method on the ℓ_1-regularized least-squares problem, a dual SVM problem in machine learning, the ℓ_0-regularized least-squares problem and a non-convex matrix completion problem.

First we consider the ℓ_1-regularized least-squares problem:

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

where $A \in \Re^{m \times N}$, $b \in \Re^m$, and $\lambda > 0$ is a regularization parameter. Clearly, this problem is a special case of the general model (1) with $f(x) = \|Ax - b\|_2^2/2$ and $\Psi(x) = \lambda\|x\|_1$ and thus our proposed RNBPG method can be suitably applied to solve it.

We generated a random instance with $m = 1000$ and $N = 2000$ following the procedure described in [20, Section 6]. The advantage of this procedure is that an optimal solution x^* is generated together with A and b, and hence the optimal value F^* is known. We generated an instance where the optimal solution x^* has only 200 nonzero entries, so this can be considered as a sparse recovery problem. We compare RNBPG with the following methods:

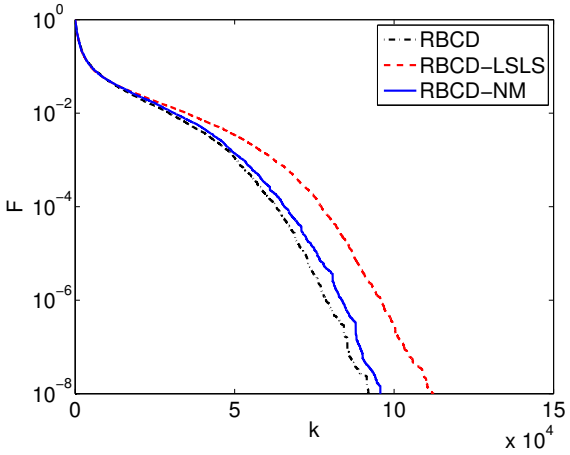
- RBCD: The RBCD method [24] with constant step sizes $1/L_i$ determined by the Lipschitz constants L_i. Here, $L_i = \|A_{:,i}\|_2^2$ where $A_{:,i}$ is the ith column block corresponding to the block partitions of x_i and $\|\cdot\|_2$ is the matrix spectral norm.

- RBCD-LS: A variant of RBCD method with variable stepsizes that are determined by a block-coordinate-wise backtracking line search scheme. This method can also be regarded as a variant of RNBPG with $M = 0$, but which has the property of monotone descent.

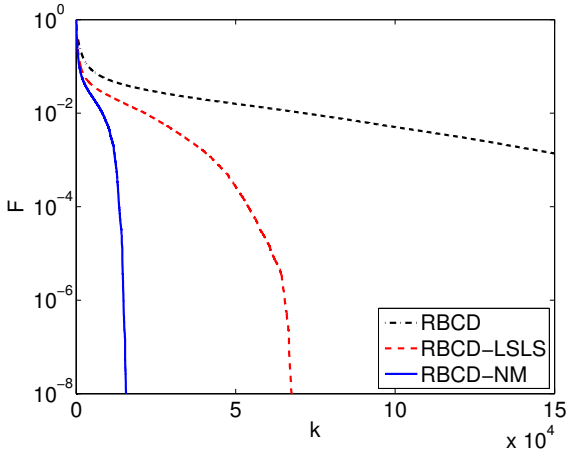
As discussed in [19], the structure of the least-squares function $f(x) = \|Ax - b\|_2^2/2$ allows efficient computation of coordinate gradients, with cost of $O(mN_i)$ operations for block i as opposed to $O(mN)$ for computing the full gradient. We note that the same structure also allows efficient computation of the function value, which costs the same order of operations as computing coordinate gradients. Therefore the backtracking line search used in RBCD-LS as well as the nonmonotone line search used in RNBPG (both relies on computing function values), have the same order computational cost as evaluating coordinate gradients at each iteration. Therefore we can focus on comparing their required number of iterations to obtain the same accuracy in reducing the objective value.

We run each algorithm with four different block coordinate sizes $N_i = 1, 20, 200, 2000$ for all i. For each blocksize, we pick the block coordinates uniformly at random at each iteration. Note that $N_i = 2000 = N$ gives the full gradient versions of the methods considered, which are deterministic algorithms. We choose the same initial point $x^0 = 0$ for all three methods.

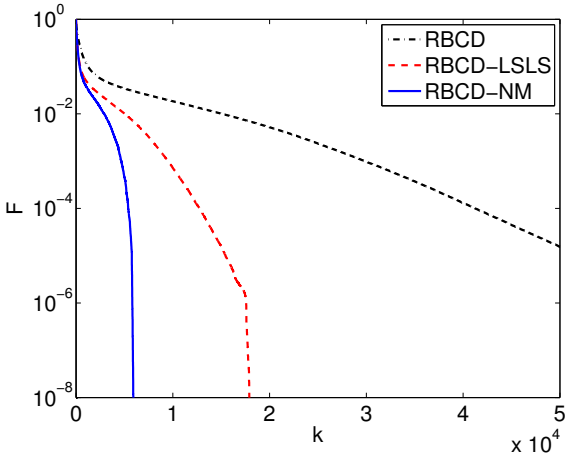
For the RNBPG method, we used the parameters $M = 10$, $\eta = 1.1$, $\underline{\theta} = 10^{-8}$, $\bar{\theta} = 10^8$ and $\sigma = 10^{-4}$. In addition, we adopted the Barzilai-Borwein spectral method [3] to compute the



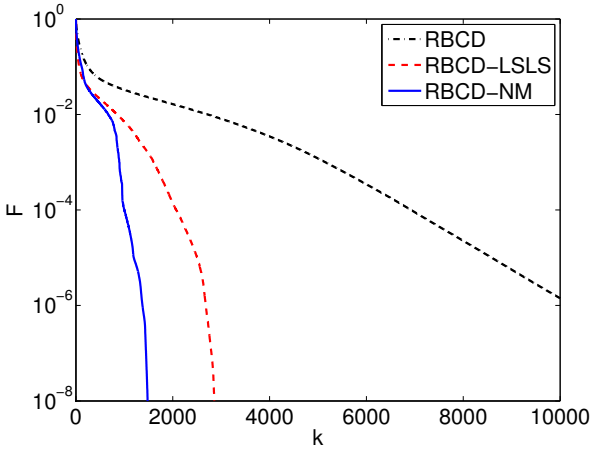
(a) Blocksize $N_i = 1$.



(b) Blocksize $N_i = 20$.



(c) Blocksize $N_i = 200$.



(d) Blocksize $N_i = 2000$.

Figure 1: Comparison of different methods with block coordinate sizes $N_i = 1, 20, 200, 2000$.

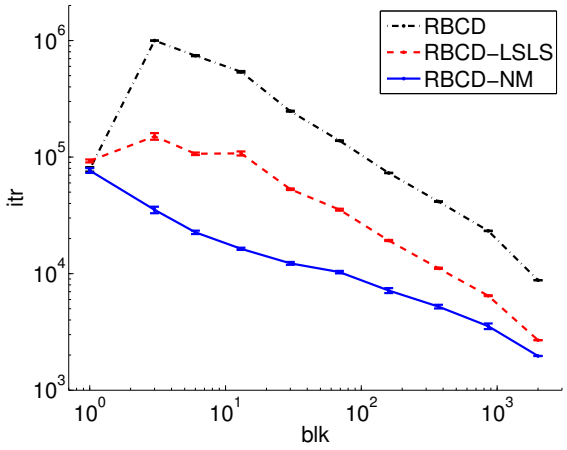
initial estimate θ_k^0 that was motivated from [1]. That is, we choose

$$\theta_k^0 = \frac{\|A_{:,i_k} u\|_2^2}{\|u\|_2^2},$$

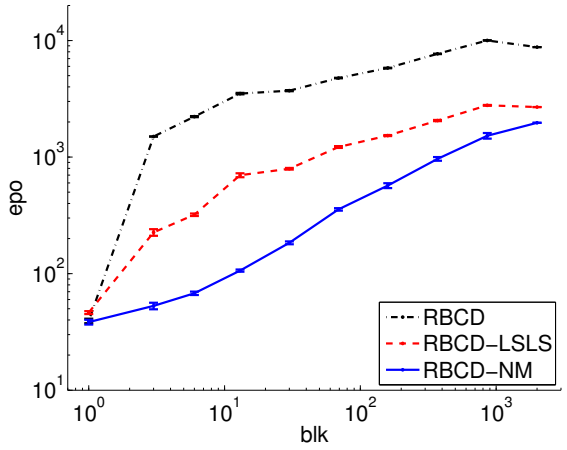
where

$$u = \arg\min_s \left\{ \nabla_{i_k} f(x^k)^T s + \frac{L_{i_k}}{2} \|s\|^2 + \Psi_{i_k}(x_{i_k}^k + s) \right\}, \quad L_{i_k} = \|A_{:,i_k}\|_2^2.$$

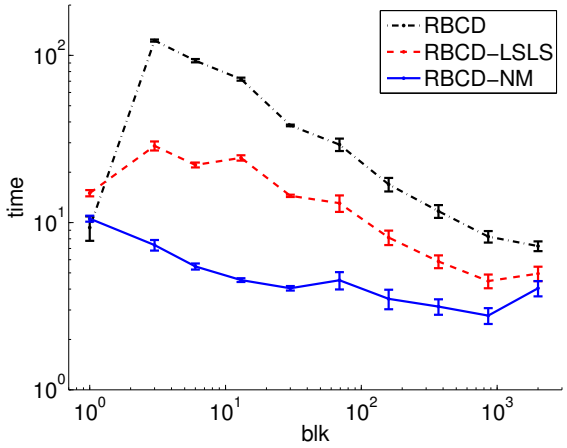
Figure 1 shows the behavior of different algorithms with the four different block coordinate sizes. For $N_i = 1$ in Figure 1(a), RBCD has slightly better convergence speed than RBCD-LS and RNBPG. The reason is that in this case, along each block f becomes an one-dimensional quadratic function, and the value $L_i = \|A_{:,i}\|_2^2$ gives the accurate second partial derivative of



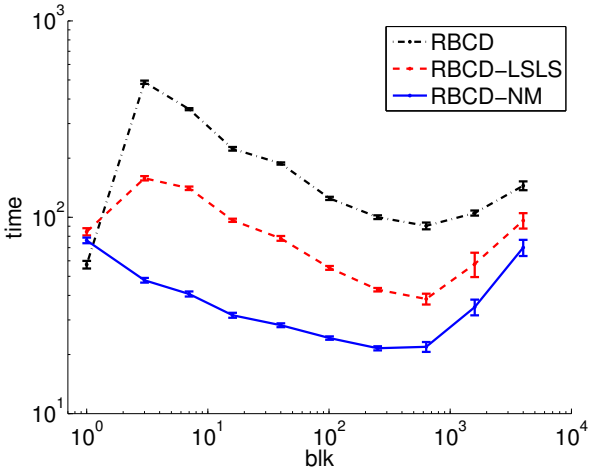
(a) Number of iterations versus blocksize.



(b) Number of epochs versus blocksize.



(c) Computation time versus blocksize.



(d) Computation time versus blocksize for a different problem instance with $m = 5000$ and $N = 1000$.

Figure 2: Comparison of different methods when varying the block coordinate size N_i.

f along each dimension. Therefore in this case the RBCD method essentially uses the best step size, which is generally better than the ones used in RBCD-LS and RNBPG.

When the blocksize N_i is larger than one, the value $L_i = \|A_{:,i}\|_2^2$ is the magnitude of second derivative along the most curved direction. Line search based methods may take advantage of the possibly much smaller local curvature along the search direction by taking larger step sizes. Figure 1 (b), (c) and (d) show that RBCD-LS converges much faster that RBCD while RNBPG (with $M = 10$) converges substantially faster than RBCD-LS.

Figure 2 shows more comprehensive study of the performance of the three methods: RBCD, RBCD-LS, and RNBPG. Figure 2(a) shows the number of iterations of different methods

	number of samples N	number of features d	sparsity	λ
RCV1	20,242	47,236	0.16%	0.0001
News20	19,996	1,355,191	0.04%	0.0001

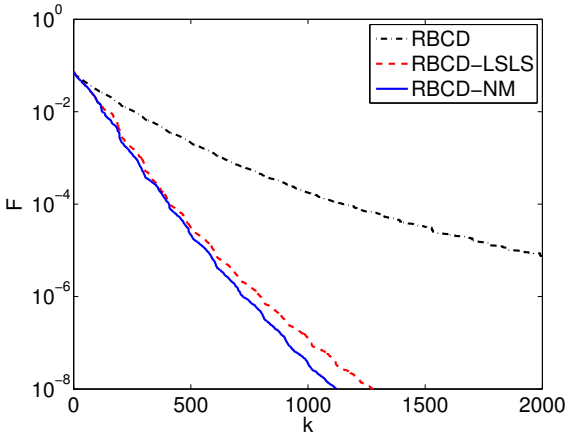
Table 1: Characteristics of two sparse datasets from the LIBSVM web site [7].

required to reach the precision $F(x^k) - F(x^*) \leq 10^{-6}$, when using 10 different block sizes ranging from 1 to 2000 with equal logarithmic spacing. Figure 2(b) shows the number of epochs required to reach the same precision, where each epoch corresponds to one equivalent pass over the dataset $A \in \Re^{m\times N}$, that is, equivalent to N/N_i iterations. For each method and each block size, we record the results of 10 runs with different random sequences to pick the block coordinates, and plot the mean with the standard deviation as error bars. As we can see, the number of iterations in general decreases when we increase the block size, because each iteration involves more coordinates and more computation. On the other hand, the number of epochs required increases with the block size, meaning that larger block size updates are less efficient than small block size updates.

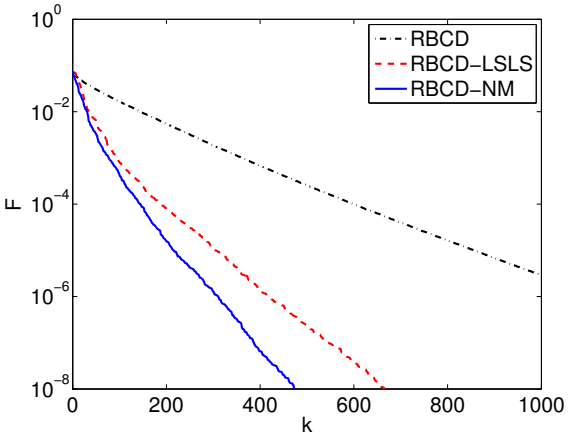
The above observations suggest that using larger block sizes is less efficient in terms of the overall computation work (e.g., measured in total flops). However, this does not mean longer computation time. In particular, using larger block sizes may better take advantage of modern multi-core computers for parallel computing, thus may take less computation time. Figure 2(c) shows the computation time required to reach the same precision on a 12 core Intel Xeon computer. We used the Intel Math Kernel Library (MKL) to carry out parallel dense matrix and vector operations. The results suggest that using appropriate large block size may take the least amount of computation time. We note that such timing results heavily depend on the specific architecture of the computer, in particular its cache size for fast access, the relative size of the data matrix A, and other implementation details. For example, Figure 2(d) shows the timing results on the same computer for a different problem instance with $m = 2000$ and $N = 4000$. Here the best block size is smaller than one shown in Figure 2(c), because the size of each column of A is doubled and the operations involved in each coordinate (corresponding to a column of the matrix) has increased. However, for any fixed block size, the relative performance of the three algorithms are consistent; in particular, RNBPG substantially outperforms the other two methods in most cases.

We also conducted experiments on using randomized block coordinate methods to solve a dual SVM problem in machine learning (specifically, the dual of a smoothed SVM problem described in [28, Section 6.2]). We used two real datasets from the LIBSVM web site [7], whose characteristics are summarized in Table 1. In the dual SVM problem, the dimension of the dual variables are the same as the number of samples N, and we partition the dual variables into blocks to apply the three randomized block coordinate gradient methods. Figure 3 shows the reduction of the objective value with the three methods on the two datasets, each illustrated with two block sizes: $N_i = 100$ and $N_i = 1000$. We observe that the RNBPG method converges faster than the other two methods, especially with relatively larger block sizes.

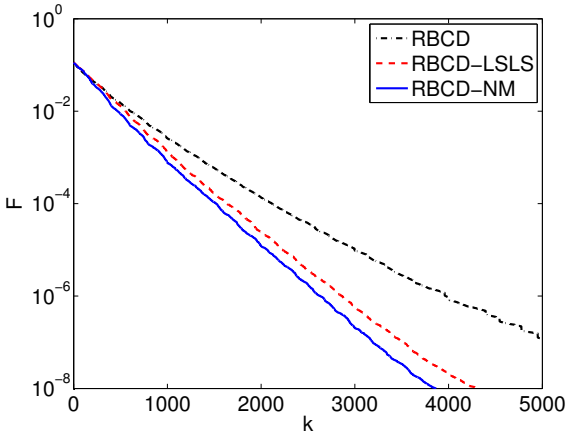
To conclude, our experiments on both synthetic and real datasets clearly demonstrate the



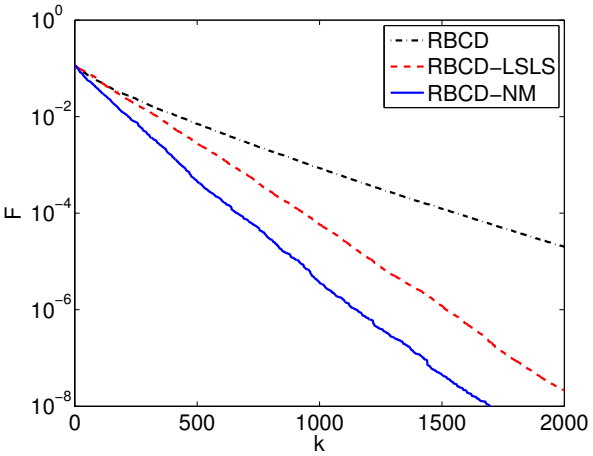
(a) RCV1 dataset with blocksize $N_i = 100$.



(b) RCV1 dataset with blocksize $N_i = 1000$.



(c) News20 dataset with blocksize $N_i = 100$.



(d) News20 dataset with blocksize $N_i = 1000$.

Figure 3: Comparison on the dual empirical risk minimization problem with real datasets.

advantage of the nonmonotone line search strategy (with spectral initialization) for randomized block coordinate gradient methods.

In the following experiment we compare the performance of RBCD and RNBPG for solving the ℓ_0-regularized least-squares problem:

$$\min_{x \in \Re^N} \ \left\{ \|Ax - b\|^2 + \lambda \|x\|_0 \right\}, \tag{63}$$

where the matrix $A \in \Re^{m \times N}$ and the vector $b \in \Re^m$ are randomly generated according to the standard normal distribution, and λ is set to 10^{-2}. We choose $x^0 = 0$ as the initial point for both methods. In addition, the decision vector x is divided sequentially into 10 blocks of equal size for both methods. At each iteration both methods choose a block uniformly

Table 2: RBCD and RNBPG for ℓ_0-regularized least squares

Problem		Objective Value		CPU Time	
m	n	RBCD	RNBPG	RBCD	RNBPG
100	500	0.38	0.29	0.01	0.01
200	1000	0.69	0.60	0.04	0.02
300	1500	1.16	0.94	0.05	0.04
400	2000	1.36	1.19	0.09	0.07
500	2500	1.92	1.54	0.13	0.11
600	3000	2.03	1.64	0.16	0.16
700	3500	2.48	1.95	0.30	0.24
800	4000	2.93	2.31	0.36	0.31
900	4500	3.22	3.06	0.41	0.58
1000	5000	3.86	2.95	0.45	0.48

at random for updating. The aforementioned parameters are used for both methods. We terminate them once the change of objective over two consecutive iterations is within 10^{-8}. The computational results are presented in Table 2. In detail, the parameters m and n of each instance are listed in the first two columns, respectively. The objective function values of (63) for both methods are given in the next two columns, and CPU times (in seconds) are given in the last two columns, respectively. One can observe that both methods are comparable in terms of CPU time, but RNBPG substantially outperforms RBCD in terms of objective value.

We next compare the performance of RBCD and RNBPG for solving a regularized nonconvex matrix completion model:

$$\min \left\{ \|\mathcal{P}_\Omega(M - U^T V)\|_F^2 + \lambda(\|U\|_F^2 + \|V\|_F^2) : U \in \Re^{r\times m}, V \in \Re^{r\times n} \right\} \tag{64}$$

for some $\lambda > 0$, where $M \in \Re^{m\times n}$, $r \leq \min(m, n)$, Ω is a subset of index pairs (i, j), and $\mathcal{P}_\Omega(\cdot)$ is the projection onto the subspace of matrices with nonzeros restricted to the index subset Ω. This model can be used to approximately recover a random matrix $M \in \Re^{m\times n}$ with rank r based on a subset of entries $\{M_{ij}\}_{(i,j)\in\Omega}$. We randomly generate M and Ω by a similar procedure as described in [18]. More specifically, we first generate random matrices $M_L \in \Re^{m\times r}$ and $M_R \in \Re^{n\times r}$ with i.i.d. standard Gaussian entries and let $M = M_L M_R^T$. We then sample a subset Ω with sampling ratio SR uniformly at random, where $SR = |\Omega|/(mn)$. In our experiment, we set $\lambda = 10^{-1}$, $m = n = 200$ and generate Ω with three different values of SR, which are 0.2, 0.5 and 0.8.

For each sample ratio SR and rank r, we apply RBCD and RNBPG to solve (64) on the instances randomly generated above. We partition the decision variable (U, V) into $m + n$ blocks and each block is a column of U or V. At each iteration both methods choose a block uniformly at random for updating. For each instance, we choose the same initial point (U^0, V^0) for both methods and terminate them once the change of objective over two consecutive iterations is within 10^{-8}. In particular, we choose U^0 and V^0 to be the matrices whose ith

Table 3: RBCD and RNBPG for problem (64) with $SR = 0.2$

r	Objective Value		Relative Error		CPU Time	
	RBCD	RNBPG	RBCD	RNBPG	RBCD	RNBPG
1	42.47	42.47	0.0034	0.0032	91.7	9.5
2	74.05	74.05	0.0039	0.0038	65.8	4.4
3	116.09	116.08	0.0052	0.0044	74.8	6.9
4	163.40	163.39	0.0045	0.0042	110.2	8.5
5	199.05	199.05	0.0050	0.0047	139.3	10.8
6	241.48	241.48	0.0055	0.0053	167.7	14.3
7	273.34	273.33	0.0064	0.0061	228.4	14.8
8	310.20	310.19	0.0069	0.0066	289.6	16.1
9	358.27	358.26	0.0082	0.0079	311.2	21.8
10	396.26	396.25	0.0094	0.0089	309.5	23.9

Table 4: RBCD and RNBPG for problem (64) with $SR = 0.5$

r	Objective Value		Relative Error		CPU Time	
	RBCD	RNBPG	RBCD	RNBPG	RBCD	RNBPG
1	42.50	42.50	0.0013	0.0012	86.4	7.3
2	76.56	76.56	0.0014	0.0014	66.2	13.0
3	114.68	114.69	0.0015	0.0014	90.3	17.7
4	163.23	163.23	0.0015	0.0014	136.0	19.5
5	200.70	200.70	0.0015	0.0014	198.8	25.2
6	268.96	268.97	0.0014	0.0014	236.1	37.8
7	282.63	282.63	0.0016	0.0015	218.8	38.8
8	313.95	313.95	0.0017	0.0017	241.7	43.4
9	342.25	342.25	0.0019	0.0018	277.8	52.8
10	384.36	384.36	0.0018	0.0018	314.8	54.6

column is $e_{\mathrm{mod}(i,r)+1}$, where e_j is the r-dimensional jth coordinate vector and mod is the standard modulo operation.

The computational results are presented in Tables 3-5. In detail, the parameter r of each instance is listed in the first column. The objective function values of (64) at the approximate solutions found by RBCD and RNBPG are given in the second and third columns. The relative errors for them are given in the next two columns, where the relative error for an approximate solution (U, V) is defined as $\|M - U^T V\|_F / \|M\|_F$. The CPU times (in seconds) of both methods are given in the last two columns. One can observe that both methods are comparable in terms of objective function value and relative error, but RNBPG substantially outperforms RBCD in terms of CPU time.

Table 5: RBCD and RNBPG for problem (64) with $SR = 0.8$

r	Objective Value		Relative Error		CPU Time	
	RBCD	RNBPG	RBCD	RNBPG	RBCD	RNBPG
1	42.52	42.51	0.0009	0.0009	86.6	25.2
2	78.47	78.48	0.0010	0.0010	74.1	17.5
3	121.91	121.91	0.0009	0.0009	107.9	23.2
4	159.63	159.63	0.0010	0.0010	125.8	31.4
5	201.15	201.15	0.0010	0.0010	186.0	40.8
6	237.14	237.14	0.0010	0.0010	225.9	55.6
7	273.38	273.38	0.0011	0.0011	234.4	64.9
8	306.68	306.68	0.0011	0.0011	268.5	72.7
9	351.22	351.22	0.0010	0.0010	330.6	85.2
10	393.81	393.80	0.0011	0.0011	375.1	84.9

5 Concluding remarks

In this paper we proposed 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. In contrast to the usual randomized block coordinate descent (RBCD) method [24, 21], our method is typically nonmonotone and moreover it uses variable stepsizes that can partially utilize the local curvature information of the smooth component of objective function. We establish the global convergence of the algorithm under suitable assumptions and also its rate of convergence. Our analysis overcame the key difficulty brought by the interplay between randomness and nonmonotonicity in the coordinate descent setting, and established convergence results under general non-degenerate probability distributions of picking the block coordinates during each iteration.

Notice that at each iteration RNBPG may need to evaluate objective values multiple times in order to determine a suitable stepsize. For many interesting problems arising in machine learning, such as training linear predictors (including SVM and logistic regression), once the partial gradient of the objective is computed, it takes little extra cost to compute the objective. Thus, the nonmonotone line search in RNBPG can be carried out quite efficiently.

Our preliminary computational results show that with nontrivial block size (for which block-wise Lipschitz constant is hard to approximate), RNBPG is consistently more efficient than RBCD in terms of both number of iterations and number of epochs, and also in computation time. Depending on the multi-core computer architecture, the RNBPG method can substantially outperform the usual RBCD methods with a suitable range of block coordinate sizes. It is worthy of a further research on how to design algorithms that can automatically tuned to have the best performance by exploiting the multi-core architecture of modern computers.

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 Journal on Optimization*, 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, Philadelphia, PA, USA, 1990.

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

[7] R.-E. Fan and C.-J. Lin. LIBSVM data: Classification, regression and multi-label. URL: http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets, 2011.

[8] 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.

[9] 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.

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

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

[12] 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.

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

[14] 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.

[15] Z. Lu and L. Xiao. On the complexity analysis of randomized block coordinate descent methods. *Mathematical Programming*, 152(1): 615–642, 2015.

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

[17] 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.

[18] S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Math. Program.*, 128(1):321-353, 2011.

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

[20] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125–161, 2013.

[21] A. Patrascu and I. Necoara. Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization. *Journal of Global Optimization*, 61(1): 19–46, 2015.

[22] Z. Qin, K. Scheinberg, and D. Goldfarb. Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation*, 5(2), 143–169, 2013.

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

[24] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

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

[26] 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.

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

[28] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14:567–599, 2013.

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

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

[31] 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.

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

[33] E. Van Den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890-912, 2008.

[34] 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.

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

[36] S. J. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Image Processing*, 57:2479–2493, 2009.

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

[38] 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.

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