Randomized block proximal damped Newton method for composite self-concordant minimization

Zhaosong Lu *

June 30, 2016

Abstract

In this paper we consider the composite self-concordant (CSC) minimization problem, which minimizes the sum of a self-concordant function f and a (possibly nonsmooth) proper closed convex function g. The CSC minimization is the cornerstone of the path-following interior point methods for solving a broad class of convex optimization problems. It has also found numerous applications in machine learning. The proximal damped Newton (PDN) methods have been well studied in the literature for solving this problem that enjoy a nice iteration complexity. Given that at each iteration these methods typically require evaluating or accessing the Hessian of f and also need to solve a proximal Newton subproblem, the cost per iteration can be prohibitively high when applied to large-scale problems. Inspired by the recent success of block coordinate descent methods, we propose a randomized block proximal damped Newton (RBPDN) method for solving the CSC minimization. Compared to the PDN methods, the computational cost per iteration of RBPDN is usually significantly lower. The computational experiment on a class of regularized logistic regression problems demonstrate that RBPDN is indeed promising in solving large-scale CSC minimization problems. The convergence of RBPDN is also analyzed in the paper. In particular, we show that RBPDN is globally convergent when g is Lipschitz continuous. It is also shown that RBPDN enjoys a local linear convergence. Moreover, we show that for a class of g including the case where g is smooth (but not necessarily self-concordant) and ∇g is Lipschitz continuous in its domain, RBPDN enjoys a global linear convergence. As a striking consequence, it shows that the classical damped Newton methods [22, 40] and the PDN [31] for such g are globally linearly convergent, which was previously unknown in the literature. Moreover, this result can be used to sharpen the existing iteration complexity of these methods.

Keywords: Composite self-concordant minimization, damped Newton method, proximal damped Newton method, randomized block proximal damped Newton method.

AMS subject classifications: 49M15, 65K05, 90C06, 90C25, 90C51

1 Introduction

In this paper we are interested in the composite self-concordant minimization:

$$F^* = \min_x \{F(x) := f(x) + g(x)\}, \tag{1.1}$$

*Department of Mathematics, Simon Fraser University, Canada (Email: zhaosong@sfu.ca). This author was supported in part by NSERC Discovery Grant.

where $f : \Re^N \to \bar{\Re} := \Re \cup \{\infty\}$ is a self-concordant function with parameter $M_f \geq 0$ and $g : \Re^N \to \bar{\Re}$ is a (possibly nonsmooth) proper closed convex function. Specifically, by the standard definition of a self-concordant function (e.g., see [25, 22]), f is convex and three times continuously differentiable in its domain denoted by $\mathrm{dom}(f)$, and moreover,

$$|\psi'''(0)| \leq M_f(\psi''(0))^{3/2}$$

holds for every $x \in \mathrm{dom}(f)$ and $u \in \Re^N$, where $\psi(t) = f(x + tu)$ for any $t \in \Re$. In addition, f is called a standard self-concordant function if $M_f = 2$.

It is well-known that problem (1.1) with $g = 0$ is the cornerstone of the path-following interior point methods for solving a broad class of convex optimization problems. Indeed, in the seminal work by Nesterov and Nemirovski [25], many convex optimization problems can be recast into the problem:

$$\min_{x\in\Omega} \langle c, x\rangle, \tag{1.2}$$

where $c \in \Re^N$, $\Omega \subseteq \Re^N$ is a closed convex set equipped with a self-concordant barrier function B, and $\langle\cdot,\cdot\rangle$ denotes the standard inner product. It has been shown that an approximate solution of problem (1.2) can be found by solving approximately a sequence of barrier problems:

$$\min_x \{f_t(x) := \langle c, x\rangle + tB(x)\},$$

where $t > 0$ is updated with a suitable scheme. Clearly, these barrier problems are a special case of (1.1) with $f = f_t$ and $g = 0$.

Recently, Tran-Dinh et al. [30] extended the aforementioned path-following scheme to solve the problem

$$\min_{x\in\Omega} g(x),$$

where g and Ω are defined as above. They showed that an approximate solution of this problem can be obtained by solving approximately a sequence of composite barrier problems:

$$\min_x tB(x) + g(x),$$

where $t > 0$ is suitably updated. These problems are also a special case of (1.1) with $f = tB$.

In addition, numerous models in machine learning are also a special case of (1.1). For example, in the context of supervised learning, each sample is recorded as (w, y), where $w \in \Re^N$ is a sample feature vector and $y \in \Re$ is usually a target response or a binary (+1 or -1) label. A loss function $\phi(x; w, y)$ is typically associated with each (w, y). Some popular loss functions include, but are not limited to:

- squared loss: $\phi(x; w, y) = (y - \langle w, x\rangle)^2$;
- logistic loss: $\phi(x; w, y) = \log(1 + \exp(-y\langle w, x\rangle))$.

A linear predictor is often estimated by solving the empirical risk minimization model:

$$\min_x \underbrace{\frac{1}{m}\sum_{i=1}^{m} \phi(x; w^i, y_i) + \frac{\mu}{2}\|x\|^2}_{\tilde{f}(x)} + g(x),$$

where m is the sample size and g is a regularizer such as ℓ_1 norm. For stability purpose, the regularization term $\mu\|x\|^2/2$, where $\mu > 0$ and $\|\cdot\|$ is the Euclidean norm, is often included to make the model strongly convex (e.g., see [40, 41]). It is easy to observe that when ϕ is the squared loss, the associated $\tilde{f}$ is self-concordant with parameter $M_{\tilde{f}} = 0$. In addition, when ϕ is the logistic loss, $y_i \in \{-1, 1\}$ for all i and $\mu > 0$, Zhang and Xiao [40, 41] showed that the associated $\tilde{f}$ is self-concordant with parameter $M_{\tilde{f}} = R/\sqrt{\mu}$, where $R = \max_i \|w^i\|$. Besides, they proved that the associated $\tilde{f}$ for a general class of loss functions ϕ is self-concordant, which includes a smoothed hinge loss.

As another example, the graphical model is often used in statistics to estimate the conditional independence of a set of random variables (e.g., see [39, 6, 9, 17]), which is in the form of:

$$\min_{X \in \mathcal{S}_{++}^N} \langle S, X \rangle - \log\det(X) + \rho \sum_{i \neq j} |X_{ij}|,$$

where $\rho > 0$, S is a sample covariance matrix, and $\mathcal{S}_{++}^N$ is the set of $N \times N$ positive definite matrices. Given that $-\log\det(X)$ is a self-concordant function in $\mathcal{S}_{++}^N$ (e.g., see [22]), it is clear to see that the graphical model is also a special case of (1.1).

When $g = 0$, problem (1.1) can be solved by a damped Newton (DN) method or a mixture of DN and Newton methods (e.g., see [22, Section 4.1.5]). To motivate our study, we now briefly review these methods for solving (1.1) with $g = 0$. In particular, given an initial point $x^0 \in \mathrm{dom}(F)$, the DN method updates the iterates according to

$$x^{k+1} = x^k + \frac{d^k}{1 + \lambda_k}, \qquad \forall k \geq 0,$$

where d^k is the Newton direction and λ_k is the local norm of d^k at x^k, which are given by:

$$d^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k), \qquad \lambda_k = \sqrt{(d^k)^T \nabla^2 f(x^k) d^k}. \tag{1.3}$$

The mixture of DN and Newton first applies DN and then switches to the standard Newton method (i.e., setting the step length to 1) once an iterate is sufficiently close to the optimal solution. The discussion in [22, Section 4.1.5] has a direct implication that both DN and the mixture of DN and Newton find an approximate solution x^k satisfying $\lambda_k \leq \epsilon$ in at most

$$O\left(F(x^0) - F^* + \log\log\epsilon^{-1}\right)$$

iterations. This complexity can be obtained by considering two phases of these methods. The first phase consists of the iterations executed by DN for generating a point lying in a certain neighborhood of the optimal solution in which the local quadratic convergence of DN or the standard Newton method is ensured to occur, while the second phase consists of the rest of the iterations. Indeed, $O\left(F(x^0) - F^*\right)$ and $O(\log\log\epsilon^{-1})$ are an estimate of the number of iterations of these two phases, respectively.

Recently, Zhang and Xiao [40, 41] proposed an inexact damped Newton (IDN) method for solving (1.1) with $g = 0$. Their method is almost identical to DN except that the search direction d^k defined in (1.3) is inexactly computed by solving approximately the linear system

$$\nabla^2 f(x^k) d = -\nabla f(x^k).$$

By controlling suitably the inexactness on d^k and considering the similar two phases as above, they showed that IDN can find an approximate solution x^k satisfying $F(x^k) - F^* \leq \epsilon$ in at most

$$O\left(F(x^0) - F^* + \log\epsilon^{-1}\right) \tag{1.4}$$

iterations.

In addition, Tran-Dinh et al. [31] recently proposed a proximal damped Newton (PDN) method and a proximal Newton method for solving (1.1). These methods are almost the same as the aforementioned DN and the mixture of DN and Newton except that d^k is chosen as the following proximal Newton direction:

$$d^k = \arg\min_d \left\{ f(x^k) + \langle \nabla f(x^k), d \rangle + \frac{1}{2} \langle d, \nabla^2 f(x^k) d \rangle + g(x^k + d) \right\}. \tag{1.5}$$

It has essentially been shown in [31, Theorems 6, 7] that the PDN and the proximal Newton method can find an approximate solution x^k satisfying $\lambda_k \le \epsilon$ in at most

$$O\left(F(x^0) - F^* + \log\log \epsilon^{-1}\right) \tag{1.6}$$

iterations, where $\lambda_k = \sqrt{(d^k)^T \nabla^2 f(x^k) d^k}$. This complexity was derived similarly as for the DN and the mixture of DN and Newton by considering the two phases mentioned above.

Besides, proximal gradient type methods and proximal Newton type methods have been proposed in the literature for solving a class of composite minimization problems in the form of (1.1) (e.g., see [1, 23, 8, 3, 12]). At each iteration, proximal gradient type methods require the gradient of f while proximal Newton type methods need to access the Hessian of f or its approximation. Though the proximal Newton type methods [3, 12] are applicable to solve (1.1), they typically require a linear search procedure to determine a suitable step length, which may be expensive for solving large-scale problems. In this paper we are only interested in a line-search free method for solving problem (1.1).

It is known from [31] that PDN has a better iteration complexity than the accelerated proximal gradient methods [1, 23]. The cost per iteration of PDN is, however, generally much higher because it computes the search direction d^k according to (1.5) that involves $\nabla^2 f(x^k)$. This can bring an enormous challenge to PDN for solving large-scale problems. Inspired by the recent success of block coordinate descent methods, block proximal gradient methods and block quasi-Newton type methods (e.g., see [2, 5, 7, 11, 13, 14, 15, 16, 19, 20, 24, 26, 27, 28, 29, 32, 34, 35]) for solving large-scale problems, we propose a randomized block proximal damped Newton (RBPDN) method for solving (1.1) with

$$g(x) = \sum_{i=1}^n g_i(x_i), \tag{1.7}$$

where each x_i denotes a subvector of x with dimension N_i, $\{x_i : i = 1, \ldots, n\}$ form a partition of the components of x, and each $g_i : \Re^{N_i} \to \bar{\Re}$ is a proper closed convex function. Briefly speaking, suppose that $p_1, \ldots, p_n > 0$ are a set of probabilities such that $\sum_i p_i = 1$. Given a current iterate x^k, we randomly choose $\iota \in \{1, \ldots, n\}$ with probability p_ι. The next iterate x^{k+1} is obtained by setting $x_j^{k+1} = x_j^k$ for $j \ne \iota$ and

$$x_\iota^{k+1} = x_\iota^k + \frac{d_\iota(x^k)}{1 + \lambda_\iota(x^k)},$$

where $d_\iota(x^k)$ is an approximate solution to the subproblem

$$\min_{d_\iota} \left\{ f(x^k) + \langle \nabla_\iota f(x^k), d_\iota \rangle + \frac{1}{2} \langle d_\iota, \nabla^2_{\iota\iota} f(x^k), d_\iota \rangle + g_\iota(x_\iota^k + d_\iota) \right\}, \tag{1.8}$$

$\lambda_\iota(x^k) = \sqrt{\langle d_\iota(x^k), \nabla^2_{\iota\iota} f(x^k) d_\iota(x^k) \rangle}$, and $\nabla_\iota f(x^k)$ and $\nabla^2_{\iota\iota} f(x^k)$ are respectively the subvector and the submatrix of $\nabla f(x^k)$ and $\nabla^2 f(x^k)$ corresponding to x_ι.

In contrast with the (full) PDN [31], the cost per iteration of RBPDN can be considerably lower because: (i) only the *submatrix* $\nabla^2_{\iota\iota} f(x^k)$ rather than the *full* $\nabla^2 f(x^k)$ needs to be accessed and/or evaluated; and (ii) the dimension of subproblem (1.8) is much smaller than that of (1.5) and thus the computational cost for solving (1.8) can also be substantially lower. In addition, compared to the randomized block accelerated proximal gradient (RBAPG) method [7, 15], RBPDN utilizes the entire curvature information in the random subspace (i.e., $\nabla^2_{\iota\iota} f(x^k)$) while RBAPG only uses the partial curvature information, particularly, the extreme eigenvalues of $\nabla^2_{\iota\iota} f(x^k)$. It is thus expected that RBPDN takes less number of iterations than RBAPG for finding an approximate solution of similar quality, which is indeed demonstrated in our numerical experiments. Overall, RBPDN can be much faster than RBAPG, provided that the subproblem (1.8) is efficiently solved.

The convergence of RBPDN is analyzed in this paper. In particular, we show that when g is Lipschitz continuous in

$$\mathcal{S}(x^0) := \{x : F(x) \leq F(x^0)\}, \tag{1.9}$$

RBPDN is globally convergent, that is, $\mathbf{E}[F(x^k)] \to F^*$ as $k \to \infty$. It is also shown that RBPDN enjoys a local linear convergence. Moreover, we show that for a class of g including the case where g is smooth (but not necessarily self-concordant) and ∇g is Lipschitz continuous in $\mathcal{S}(x^0)$, RBPDN enjoys a global linear convergence, that is, there exists some $q \in (0, 1)$ such that

$$\mathbf{E}[F(x^k) - F^*] \leq q^k (F(x^0) - F^*), \qquad \forall k \geq 0,$$

Notice that the DN [22] and PDN [31] are a special case of RBPDN with $n = 1$. As a striking consequence, it follows that they are globally linearly convergent for such g, which was previously unknown in the literature. Moreover, this result can be used to sharpen the existing iteration complexity of the first phase of DN [22], IDN [40], PDN [31], the proximal Newton method [31] and the mixture of DN and Newton [22].

The rest of this paper is organized as follows. In Subsection 1.1, we present some assumption, notation and also some known facts. We propose in Section 2 a RBPDN method for solving problem (1.1) in which g is in the form of (1.7). In Section 3, we provide some technical preliminaries. The convergence analysis of RBPDN is given in Section 4. Numerical results are presented in Section 5.

1.1 Assumption, notation and facts

Throughout this paper, we make the following assumption for problem (1.1).

Assumption 1 *(i) f is a standard self-concordant function*[1] *and g is in the form of* (1.7).

(ii) $\nabla^2 f$ is continuous and positive definite in the domain of F.

(iii) Problem (1.1) *has a unique optimal solution x^*.*

Let $\Re^N$ denote the Euclidean space of dimension N that is equipped with the standard inner product $\langle \cdot, \cdot \rangle$. For every $x \in \Re^N$, let x_i denote a subvector of x with dimension N_i, where $\{x_i : i = 1, \ldots, n\}$ form a particular partition of the components of x.

$\| \cdot \|$ denotes the Euclidean norm of a vector or the spectral norm of a matrix. The local norm and its dual norm at any $x \in \mathrm{dom}(f)$ are given by

$$\|u\|_x := \sqrt{\langle u, \nabla^2 f(x) u \rangle}, \quad \|v\|_x^* := \sqrt{\langle v, (\nabla^2 f(x))^{-1} v \rangle}, \qquad \forall u, v \in \Re^N.$$

It is easy to see that

$$|\langle u, v \rangle| \leq \|u\|_x \cdot \|v\|_x^*, \qquad \forall u, v \in \Re^N. \tag{1.10}$$

[1] It follows from [22, Corollary 4.1.2] that if f is self-concordant with parameter M_f, then $\frac{M_f^2}{4} f$ is a standard self-concordant function. Therefore, problem (1.1) can be rescaled into an equivalent problem for which Assumption 1 (i) holds.

For any $i \in \{1, \ldots, n\}$, let $\nabla^2_{ii} f(x)$ denote the submatrix of $\nabla^2 f(x)$ corresponding to the subvector x_i. The local norm and its dual norm of x restricted to the subspace of x_i are defined as

$$\|y\|_{x_i} := \sqrt{\langle y, \nabla^2_{ii} f(x) y\rangle}, \quad \|z\|^*_{x_i} := \sqrt{\langle z, (\nabla^2_{ii} f(x))^{-1} z\rangle}, \qquad \forall y, z \in \Re^{N_i}. \tag{1.11}$$

In addition, for any symmetric positive definite matrix M, the weighted norm and its dual norm associated with M are defined as

$$\|u\|_M := \sqrt{\langle u, Mu\rangle}, \qquad \|v\|^*_M := \sqrt{\langle v, M^{-1} v\rangle}. \tag{1.12}$$

It is clear that

$$|\langle u, v\rangle| \leq \|u\|_M \cdot \|v\|^*_M. \tag{1.13}$$

The following two functions have played a crucial role in studying some properties of a standard self-concordant function (e.g., see [22]):

$$\omega(t) = t - \ln(1+t), \qquad \omega_*(t) = -t - \ln(1-t). \tag{1.14}$$

It is not hard to observe that $\omega(t) \geq 0$ for all $t > -1$ and $\omega_*(t) \geq 0$ for every $t < 1$, and moreover, ω and ω_* are strictly increasing in $[0, \infty)$ and $[0, 1)$, respectively. In addition, they are conjugate of each other, which implies that for any $t \geq 0$ and $\tau \in [0, 1)$,

$$\omega(t) = t\omega'(t) - \omega_*(\omega'(t)), \qquad \omega(t) + \omega_*(\tau) \geq \tau t \tag{1.15}$$

(e.g., see [22, Lemma 4.1.4]).

It is known from [22, Theorems 4.1.7, 4.1.8]) that f satisfies:

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x\rangle + \omega(\|y - x\|_x), \quad \forall x \in \mathrm{dom}(f), \forall y; & (1.16)\\ f(y) &\leq f(x) + \langle \nabla f(x), y - x\rangle + \omega_*(\|y - x\|_x) \quad \forall x, y \in \mathrm{dom}(f),\ \|y - x\|_x < 1. & (1.17) \end{aligned}$$

2 Randomized block proximal damped Newton method

In this section we propose a randomized block proximal damped Newton (RBPDN) method for solving problem (1.1) in which g is in the form of (1.7).

RBPDN method for solving (1.1):

Choose $x^0 \in \mathrm{dom}(F)$, $\eta \in [0, 1/4]$, and $p_i > 0$ for $i = 1, \ldots, n$ such that $\sum_{i=1}^n p_i = 1$. Set $k = 0$.

1) Pick $\iota \in \{1, \ldots, n\}$ randomly with probability p_ι.

2) Find an approximate solution $d_\iota(x^k)$ to the subproblem

$$\min_{d_\iota} \left\{ f(x^k) + \langle \nabla_\iota f(x^k), d_\iota\rangle + \frac{1}{2}\langle d_\iota, \nabla^2_{\iota\iota} f(x^k), d_\iota\rangle + g_\iota(x^k_\iota + d_\iota) \right\} \tag{2.1}$$

such that

$$\begin{aligned} -v_\iota &\in \nabla_\iota f(x^k) + \nabla^2_{\iota\iota} f(x^k) d_\iota(x^k) + \partial g_\iota(x^k_\iota + d_\iota(x^k)), & (2.2)\\ \|v_\iota\|^*_{x^k_\iota} &\leq \eta \|d_\iota(x^k)\|_{x^k_\iota} & (2.3) \end{aligned}$$

for some v_ι.

3) Set $x_j^{k+1} = x_j^k$ for $j \neq \iota$, $x_\iota^{k+1} = x_\iota^k + d_\iota(x^k)/(1+\lambda_\iota(x^k))$, $k \leftarrow k+1$ and go to step 1), where $\lambda_\iota(x^k) = \sqrt{\langle d_\iota(x^k), \nabla^2_{\iota\iota} f(x^k) d_\iota(x^k)\rangle}$.

end

Remark:

(i) The constant η controls the inexactness of solving subproblem (2.1). Clearly, $d_\iota(x^k)$ is the optimal solution to (2.1) if $\eta = 0$.

(ii) For various g, the above $d_\iota(x^k)$ can be efficiently found. For example, when $g = 0$, $d_\iota(x^k)$ can be computed by conjugate gradient method. In addition, when $g = \|\cdot\|_{\ell_1}$, it can be found by numerous methods (e.g., see [1, 23, 10, 33, 36, 38, 37, 21, 4, 18]).

(iii) To verify (2.3), one has to compute $\|v_\iota\|^*_{x_\iota^k}$, which can be expensive since $(\nabla^2_{\iota\iota} f(x^k))^{-1}$ is involved. Alternatively, we may replace (2.3) by a relation that can be cheaply verified and also ensures (2.3). Indeed, as seen later, the sequence $\{x^k\}$ lies in the compact set $\mathcal{S}(x^0)$ and $\nabla^2 f(x)$ is positive definite for all $x \in \mathcal{S}(x^0)$. It follows that

$$\sigma_f := \min_{x \in \mathcal{S}(x^0)} \lambda_{\min}(\nabla^2 f(x)) \tag{2.4}$$

is well-defined and positive, where $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue of the associated matrix. One can observe from (1.11) and (2.4) that

$$\|v_\iota\|^*_{x_\iota^k} = \sqrt{v_\iota^T (\nabla^2_{\iota\iota} f(x^k))^{-1} v_\iota} \leq \frac{\|v_\iota\|}{\sqrt{\sigma_f}}.$$

It follows that if $\|v_\iota\| \leq \eta\sqrt{\sigma_f}\|d_\iota(x^k)\|_{x_\iota^k}$ holds, so does (2.3). Therefore, for a cheaper computation, one can replace (2.3) by

$$\|v_\iota\| \leq \eta\sqrt{\sigma_f}\|d_\iota(x^k)\|_{x_\iota^k},$$

provided that σ_f is known or can be bounded from below.

(iv) The convergence of RBPDN will be analyzed in Section 4. In particular, we show that if g is Lipschitz continuous in $\mathcal{S}(x^0)$, then RBPDN is globally convergent. It is also shown that RBPDN enjoys a local linear convergence. Moreover, we show that for a class of g including the case where g is smooth (but not necessarily self-concordant) and ∇g is Lipschitz continuous in $\mathcal{S}(x^0)$, RBPDN enjoys a global linear convergence.

3 Technical preliminaries

In this section we establish some technical results that will be used later to study the convergence of RBPDN.

For any $x \in \mathrm{dom}(F)$, let $\hat{d}(x)$ be an inexact proximal Newton direction, which is an approximate solution of

$$\min_d \left\{ f(x) + \langle \nabla f(x), d\rangle + \frac{1}{2}\langle d, \nabla^2 f(x) d\rangle + g(x+d) \right\}$$

such that

$$-\hat{v} \in \nabla f(x) + \nabla^2 f(x)\hat{d}(x) + \partial g(x + \hat{d}(x)) \tag{3.1}$$

for some $\hat{v}$ satisfying $\|\hat{v}\|^*_x \leq \eta\|\hat{d}(x)\|_x$ with $\eta \in [0, 1/4]$.

The following theorem provides an estimate on the reduction of the objective value resulted from an inexact proximal damped Newton step.

Lemma 3.1 *Let $x \in \mathrm{dom}(F)$ and $\hat{d}(x)$ be defined above with $\eta \in [0, 1/4]$. Then*

$$F\left(x + \frac{\hat{d}}{1+\hat{\lambda}}\right) \leq F(x) - \frac{1}{2}\omega(\hat{\lambda}),$$

where $\hat{d} = \hat{d}(x)$ and $\hat{\lambda} = \|\hat{d}(x)\|_x$.

Proof. By the definition of $\hat{d}$ and $\hat{\lambda}$, one can observe that

$$\|\hat{d}\|_x/(1+\hat{\lambda}) = \hat{\lambda}/(1+\hat{\lambda}) < 1.$$

It then follows from (1.17) that

$$f\left(x + \frac{\hat{d}}{1+\hat{\lambda}}\right) \leq f(x) + \frac{1}{1+\hat{\lambda}}\langle \nabla f(x), \hat{d}\rangle + \omega_*\left(\frac{\hat{\lambda}}{1+\hat{\lambda}}\right). \tag{3.2}$$

In view of (3.1) and $\hat{d} = \hat{d}(x)$, there exists $s \in \partial g(x + \hat{d})$ such that

$$\nabla f(x) + \nabla^2 f(x)\hat{d} + \hat{v} + s = 0. \tag{3.3}$$

By the convexity of g, one has

$$g\left(x + \frac{\hat{d}}{1+\hat{\lambda}}\right) \leq \frac{g(x+\hat{d})}{1+\hat{\lambda}} + \frac{\hat{\lambda}g(x)}{1+\hat{\lambda}} \leq \frac{1}{1+\hat{\lambda}}[g(x) + \langle s, \hat{d}\rangle] + \frac{\hat{\lambda}g(x)}{1+\hat{\lambda}} = g(x) + \frac{\langle s, \hat{d}\rangle}{1+\hat{\lambda}}. \tag{3.4}$$

Summing up (3.2) and (3.4), and using (3.3), we have

$$\begin{aligned} F\left(x + \frac{\hat{d}}{1+\hat{\lambda}}\right) &\leq F(x) + \frac{1}{1+\hat{\lambda}}\langle \nabla f(x) + s, \hat{d}\rangle + \omega_*\left(\frac{\hat{\lambda}}{1+\hat{\lambda}}\right) \\ &= F(x) + \frac{1}{1+\hat{\lambda}}\langle -\nabla^2 f(x)\hat{d} - \hat{v}, \hat{d}\rangle + \omega_*\left(\frac{\hat{\lambda}}{1+\hat{\lambda}}\right) \\ &\leq F(x) - \frac{\hat{\lambda}^2}{1+\hat{\lambda}} + \frac{\hat{\lambda}}{1+\hat{\lambda}}\|v\|_x^* + \omega_*\left(\frac{\hat{\lambda}}{1+\hat{\lambda}}\right), \end{aligned} \tag{3.5}$$

where the last relation is due to the definition of $\hat{\lambda}$ and (1.13). In addition, observe from (1.14) that $\omega'(\hat{\lambda}) = \hat{\lambda}/(1+\hat{\lambda})$. It follows from this and (1.15) that

$$-\frac{\hat{\lambda}^2}{1+\hat{\lambda}} + \omega_*\left(\frac{\hat{\lambda}}{1+\hat{\lambda}}\right) = -\hat{\lambda}\omega'(\hat{\lambda}) + \omega_*\left(\omega'(\hat{\lambda})\right) = -\omega(\hat{\lambda}),$$

which along with (3.5), $\|\hat{v}\|_x^* \leq \eta\|\hat{d}\|_x$ and $\hat{\lambda} = \|\hat{d}\|_x$ implies

$$F\left(x + \frac{\hat{d}}{1+\hat{\lambda}}\right) \leq F(x) - \omega(\hat{\lambda}) + \frac{\eta\hat{\lambda}^2}{1+\hat{\lambda}}. \tag{3.6}$$

Claim that for any $\eta \in [0, 1/4]$,

$$\frac{\eta\hat{\lambda}^2}{1+\hat{\lambda}} \leq \frac{1}{2}\omega(\hat{\lambda}). \tag{3.7}$$

Indeed, let $\phi(\lambda) = \frac{1}{2}\omega(\lambda)(1+\lambda) - \eta\lambda^2$. In view of $\omega'(\lambda) = \lambda/(1+\lambda)$, (1.14) and $\eta \in [0, 1/4]$, one has that for every $\lambda \geq 0$,

$$\begin{aligned} \phi'(\lambda) &= \tfrac{1}{2}[\omega'(\lambda)(1+\lambda) + \omega(\lambda)] - 2\eta\lambda = \tfrac{1}{2}\left[\tfrac{\lambda}{1+\lambda}(1+\lambda) + \lambda - \ln(1+\lambda)\right] - 2\eta\lambda \\ &= (1-2\eta)\lambda - \tfrac{1}{2}\ln(1+\lambda) \geq \tfrac{1}{2}[\lambda - \ln(1+\lambda)] = \tfrac{1}{2}\omega(\lambda) \geq 0. \end{aligned}$$

This together with $\phi(0) = 0$ implies $\phi(\lambda) \geq 0$. Thus (3.7) holds as claimed. The conclusion of this lemma then immediately follows from (3.6) and (3.7). ∎

We next provide some lower and upper bounds on the optimality gap.

Lemma 3.2 *Let $x \in \operatorname{dom}(F)$ and $\bar{\lambda}(x)$ be defined as*

$$\bar{\lambda}(x) := \min_{s \in \partial F(x)} \|s\|_x^*. \tag{3.8}$$

Then

$$\omega(\|x - x^*\|_{x^*}) \leq F(x) - F^* \leq \omega_*(\bar{\lambda}(x)), \tag{3.9}$$

where the second inequality is valid only when $\bar{\lambda}(x) < 1$.

Proof. Since x^* is the optimal solution of problem (1.1), we have $-\nabla f(x^*) \in \partial g(x^*)$. This together with the convexity of g implies $g(x) \geq g(x^*) + \langle -\nabla f(x^*), x - x^*\rangle$. Also, by (1.16), one has

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^*\rangle + \omega(\|x - x^*\|_{x^*}).$$

Summing up these two inequalities yields the first inequality of (3.9).

Suppose $\bar{\lambda}(x) < 1$. We now prove the second inequality of (3.9). Indeed, by (1.16), one has

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \omega(\|y - x\|_x), \quad \forall y.$$

By (3.8), there exists $s \in \partial F(x)$ such that $\|s\|_x^* = \bar{\lambda}(x) < 1$. Clearly, $s - \nabla f(x) \in \partial g(x)$. In view of this and the convexity of g, we have

$$g(y) \geq g(x) + \langle s - \nabla f(x), y - x\rangle, \quad \forall y.$$

Summing up these two inequalities gives

$$F(y) \geq F(x) + \langle s, y - x\rangle + \omega(\|y - x\|_x), \quad \forall y.$$

It then follows from this, (1.10) and (1.15) that

$$\begin{aligned} F^* &= \min_y F(y) \;\geq\; \min_y \{F(x) + \langle s, y - x\rangle + \omega(\|y - x\|_x)\}, \\ &\geq \min_y \{F(x) - \|s\|_x^* \cdot \|y - x\|_x + \omega(\|y - x\|_x)\}, \\ &\geq F(x) - \omega_*(\|s\|_x^*) = F(x) - \omega_*(\bar{\lambda}(x)), \end{aligned}$$

where the last inequality uses (1.15). Thus the second inequality of (3.9) holds. ∎

For the further discussion, we denote by $\tilde{d}(x)$ and $\tilde{\lambda}(x)$ the (exact) proximal Newton direction and its local norm at $x \in \operatorname{dom}(F)$, that is,

$$\tilde{d}(x) := \arg\min_d \left\{ f(x) + \langle \nabla f(x), d\rangle + \frac{1}{2}\langle d, \nabla^2 f(x) d\rangle + g(x+d) \right\}, \tag{3.10}$$

$$\tilde{\lambda}(x) := \|\tilde{d}(x)\|_x. \tag{3.11}$$

The following result provides an estimate on the reduction of the objective value resulted from the exact proximal damped Newton step.

Lemma 3.3 *Let $x \in \mathrm{dom}(F)$, $\tilde{d}(x)$ and $\tilde{\lambda}(x)$ be defined respectively in* (3.10) *and* (3.11), *and* $\tilde{x} = x + \tilde{d}(x)/(1+\tilde{\lambda}(x))$. *Then*

$$F(\tilde{x}) \leq F(x) - \omega(\tilde{\lambda}(x)), \tag{3.12}$$

$$F(x) - F^* \geq \omega(\tilde{\lambda}(x)). \tag{3.13}$$

Proof. The relation (3.12) follows from [31, Theorem 5]. In addition, the relation (3.13) holds due to (3.12) and $F(\tilde{x}) \geq F^*$. ∎

Throughout the remainder of the paper, let $d_i(x)$ be an approximate solution of the problem

$$\min_{d_i} \left\{ f(x) + \langle \nabla_i f(x), d_i \rangle + \frac{1}{2}\langle d_i, \nabla^2_{ii} f(x), d_i \rangle + g_i(x_i + d_i) \right\}, \tag{3.14}$$

which satisfies the following conditions:

$$-v_i \in \nabla_i f(x) + \nabla^2_{ii} f(x) d_i(x) + \partial g_i(x_i + d_i(x)), \tag{3.15}$$

$$\|v_i\|^*_{x_i} \leq \eta \|d_i(x)\|_{x_i} \tag{3.16}$$

for some v_i and $\eta \in [0, 1/4]$. Define

$$d(x) := (d_1(x), \ldots, d_n(x)), \qquad v := (v_1, \ldots, v_n), \tag{3.17}$$

$$\lambda_i(x) := \|d_i(x)\|_{x_i}, \ i = 1, \ldots, n, \tag{3.18}$$

$$H(x) := \mathrm{Diag}(\nabla^2_{11} f(x), \ldots, \nabla^2_{nn} f(x)), \tag{3.19}$$

where $H(x)$ is a block diagonal matrix, whose diagonal blocks are $\nabla^2_{11} f(x), \ldots, \nabla^2_{nn} f(x)$. It then follows that

$$-(\nabla f(x) + v + H(x)d(x)) \in \partial g(x + d(x)). \tag{3.20}$$

The following result builds some relationship between $\|d(x)\|_{H(x)}$ and $\sum_{i=1}^n \lambda_i(x)$.

Lemma 3.4 *Let $x \in \mathrm{dom}(F)$, $d(x)$, $\lambda_i(x)$ and $H(x)$ be defined in* (3.17), (3.18) *and* (3.19), *respectively. Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \lambda_i(x) \leq \|d(x)\|_{H(x)} \leq \sum_{i=1}^n \lambda_i(x). \tag{3.21}$$

Proof. By (1.11), (1.12), (3.17) and (3.19), one has

$$\|d(x)\|_{H(x)} = \sqrt{\sum_{i=1}^n \left\| (\nabla^2_{ii} f(x))^{\frac{1}{2}} d_i(x) \right\|^2} \geq \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\| (\nabla^2_{ii} f(x))^{\frac{1}{2}} d_i(x) \right\| = \frac{1}{\sqrt{n}} \sum_{i=1}^n \lambda_i(x),$$

$$\sum_{i=1}^n \lambda_i(x) = \sum_{i=1}^n \left\| (\nabla^2_{ii} f(x))^{\frac{1}{2}} d_i(x) \right\| \geq \sqrt{\sum_{i=1}^n \left\| (\nabla^2_{ii} f(x))^{\frac{1}{2}} d_i(x) \right\|^2} = \|d(x)\|_{H(x)}.$$

∎

The following lemma builds some relationship between $\|d(x)\|_{H(x)}$ and $\|\tilde{d}(x)\|_x$.

Lemma 3.5 *Let $x \in \mathrm{dom}(F)$, $\tilde{d}(x)$, $d(x)$ and $H(x)$ be defined in* (3.10), (3.17) *and* (3.19), *respectively. Then*

$$\|d(x)\|_{H(x)} \leq \frac{\|\tilde{d}(x)\|_x}{1-\eta} \left((1+\eta) \|H(x)^{\frac{1}{2}} (\nabla^2 f(x))^{-\frac{1}{2}}\| + \|H(x)^{-\frac{1}{2}} (\nabla^2 f(x))^{\frac{1}{2}}\| \right), \tag{3.22}$$

$$\|\tilde{d}(x)\|_x \leq \left((1+\eta) \|H(x)^{\frac{1}{2}} (\nabla^2 f(x))^{-\frac{1}{2}}\| + \|H(x)^{-\frac{1}{2}} (\nabla^2 f(x))^{\frac{1}{2}}\| \right) \|d(x)\|_{H(x)}. \tag{3.23}$$

Proof. For convenience, let $d = d(x)$, $\tilde{d} = \tilde{d}(x)$, $H = H(x)$ and $\tilde{H} = \nabla^2 f(x)$. Then it follows from (3.20) and (3.10) that

$$
\begin{aligned}
&-(\nabla f(x) + v + Hd) \in \partial g(x+d), \\
&-(\nabla f(x) + \tilde{H}\tilde{d}) \in \partial g(x+\tilde{d}).
\end{aligned}
$$

In view of these and the monotonicity of ∂g, one has $\langle d - \tilde{d}, -v - Hd + \tilde{H}\tilde{d}\rangle \geq 0$, which together with (1.12) and (1.13) implies that

$$
\begin{aligned}
\|d\|_H^2 + \|\tilde{d}\|_{\tilde{H}}^2 &\leq \langle v, \tilde{d} - d\rangle + \langle d, (H+\tilde{H})\tilde{d}\rangle \\
&\leq \|v\|_H^*(\|d\|_H + \|\tilde{d}\|_H) + \|d\|_H \cdot \|\tilde{d}\|_{\tilde{H}} \cdot \|H^{-\frac{1}{2}}(H+\tilde{H})\tilde{H}^{-\frac{1}{2}}\|. \qquad (3.24)
\end{aligned}
$$

Notice that

$$
\|\tilde{d}\|_H \leq \|H^{\frac{1}{2}}\tilde{H}^{-\frac{1}{2}}\| \cdot \|\tilde{d}\|_{\tilde{H}}. \qquad (3.25)
$$

Let $H_i = \nabla^2_{ii} f(x)$. Observe that $\|v_i\|_{H_i}^* = \|v_i\|_{x_i}^*$ and $\|d_i\|_{H_i} = \|d_i\|_{x_i}$. These and (3.16) yield $\|v_i\|_{H_i}^* \leq \eta\|d_i\|_{H_i}$. In view of this and (3.19), one has

$$
\|v\|_H^* = \sqrt{\sum_i (\|v_i\|_{H_i}^*)^2} \leq \sqrt{\sum_i \eta^2 \|d_i\|_{H_i}^2} = \eta\|d\|_H. \qquad (3.26)
$$

It follows from this, (3.24) and (3.25) that

$$
\begin{aligned}
\|d\|_H^2 + \|\tilde{d}\|_{\tilde{H}}^2 &\leq \eta\|d\|_H \left(\|d\|_H + \|H^{\frac{1}{2}}\tilde{H}^{-\frac{1}{2}}\| \cdot \|\tilde{d}\|_{\tilde{H}}\right) + \|d\|_H \cdot \|\tilde{d}\|_{\tilde{H}} \cdot \|H^{-\frac{1}{2}}(H+\tilde{H})\tilde{H}^{-\frac{1}{2}}\|, \\
&\leq \eta\|d\|_H^2 + \left((1+\eta)\|H^{\frac{1}{2}}\tilde{H}^{-\frac{1}{2}}\| + \|H^{-\frac{1}{2}}\tilde{H}^{\frac{1}{2}}\|\right)\|d\|_H \cdot \|\tilde{d}\|_{\tilde{H}}, \qquad (3.27)
\end{aligned}
$$

where the second inequality uses the relation

$$
\|H^{-\frac{1}{2}}(H+\tilde{H})\tilde{H}^{-\frac{1}{2}}\| \leq \|H^{\frac{1}{2}}\tilde{H}^{-\frac{1}{2}}\| + \|H^{-\frac{1}{2}}\tilde{H}^{\frac{1}{2}}\|.
$$

Clearly, (3.27) is equivalent to

$$
(1-\eta)\|d\|_H^2 + \|\tilde{d}\|_{\tilde{H}}^2 \leq \left((1+\eta)\|H^{\frac{1}{2}}\tilde{H}^{-\frac{1}{2}}\| + \|H^{-\frac{1}{2}}\tilde{H}^{\frac{1}{2}}\|\right)\|d\|_H \cdot \|\tilde{d}\|_{\tilde{H}}.
$$

This, along with $d = d(x)$, $\tilde{d} = \tilde{d}(x)$, $H = H(x)$, $\tilde{H} = \nabla^2 f(x)$ and $\|\tilde{d}\|_x = \|\tilde{d}\|_{\tilde{H}}$, yields (3.22) and (3.23). ∎

The following results will be used subsequently to study the convergence of RBPDN.

Lemma 3.6 *Let $\mathcal{S}(x^0)$, σ_f, $\tilde{d}(x)$, $d(x)$, $\lambda_i(x)$ and $H(x)$ be defined in* (1.9), (2.4), (3.10), (3.17), (3.18) *and* (3.19), *respectively. Then*

(i) $\mathcal{S}(x^0)$ is a nonempty convex compact set.

(ii)

$$
\|x - x^*\| \leq 2(L_f/\sigma_f)\|\tilde{d}(x)\|, \quad \forall x \in \mathcal{S}(x^0), \qquad (3.28)
$$

where

$$
L_f = \max_{x \in \mathcal{S}(x^0)} \|\nabla^2 f(x)\|. \qquad (3.29)
$$

(iii)

$$F(x) - F^* \geq \omega\left(c_1 \sum_{i=1}^{n} \lambda_i(x)\right), \quad \forall x \in \mathcal{S}(x^0), \tag{3.30}$$

where

$$c_1 = \frac{1-\eta}{\sqrt{n} \max\limits_{x \in \mathcal{S}(x^0)} \left\{(1+\eta)\|H(x)^{\frac{1}{2}}(\nabla^2 f(x))^{-\frac{1}{2}}\| + \|H(x)^{-\frac{1}{2}}(\nabla^2 f(x))^{\frac{1}{2}}\|\right\}}. \tag{3.31}$$

(iv)

$$\|\tilde{d}(x)\| \leq \frac{1-\eta}{c_1\sqrt{n\sigma_f}}\|d(x)\|_{H(x)}, \quad \forall x \in \mathcal{S}(x^0). \tag{3.32}$$

(v)

$$\|\tilde{d}(x)\| \leq \frac{1-\eta}{c_1\sqrt{n\sigma_f}} \sum_{i=1}^{n} \lambda_i(x), \quad \forall x \in \mathcal{S}(x^0). \tag{3.33}$$

Proof. (i) Clearly, $\mathcal{S}(x^0) \neq \emptyset$ due to $x^0 \in \mathcal{S}(x^0)$. By (1.9) and the first inequality of (3.9), one can observe that $\mathcal{S}(x^0) \subseteq \left\{x : \omega(\|x - x^*\|_{x^*}) \leq F(x^0) - F^*\right\}$. This together with the strict monotonicity of ω in $[0, \infty)$ implies that $\mathcal{S}(x^0)$ is a bounded set. In addition, we know that F is a closed convex function. Hence, $\mathcal{S}(x^0)$ is closed and convex.

(ii) By Assumption 1, we know that $\nabla^2 f$ is continuous and positive definite in $\mathrm{dom}(F)$. It follows from this and the compactness of $\mathcal{S}(x^0)$ that σ_f and L_f are well-defined in (2.4) and (3.29) and moreover they are positive. For convenience, let $\tilde{d} = \tilde{d}(x)$ and $\tilde{H} = \nabla^2 f(x)$. By the optimality condition of (1.1) and (3.10), one has

$$-(\nabla f(x) + \tilde{H}\tilde{d}) \in \partial g(x + \tilde{d}), \qquad -\nabla f(x^*) \in \partial g(x^*),$$

which together with the monotonicity of ∂g yield

$$\langle x + \tilde{d} - x^*, -\nabla f(x) - \tilde{H}\tilde{d} + \nabla f(x^*)\rangle \geq 0.$$

Hence, we have that for all $x \in \mathcal{S}(x^0)$,

$$\begin{aligned}
\sigma_f\|x - x^*\|^2 &\leq \langle x - x^*, \nabla f(x) - \nabla f(x^*)\rangle \leq -\langle \tilde{d}, \nabla f(x) - \nabla f(x^*)\rangle - \langle x - x^*, \tilde{H}\tilde{d}\rangle \\
&\leq \|\nabla f(x) - \nabla f(x^*)\| \cdot \|\tilde{d}\| + \|\tilde{H}\| \cdot \|x - x^*\| \cdot \|\tilde{d}\| \leq 2L_f\|x - x^*\| \cdot \|\tilde{d}\|,
\end{aligned}$$

which immediately implies (3.28).

(iii) In view of (3.11), (3.21), (3.22) and (3.31), one can observe that

$$\tilde{\lambda}(x) = \|\tilde{d}(x)\|_x \geq c_1 \sum_{i=1}^{n} \lambda_i(x), \quad \forall x \in \mathcal{S}(x^0),$$

which, together with (3.13) and the monotonicity of ω in $[0, \infty)$, implies that (3.30) holds.

(iv) One can observe that

$$\|\tilde{d}(x)\| \leq \left\|(\nabla^2 f(x))^{-\frac{1}{2}}\right\| \cdot \|\tilde{d}(x)\|_x \leq \frac{1}{\sqrt{\sigma_f}}\|\tilde{d}(x)\|_x, \quad \forall x \in \mathcal{S}(x^0), \tag{3.34}$$

where the last inequality is due to (2.4). This, (3.23) and (3.31) lead to (3.32).

(v) The relation (3.33) follows from (3.21) and (3.32). ∎

4 Convergence results

In this section we establish some convergence results for RBPDN. In particular, we show in Subsection 4.1 that if g is Lipschitz continuous in $\mathcal{S}(x^0)$, then RBPDN is globally convergent. In Subsection 4.2, we show that RBPDN enjoys a local linear convergence. In Subsection 4.3, we show that for a class of g including the case where g is smooth (but not necessarily self-concordant) and ∇g is Lipschitz continuous in $\mathcal{S}(x^0)$, RBPDN enjoys a global linear convergence.

4.1 Global convergence

In this subsection we study the global convergence of RBPDN. To proceed, we first establish a certain reduction on the objective values over every two consecutive iterations.

Lemma 4.1 *Let $\{x^k\}$ be generated by RBPDN. Then*

$$\mathbf{E}_\iota[F(x^{k+1})] \le F(x^k) - \frac{1}{2}\omega\left(p_{\min}\sum_{i=1}^n \lambda_i(x^k)\right), \qquad k \ge 0, \tag{4.1}$$

where $\lambda_i(\cdot)$ is defined in (3.18) *and*

$$p_{\min} := \min_{1\le i\le n} p_i. \tag{4.2}$$

Proof. Recall that $\iota \in \{1,\ldots,n\}$ is randomly chosen at iteration k with probability p_ι. Since f is a standard self-concordant function, it is not hard to observe that $f(x_1^k,\ldots,x_{\iota-1}^k, z, x_{\iota+1}^k,\ldots,x_n^k)$ is also a standard self-concordant function of z. In view of this and Lemma 3.1 with F replaced by $F(x_1^k,\ldots,x_{\iota-1}^k, z, x_{\iota+1}^k,\ldots,x_n^k)$, one can obtain that

$$F(x^{k+1}) \le F(x^k) - \frac{1}{2}\omega(\lambda_\iota(x^k)). \tag{4.3}$$

Taking expectation with respect to ι and using the convexity of ω, one has

$$\begin{aligned}
\mathbf{E}_\iota[F(x^{k+1})] &\le F(x^k) - \tfrac{1}{2}\sum_{i=1}^n p_i\omega(\lambda_i(x^k)) \le F(x^k) - \tfrac{1}{2}\omega\left(\sum_{i=1}^n p_i\lambda_i(x^k)\right) \\
&\le F(x^k) - \tfrac{1}{2}\omega\left(p_{\min}\sum_{i=1}^n \lambda_i(x^k)\right),
\end{aligned}$$

where the last inequality follows from (4.2) and the monotonicity of ω in $[0,\infty)$. ∎

We next show that under a mild assumption RBPDN is globally convergent.

Theorem 4.1 *Assume that g is Lipschitz continuous in $\mathcal{S}(x^0)$. Then*

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

Proof. It follows from (4.1) that

$$\begin{aligned}
\mathbf{E}[F(x^{k+1})] &\le \mathbf{E}[F(x^k)] - \tfrac{1}{2}\mathbf{E}\left[\omega\left(p_{\min}\sum_{i=1}^n \lambda_i(x^k)\right)\right] \\
&\le \mathbf{E}[F(x^k)] - \tfrac{1}{2}\omega\left(p_{\min}\mathbf{E}\left[\sum_{i=1}^n \lambda_i(x^k)\right]\right),
\end{aligned}$$

where the last relation follows from Jensen's inequality. Hence, we have

$$0 \le \sum_k \omega\left(p_{\min}\mathbf{E}\left[\sum_{i=1}^n \lambda_i(x^k)\right]\right) \le F(x^0) - F^*. \tag{4.4}$$

Notice from (1.14) that $\omega(t) \geq 0$ for all $t \geq 0$ and $\omega(t) = 0$ if and only if $t = 0$. This and (4.4) imply that

$$\lim_{k\to\infty} \mathbf{E}\left[\sum_{i=1}^{n} \lambda_i(x^k)\right] = 0. \tag{4.5}$$

In view of $x^0 \in \mathcal{S}(x^0)$ and (4.3), one can observe that $x^k \in \mathcal{S}(x^0)$ for all $k \geq 0$. Due to the continuity of ∇f and the compactness of $\mathcal{S}(x^0)$, one can observe that f is Lipschitz continuous in $\mathcal{S}(x^0)$. This along with the assumption of Lipschitz continuity of g in $\mathcal{S}(x^0)$ implies that F is Lipschitz continuous in $\mathcal{S}(x^0)$ with some Lipschitz constant $L_F \geq 0$. Using this, (3.28) and (3.33), we obtain that for all $k \geq 0$,

$$\begin{aligned} F(x^k) &\leq F^* + L_F\|x^k - x^*\| \leq F^* + \tfrac{2L_f L_F}{\sigma_f}\|\tilde{d}(x^k)\| \\ &\leq F^* + \tfrac{2(1-\eta)L_f L_F}{c_1\sqrt{n}\sigma_f^{3/2}} \sum_{i=1}^{n} \lambda_i(x^k), \end{aligned}$$

where the last two inequalities follow from (3.28) and (3.33), respectively. This together with (4.5) and $F(x^k) \geq F^*$ implies that the conclusion holds. ∎

4.2 Local linear convergence

In this subsection we show that RBPDN enjoys a local linear convergence.

Theorem 4.2 *Let $\{x^k\}$ be generated by RBPDN. Suppose $F(x^0) \leq F^* + \omega(c_1/p_{\min})$, where c_1 and $p_{\min}$ are defined in* (3.31) *and* (4.2)*, respectively. Then*

$$\mathbf{E}[F(x^k) - F^*] \leq \left[\frac{6c_2 + p_{\min}^2(1-\theta)}{6c_2 + p_{\min}^2}\right]^k (F(x^0) - F^*), \qquad \forall k \geq 0,$$

where

$$c_2 := \left|\theta\left[\left(\frac{L_f}{\sigma_f}\right)^{3/2} \frac{2(1-\eta^2)}{c_1\sqrt{n}} - 1\right] + \left(\frac{1}{2} + \eta\right) p_{\max}\right|, \tag{4.6}$$

$$p_{\max} := \max_{1\leq i\leq n} p_i, \qquad \theta := \min_{1\leq i\leq n} \inf_{x\in\mathcal{S}(x^0)} \frac{p_i}{1+\lambda_i(x)} \in (0,1), \tag{4.7}$$

and σ_f, L_f and c_1 are defined respectively in (2.4)*,* (3.29) *and* (3.31)*.*

Proof. Let $k \geq 0$ be arbitrarily chosen. For convenience, let $x = x^k$ and $x^+ = x^{k+1}$. By the updating scheme of x^{k+1}, one can observe that $x_j^+ = x_j$ for $j \neq \iota$ and

$$x_\iota^+ = x_\iota + \frac{d_\iota(x)}{1+\lambda_\iota(x)},$$

where $\iota \in \{1, \ldots, n\}$ is randomly chosen with probability p_ι and $d_\iota(x)$ is an approximate solution to problem (3.14) that satisfies (3.15) and (3.16) for some v_ι and $\eta \in [0, 1/4]$. To prove this theorem, it suffices to show that

$$\mathbf{E}_\iota[F(x^+) - F^*] \leq \left(\frac{6c_2 + p_{\min}^2(1-\theta)}{6c_2 + p_{\min}^2}\right)(F(x) - F^*). \tag{4.8}$$

To this end, we first claim that θ is well-defined in (4.7) and moreover $\theta \in (0,1)$. Indeed, given any $i \in \{1, \ldots, n\}$, let $y \in \Re^N$ be defined as follows:

$$y_i = x_i + \frac{d_i(x)}{1+\lambda_i(x)}, \qquad y_j = x_j, \ \forall j \neq i,$$

where $\lambda_i(\cdot)$ is defined in (3.18). By a similar argument as for (4.3), one has

$$F(y) \le F(x) - \frac{1}{2}\omega(\lambda_i(x)).$$

Using this, $x \in \mathcal{S}(x^0)$, $F(y) \ge F^*$ and the monotonicity of ω^{-1}, we obtain that

$$\lambda_i(x) \le \omega^{-1}(2[F(x) - F(y)]) \le \omega^{-1}(2[F(x^0) - F^*]),$$

where ω^{-1} is the inverse function of ω when restricted to the interval $[0, \infty)$.[2] It thus follows that θ is well-defined in (4.7) and moreover $\theta \in (0, 1)$.

For convenience, let $\lambda_i = \lambda_i(x)$, $d_i = d_i(x)$ and $H_i = \nabla^2_{ii} f(x)$ for $i = 1, \ldots, n$ and $H = \mathrm{Diag}(H_1, \ldots, H_n)$. In view of $x \in \mathcal{S}(x^0)$ and (3.29), one can observe that

$$\|H\| \le \|\nabla^2 f(x)\| \le L_f,$$

which along with (3.28) and (3.32) implies

$$\begin{aligned} \|x - x^*\|_H &\le \|H\|^{1/2}\|x - x^*\| \le 2(L_f^{3/2}/\sigma_f)\|\tilde{d}(x)\|, \\ &\le 2\left(\frac{L_f}{\sigma_f}\right)^{3/2} \frac{1-\eta}{c_1\sqrt{n}}\|d\|_H. \end{aligned} \tag{4.9}$$

It follows from (3.15) that there exists $s_i \in \partial g_i(x_i + d_i)$ such that

$$\nabla_i f(x) + H_i d_i + s_i + v_i = 0, \qquad i = 1, \ldots, n, \tag{4.10}$$

which together with the definition of H and v yields

$$\nabla f(x) + Hd + s + v = 0,$$

where $s = (s_1, \ldots, s_n) \in \partial g(x + d)$.

By the convexity of f, one has

$$f(x) \le f(x^*) + \langle \nabla f(x), x - x^* \rangle.$$

In addition, by $s \in \partial g(x + d)$ and the convexity of g, one has

$$g(x + d) \le g(x^*) + \langle s, x + d - x^* \rangle.$$

Using the last three relations, (3.26) and (4.9), we can obtain that

$$\begin{aligned} f(x) + \langle \nabla f(x) + v, d \rangle + g(x + d) &\le f(x^*) + \langle \nabla f(x), x - x^* \rangle + \langle \nabla f(x) + v, d \rangle + g(x^*) \\ &\quad + \langle s, x + d - x^* \rangle \\ &= F^* + \langle \nabla f(x) + v + s, x + d - x^* \rangle - \langle v, x - x^* \rangle \\ &= F^* + \langle -Hd, x + d - x^* \rangle - \langle v, x - x^* \rangle \\ &= F^* - \langle Hd, d \rangle - \langle Hd, x - x^* \rangle - \langle v, x - x^* \rangle \\ &\le F^* - \|d\|_H^2 + \|d\|_H \cdot \|x - x^*\|_H + \|v\|_H^* \cdot \|x - x^*\|_H \\ &\le F^* + \beta\|d\|_H^2, \end{aligned} \tag{4.11}$$

where

$$\beta = \left(\frac{L_f}{\sigma_f}\right)^{3/2} \frac{2(1-\eta^2)}{c_1\sqrt{n}} - 1. \tag{4.12}$$

[2]Observe from (1.14) that ω is strictly increasing in $[0, \infty)$. Thus, its inverse function ω^{-1} is well-defined when restricted to this interval and moreover it is strictly increasing.

By (3.16) and (4.7), we have

$$-\sum_i \frac{p_i\langle v_i, d_i\rangle}{1+\lambda_i} \le \sum_i \frac{p_i}{1+\lambda_i}\|v_i\|_{H_i}^* \cdot \|d_i\|_{H_i} \le \eta \sum_i \frac{p_i}{1+\lambda_i}\|d_i\|_{H_i}^2 \le \eta\, p_{\max}\|d\|_H^2. \tag{4.13}$$

In addition, recall that $\omega_*(t) = -t - \ln(1-t)$. It thus follows that

$$\omega_*(t) = \sum_{k=2}^{\infty} \frac{t^k}{k!} \;\le\; \frac{t^2}{2}\sum_{k=0}^{\infty} t^k \;=\; \frac{t^2}{2(1-t)}, \qquad \forall t \in [0,1).$$

This inequality implies that

$$\sum_i p_i\omega_*\left(\frac{\lambda_i}{1+\lambda_i}\right) \le \sum_i \frac{p_i(\lambda_i/(1+\lambda_i))^2}{2(1-\lambda_i/(1+\lambda_i))} = \frac{1}{2}\sum_i \frac{p_i\lambda_i^2}{1+\lambda_i} \le \frac{p_{\max}}{2}\sum_i \lambda_i^2 = \frac{p_{\max}}{2}\|d\|_H^2, \tag{4.14}$$

where $p_{\max}$ is defined in (4.7).

Recall that $s_i \in \partial g_i(x_i + d_i)$. By the convexity of g_i, one has $g_i(x_i + d_i) - g_i(x_i) \le \langle s_i, d_i\rangle$. It thus follows from this and (4.10) that for $i = 1, \ldots, n$,

$$\begin{aligned}
&\langle \nabla_i f(x) + v_i, d_i\rangle + g_i(x_i + d_i) - g_i(x_i) \le \langle \nabla_i f(x) + v_i, d_i\rangle + \langle s_i, d_i\rangle \\
&= \langle \nabla_i f(x) + s_i + v_i, d_i\rangle = -\langle d_i, H_i d_i\rangle \;\le\; 0.
\end{aligned} \tag{4.15}$$

By a similar argument as for (3.2) and the definition of x^+, one has

$$f(x^+) \le f(x) + \frac{1}{1+\lambda_\iota}\langle \nabla_\iota f(x), d_\iota\rangle + \omega_*\left(\frac{\lambda_\iota}{1+\lambda_\iota}\right).$$

It also follows from the convexity of g_ι that

$$g_\iota\left(x_\iota + \frac{d_\iota}{1+\lambda_\iota}\right) - g_\iota(x_\iota) \le \frac{1}{1+\lambda_\iota}\left[g_\iota(x_\iota + d_\iota) - g_\iota(x_\iota)\right].$$

Using the last two inequalities and the definition of x^+, we have

$$\begin{aligned}
F(x^+) &= f(x^+) + g_\iota\left(x_\iota + \tfrac{d_\iota}{1+\lambda_\iota}\right) + \sum_{j\neq\iota} g_j(x_j) \\
&= f(x^+) + g(x) + g_\iota\left(x_\iota + \tfrac{d_\iota}{1+\lambda_\iota}\right) - g_\iota(x_\iota) \\
&\le f(x) + \tfrac{1}{1+\lambda_\iota}\langle \nabla_\iota f(x), d_\iota\rangle + \omega_*\left(\tfrac{\lambda_\iota}{1+\lambda_\iota}\right) + g(x) + g_\iota\left(x_\iota + \tfrac{d_\iota}{1+\lambda_\iota}\right) - g_\iota(x_\iota) \\
&= F(x) + \tfrac{1}{1+\lambda_\iota}\langle \nabla_\iota f(x), d_\iota\rangle + \omega_*\left(\tfrac{\lambda_\iota}{1+\lambda_\iota}\right) + g_\iota\left(x_\iota + \tfrac{d_\iota}{1+\lambda_\iota}\right) - g_\iota(x_\iota) \\
&\le F(x) + \tfrac{1}{1+\lambda_\iota}\langle \nabla_\iota f(x), d_\iota\rangle + \omega_*\left(\tfrac{\lambda_\iota}{1+\lambda_\iota}\right) + \tfrac{1}{1+\lambda_\iota}\left[g_\iota(x_\iota + d_\iota) - g_\iota(x_\iota)\right] \\
&= F(x) + \tfrac{1}{1+\lambda_\iota}\left[\langle \nabla_\iota f(x) + v_\iota, d_\iota\rangle + g_\iota(x_\iota + d_\iota) - g_\iota(x_\iota)\right] - \tfrac{\langle v_\iota, d_\iota\rangle}{1+\lambda_\iota} + \omega_*\left(\tfrac{\lambda_\iota}{1+\lambda_\iota}\right).
\end{aligned}$$

Taking expectation with respect to ι on both sides and using (4.7), (4.11), (4.13), (4.14) and (4.15), one has

$$
\begin{aligned}
\mathbf{E}_\iota[F(x^+)] &\le F(x) + \sum_i \frac{p_i}{1+\lambda_i} \underbrace{[\langle \nabla_i f(x) + v_i, d_i\rangle + g_i(x_i+d_i) - g_i(x_i)]}_{\le 0 \text{ due to (4.15)}} - \sum_i \frac{p_i \langle v_i, d_i\rangle}{1+\lambda_i} + \sum_i p_i \omega_*\left(\frac{\lambda_i}{1+\lambda_i}\right) \\
&\le F(x) + \theta \sum_i [\langle \nabla_i f(x) + v_i, d_i\rangle + g_i(x_i+d_i) - g_i(x_i)] - \sum_i \frac{p_i \langle v_i, d_i\rangle}{1+\lambda_i} + \sum_i p_i \omega_*\left(\frac{\lambda_i}{1+\lambda_i}\right) \\
&= F(x) + \theta [\langle \nabla f(x) + v, d\rangle + g(x+d) - g(x)] - \sum_i \frac{p_i \langle v_i, d_i\rangle}{1+\lambda_i} + \sum_i p_i \omega_*\left(\frac{\lambda_i}{1+\lambda_i}\right) \\
&= (1-\theta)F(x) + \theta [f(x) + \langle \nabla f(x) + v, d\rangle + g(x+d)] - \sum_i \frac{p_i \langle v_i, d_i\rangle}{1+\lambda_i} + \sum_i p_i \omega_*\left(\frac{\lambda_i}{1+\lambda_i}\right) \\
&\le (1-\theta)F(x) + \theta(F^* + \beta\|d\|_H^2) + \eta\, p_{\max}\|d\|_H^2 + \frac{p_{\max}}{2}\|d\|_H^2 \\
&= (1-\theta)F(x) + \theta F^* + (\theta\beta + (1/2+\eta)p_{\max})\,\|d\|_H^2 \\
&\le (1-\theta)F(x) + \theta F^* + c_2\left(\sum_i \lambda_i\right)^2, \qquad (4.16)
\end{aligned}
$$

where the last inequality is due to (4.12), (4.6) and $\|d\|_H^2 = \sum_i \lambda_i^2 \le (\sum_i \lambda_i)^2$.

One can easily observe from (4.16) that the conclusion of this theorem holds if $c_2 = 0$. We now assume $c_2 > 0$. Let $\delta^+ = F(x^+) - F^*$ and $\delta = F(x) - F^*$. It then follows from (4.16) that

$$\mathbf{E}_\iota[\delta^+] \le (1-\theta)\delta + c_2\left(\sum_i \lambda_i\right)^2,$$

which yields

$$\left(\sum_i \lambda_i\right)^2 \ge \frac{1}{c_2}\left(\mathbf{E}_\iota[\delta^+] - (1-\theta)\delta\right) \qquad (4.17)$$

By the assumption, one has $F(x) \le F(x^0) \le F^* + \omega(c_1/p_{\min})$. By this and (3.30), we have

$$\omega(c_1 \sum_i \lambda_i) \le F(x) - F^* \le \omega(c_1/p_{\min}),$$

which together with the monotonicity of ω in $[0,\infty)$ implies $p_{\min}\sum_i \lambda_i \le 1$. Observe that

$$\omega(t) = t - \ln(1+t) = \sum_{k=2}^{\infty} \frac{(-1)^k t^k}{k!} \ge \frac{t^2}{2} - \frac{t^3}{6} \ge \frac{t^2}{3}, \qquad \forall t \in [0,1].$$

This and $p_{\min}\sum_i \lambda_i \le 1$ lead to

$$\omega\left(p_{\min}\sum_i \lambda_i\right) \ge \frac{1}{3}p_{\min}^2\left(\sum_i \lambda_i\right)^2.$$

It then follows from this and (4.1) that

$$\mathbf{E}_\iota[\delta^+] \le \delta - \frac{1}{6}p_{\min}^2\left(\sum_i \lambda_i\right)^2,$$

which together with (4.17) gives

$$\mathbf{E}_\iota[\delta^+] \le \delta - \frac{p_{\min}^2}{6c_2}\left(\mathbf{E}_\iota[\delta^+] - (1-\theta)\delta\right).$$

Hence, we obtain that

$$\mathbf{E}_\iota[\delta^+] \le \left(\frac{6c_2 + p_{\min}^2(1-\theta)}{6c_2 + p_{\min}^2}\right)\delta,$$

which proves (4.8) as desired. ■

4.3 Global linear convergence

In this subsection we show that for a class of g including the case where g is smooth (but not necessarily self-concordant) and ∇g is Lipschitz continuous in $\mathcal{S}(x^0)$,[3] RBPDN enjoys a global linear convergence. To this end, we make the following assumption throughout this subsection which, as shown subsequently, holds for a class of g.

Assumption 2 *There exists some $c_3 > 0$ such that*

$$\|\tilde{d}(x)\| \geq c_3 \bar{\lambda}(x), \qquad \forall x \in \mathcal{S}(x^0),$$

where $\mathcal{S}(x^0)$, $\bar{\lambda}(x)$ and $\tilde{d}(x)$ are defined in (1.9), (3.8) *and* (3.10), *respectively.*

The following proposition shows that Assumption 2 holds for a class of g including $g = 0$ as a special case.

Proposition 4.1 *Suppose that g is Lipschitz differentiable in $\mathcal{S}(x^0)$ with a Lipschitz constant $L_g \geq 0$. Then Assumption 2 holds with $c_3 = \sqrt{\sigma_f}/(L_f + L_g)$, where σ_f and L_f are defined in* (2.4) *and* (3.29), *respectively.*

Proof. Let $x \in \mathcal{S}(x^0)$ be arbitrarily chosen. It follows from (3.10) and the differentiability of g that

$$\nabla f(x) + \nabla^2 f(x)\tilde{d}(x) + \nabla g(x + \tilde{d}(x)) = 0,$$

which, together with (3.8), (3.29) and the Lipschitz continuity of ∇g, implies that

$$\begin{aligned} \bar{\lambda}(x) &= \|\nabla f(x) + \nabla g(x)\|_x^* \leq \tfrac{1}{\sqrt{\sigma_f}}\|\nabla f(x) + \nabla g(x)\|, \\ &= \tfrac{1}{\sqrt{\sigma_f}}\|\nabla g(x) - \nabla g(x + \tilde{d}(x)) - \nabla^2 f(x)\tilde{d}(x)\| \leq \tfrac{L_f + L_g}{\sqrt{\sigma_f}}\|\tilde{d}(x)\|. \end{aligned}$$

and hence the conclusion holds. ∎

We next provide a lower bound for $\bar{\lambda}(x)$ in terms of the optimality gap, which will play crucial role in our subsequent analysis.

Lemma 4.2 *Let $x \in \mathrm{dom}(F)$ and $\bar{\lambda}(x)$ be defined in* (3.8). *Then*

$$\bar{\lambda}(x) \geq \omega_*^{-1}(F(x) - F^*), \tag{4.18}$$

where ω_*^{-1} is the inverse function of ω_* when restricted to the interval $[0, 1)$.

Proof. Observe from (1.14) that $\omega_*(t) \in [0, \infty)$ for $t \in [0, 1)$ and ω_* is strictly increasing in $[0, 1)$. Thus its inverse function ω_*^{-1} is well-defined when restricted to this interval. It also follows that $\omega_*^{-1}(t) \in [0, 1)$ for $t \in [0, \infty)$ and ω_*^{-1} is strictly increasing in $[0, \infty)$. We divide the rest of the proof into two separable cases as follows.

Case 1): $\bar{\lambda}(x) < 1$. It follows from Theorem 3.2 that $F(x) - F^* \leq \omega_*(\bar{\lambda}(x))$. Taking ω_*^{-1} on both sides of this relation and using the monotonicity of ω_*^{-1}, we see that (4.18) holds.

Case 2): $\bar{\lambda}(x) \geq 1$. (4.18) clearly holds in this case due to $\omega_*^{-1}(t) \in [0, 1)$ for all $t \geq 0$ ∎

In what follows, we show that under Assumption 2 RBPDN enjoys a global linear convergence.

[3]This covers the case where $g = 0$, which, for instance, arises in the interior point methods for solving smooth convex optimization problems.

Theorem 4.3 *Let $\{x^k\}$ be generated by RBPDN. Suppose that Assumption 2 holds. Then*

$$\mathbf{E}[F(x^k) - F^*] \leq \left[1 - \frac{c_4^2 p_{\min}^2 (1 - \omega_*^{-1}(\delta_0))}{2(1 + c_4 p_{\min} \omega_*^{-1}(\delta_0))}\right]^k (F(x^0) - F^*), \qquad \forall k \geq 0,$$

where $\delta_0 = F(x^0) - F^*$,

$$c_4 = \frac{c_1 c_3 \sqrt{n \sigma_f}}{1 - \eta}, \tag{4.19}$$

and σ_f and c_1 are defined in (2.4) *and* (3.31), *respectively.*

Proof. Let $k \geq 0$ be arbitrarily chosen. For convenience, let $x = x^k$ and $x^+ = x^{k+1}$. By the updating scheme of x^{k+1}, one can observe that $x_j^+ = x_j$ for $j \neq \iota$ and

$$x_\iota^+ = x_\iota + \frac{d_\iota(x)}{1 + \lambda_\iota(x)},$$

where $\iota \in \{1, \ldots, n\}$ is randomly chosen with probability p_ι and $d_\iota(x)$ is an approximate solution to problem (3.14) that satisfies (3.15) and (3.16) for some v_ι and $\eta \in [0, 1/4]$. To prove this theorem, it suffices to show that

$$\mathbf{E}_\iota[F(x^+) - F^*] \leq \left[1 - \frac{c_4^2 p_{\min}^2 (1 - \omega_*^{-1}(\delta_0))}{2(1 + c_4 p_{\min} \omega_*^{-1}(\delta_0))}\right] (F(x) - F^*). \tag{4.20}$$

Indeed, it follows from (3.33), (4.19) and Assumption 2 that

$$\sum_{i=1}^n \lambda_i(x) \;\geq\; \frac{c_1 \sqrt{n \sigma_f}}{1 - \eta} \|\tilde{d}(x)\| \;\geq\; c_4 \bar{\lambda}(x).$$

This together with (4.18) yields

$$\sum_{i=1}^n \lambda_i(x) \;\geq\; c_4 \omega_*^{-1}(F(x) - F^*).$$

Using this, (4.1) and the monotonicity of ω in $[0, \infty)$, we obtain that

$$\mathbf{E}_\iota[F(x^+)] \leq F(x) - \frac{1}{2}\omega\left(c_4 p_{\min} \omega_*^{-1}(F(x) - F^*)\right).$$

Let $\delta^+ = F(x^+) - F^*$ and $\delta = F(x) - F^*$. It then follows that

$$\mathbf{E}_\iota[\delta^+] \leq \delta - \frac{1}{2}\omega\left(c_4 p_{\min} \omega_*^{-1}(\delta)\right). \tag{4.21}$$

Consider the function $t = \omega_*^{-1}(s)$. Then $s = \omega_*(t)$. Differentiating both sides with respect to s, we have

$$(\omega_*(t))' \frac{dt}{ds} = 1,$$

which along with $\omega_*(t) = -t - \ln(1 - t)$ yields

$$(\omega_*^{-1}(s))' = \frac{dt}{ds} = \frac{1}{(\omega_*(t))'} = \frac{1 - t}{t} = \frac{1 - \omega_*^{-1}(s)}{\omega_*^{-1}(s)}.$$

In view of this and $\omega(t) = t - \ln(1 + t)$, one has that for any $\alpha > 0$,

$$\frac{d}{ds}[\omega(\alpha \omega_*^{-1}(s))] = \alpha \omega'(\alpha \omega_*^{-1}(s))(\omega_*^{-1}(s))' = \alpha \cdot \frac{\alpha \omega_*^{-1}(s)}{1 + \alpha \omega_*^{-1}(s)} \cdot \frac{1 - \omega_*^{-1}(s)}{\omega_*^{-1}(s)} = \frac{\alpha^2 (1 - \omega_*^{-1}(s))}{1 + \alpha \omega_*^{-1}(s)}. \tag{4.22}$$

Notice that $\delta \leq \delta_0$ due to $x \in \mathcal{S}(x^0)$. By this and the monotonicity of ω_*^{-1}, one can see that

$$\omega_*^{-1}(s) \leq \omega_*^{-1}(\delta) \leq \omega_*^{-1}(\delta_0), \qquad \forall s \in [0, \delta],$$

which implies that

$$\frac{1-\omega_*^{-1}(s)}{1+\alpha\omega_*^{-1}(s)} \geq \frac{1-\omega_*^{-1}(\delta_0)}{1+\alpha\omega_*^{-1}(\delta_0)}, \qquad \forall s \in [0, \delta].$$

Also, observe that $\omega(\alpha\omega_*^{-1}(0)) = 0$. Using these relations and (4.22), we have

$$\omega(\alpha\omega_*^{-1}(\delta)) = \int_0^\delta \frac{d}{ds}[\omega(\alpha\omega_*^{-1}(s))]ds = \int_0^\delta \frac{\alpha^2(1-\omega_*^{-1}(s))}{1+\alpha\omega_*^{-1}(s)}ds \geq \frac{\alpha^2(1-\omega_*^{-1}(\delta_0))}{1+\alpha\omega_*^{-1}(\delta_0)}\delta.$$

This and (4.21) with $\alpha = c_4 p_{\min}$ lead to

$$\mathbf{E}_\iota[\delta^+] \leq \left[1 - \frac{c_4^2 p_{\min}^2(1-\omega_*^{-1}(\delta_0))}{2(1+c_4 p_{\min}\omega_*^{-1}(\delta_0))}\right]\delta,$$

which gives (4.20) as desired. ∎

The following result is an immediate consequence of Proposition 4.1 and Theorem 4.3.

Corollary 4.1 *Let $\{x^k\}$ be generated by RBPDN. Suppose that g is Lipschitz differentiable in $\mathcal{S}(x^0)$ with a Lipschitz constant $L_g \geq 0$. Then*

$$\mathbf{E}[F(x^k) - F^*] \leq \left[1 - \frac{\tilde{c}_4^2 p_{\min}^2(1-\omega_*^{-1}(\delta_0))}{2(1+\tilde{c}_4 p_{\min}\omega_*^{-1}(\delta_0))}\right]^k (F(x^0) - F^*), \qquad \forall k \geq 0,$$

where $\delta_0 = F(x^0) - F^*$,

$$\tilde{c}_4 = \frac{\sqrt{n}c_1\sigma_f}{(1-\eta)(L_f + L_g)},$$

and σ_f, L_f and c_1 are defined in (2.4), (3.29) *and* (3.31), *respectively.*

One can observe that RBPDN reduces to PDN [31] or DN [22] [4] by setting $n = 1$. It thus follows from Corollary 4.1 that PDN for a class of g and DN are globally linearly convergent, which is stated below. To the best of our knowledge, this result was previously unknown in the literature.

Corollary 4.2 *Suppose that g is Lipschitz differentiable in $\mathcal{S}(x^0)$. Then PDN [31] for such g and DN [22] are globally linearly convergent.*

Before ending this subsection we show that Corollary 4.2 can be used to sharpen the existing iteration complexity of some methods in [22, 40, 31].

A mixture of DN and Newton methods is presented in [22, Section 4.1.5] for solving problem (1.1) with $g = 0$. In particular, this method consists of two stages. Given an initial point x^0, $\beta \in (0, (3-\sqrt{5})/2)$ and $\epsilon > 0$, the first stage performs the DN iterations

$$x^{k+1} = x^k - \frac{\tilde{d}(x^k)}{1+\tilde{\lambda}(x^k)} \tag{4.23}$$

until finding some x^{K_1} such that $\tilde{\lambda}(x^{K_1}) \leq \beta$, where $\tilde{d}(\cdot)$ and $\tilde{\lambda}(\cdot)$ are defined in (3.10) and (3.11), respectively. The second stage executes the standard Newton iterations

$$x^{k+1} = x^k - \tilde{d}(x^k), \tag{4.24}$$

[4] PDN becomes DN if $g = 0$.

starting at x^{K_1} and terminating at some x^{K_2} such that $\tilde{\lambda}(x^{K_2}) \leq \epsilon$. As shown in [22, Section 4.1.5], the second stage converges quadratically:

$$\tilde{\lambda}(x^{k+1}) \leq \left(\frac{\tilde{\lambda}(x^k)}{1-\tilde{\lambda}(x^k)}\right)^2, \qquad \forall k \geq K_1. \tag{4.25}$$

In addition, an upper bound on K_1 is established in [22, Section 4.1.5], which is

$$K_1 \leq \left\lceil (F(x^0) - F^*)/\omega(\beta) \right\rceil. \tag{4.26}$$

In view of (4.25), one can easily show that

$$K_2 - K_1 \leq \left\lceil \log_2\left(\frac{\log\epsilon - 2\log(1-\beta)}{\log\beta - 2\log(1-\beta)}\right)\right\rceil. \tag{4.27}$$

Observe that the first stage of this method is just DN, which is a special case of RBPDN with $n = 1$ and $\eta = 0$. It thus follows from Corollary 4.2 that the first stage converges linearly. In fact, it can be shown that

$$F(x^{k+1}) - F^* \leq \left(1 - \frac{1-\omega_*^{-1}(\delta_0)}{1+\omega_*^{-1}(\delta_0)}\right)(F(x^k) - F^*), \qquad \forall k \leq K_1, \tag{4.28}$$

where $\delta_0 = F(x^0) - F^*$. Indeed, since $g = 0$, one can observe from (3.8) and (3.11) that $\tilde{\lambda}(x^k) = \bar{\lambda}(x^k)$. It then follows from this, $g = 0$ and [22, Theorem 4.1.12] that $F(x^{k+1}) \leq F(x^k) - \omega(\bar{\lambda}(x^k))$ for all $k \leq K_1$. This together with (4.18) implies that

$$F(x^{k+1}) \leq F(x^k) - \omega(\omega_*^{-1}(F(x^k) - F^*)), \qquad \forall k \leq K_1.$$

The relation (4.28) then follows from this and a similar argument as in the proof of Theorem 4.3. Let

$$\bar{K} = \left\lceil \left[\frac{\log(\omega(\beta)) - \log\delta_0}{\log\left(1 - \frac{1-\omega_*^{-1}(\delta_0)}{1+\omega_*^{-1}(\delta_0)}\right)}\right]_+ \right\rceil,$$

where $t_+ = \max(t, 0)$. In view of (4.28), one can easily verify that $F(x^{\bar{K}}) - F^* \leq \omega(\beta)$, which along with (3.13) implies that $\tilde{\lambda}(x^{\bar{K}}) \leq \beta$. By (4.26) and the definition of K_1, one can have $K_1 \leq \min\left\{\bar{K}, \lceil \delta_0/\omega(\beta) \rceil\right\}$, which sharpens the bound (4.26). Combining this relation and (4.27), we thus obtain the following new iteration complexity for finding an approximate solution of (1.1) with $g = 0$ by a mixture of DN and Newton method [22, Section 4.1.5].

Theorem 4.4 *Let $x^0 \in \mathrm{dom}(F)$, $\beta \in (0, (3-\sqrt{5})/2)$ and $\epsilon > 0$ be given. Then the mixture of DN and Newton methods [22, Section 4.1.5] for solving problem* (1.1) *with $g = 0$ requires at most*

$$\min\left\{\left\lceil \left[\frac{\log(\omega(\beta)) - \log\delta_0}{\log\left(1 - \frac{1-\omega_*^{-1}(\delta_0)}{1+\omega_*^{-1}(\delta_0)}\right)}\right]_+ \right\rceil, \left\lceil \frac{\delta_0}{\omega(\beta)} \right\rceil\right\} + \left\lceil \log_2\left(\frac{\log\epsilon - 2\log(1-\beta)}{\log\beta - 2\log(1-\beta)}\right)\right\rceil$$

iterations for finding some x^k satisfying $\tilde{\lambda}(x^k) \leq \epsilon$, where $\delta_0 = F(x^0) - F^*$.

Recently, Zhang and Xiao [40] proposed an inexact DN method for solving problem (1.1) with $g = 0$, whose iterations are updated as follows:

$$x^{k+1} = x^k - \frac{\hat{d}(x^k)}{1+\hat{\lambda}(x^k)}, \qquad \forall k \geq 0,$$

where $\hat{d}(x^k)$ is an approximation to $\tilde{d}(x^k)$ and $\hat{\lambda}(x^k) = \sqrt{\langle \hat{d}(x^k), \nabla^2 f(x^k)\hat{d}(x^k)\rangle}$ (see [40, Algorithm 1] for details). It is shown in [40, Theorem 1] that such $\{x^k\}$ satisfies

$$F(x^{k+1}) \leq F(x^k) - \frac{1}{2}\omega(\tilde{\lambda}(x^k)), \qquad \forall k \geq 0, \tag{4.29}$$

$$\omega(\tilde{\lambda}(x^{k+1})) \leq \frac{1}{2}\omega(\tilde{\lambda}(x^k)), \qquad \text{if } \tilde{\lambda}(x^k) \leq 1/6, \tag{4.30}$$

where $\tilde{\lambda}(\cdot)$ is defined in (3.11). These relations are used in [40] for deriving an iteration complexity of the inexact DN method. In particular, its complexity analysis is divided into two parts. The first part estimates the number of iterations required for generating some x^{K_1} satisfying $\tilde{\lambda}(x^{K_1}) \leq 1/6$, while the second part estimates the additional iterations needed for generating some x^{K_2} satisfying $F(x^{K_2}) - F^* \leq \epsilon$. In [40], the relation (4.29) is used to show that

$$K_1 \leq \lceil (2(F(x^0) - F^*))/\omega(1/6) \rceil, \tag{4.31}$$

while (4.30) is used to establish

$$K_2 - K_1 \leq \left\lceil \log_2\left(\frac{2\omega(1/6)}{\epsilon}\right)\right\rceil. \tag{4.32}$$

It follows from these two relations that the inexact DN method can find an approximate solution x^k satisfying $F(x^k) - F^* \leq \epsilon$ in at most

$$\left\lceil \frac{2(F(x^0) - F^*)}{\omega(1/6)} \right\rceil + \left\lceil \log_2\left(\frac{2\omega(1/6)}{\epsilon}\right)\right\rceil$$

iterations, which is stated in [40, Corollary 1].

By a similar analysis as above, one can show that the inexact DN method ([40, Algorithm 1]) is globally linearly convergent. In fact, it can be shown that

$$F(x^{k+1}) - F^* \leq \left(1 - \frac{1-\omega_*^{-1}(\delta_0)}{2(1+\omega_*^{-1}(\delta_0))}\right)(F(x^k) - F^*), \qquad \forall k \geq 0, \tag{4.33}$$

where $\delta_0 = F(x^0) - F^*$. Indeed, since $g = 0$, one has $\tilde{\lambda}(x^k) = \bar{\lambda}(x^k)$. It follows from this, (4.18) and (4.29) that

$$F(x^{k+1}) \leq F(x^k) - \frac{1}{2}\omega(\omega_*^{-1}(F(x^k) - F^*)), \qquad \forall k \geq 0.$$

The relation (4.33) then follows from this and a similar derivation as in the proof of Theorem 4.3. By (4.31), (4.33) and a similar argument as above, one can have

$$K_1 \leq \min\left\{ \left\lceil \left[\frac{\log(\frac{1}{2}\omega(1/6)) - \log\delta_0}{\log\left(1 - \frac{1-\omega_*^{-1}(\delta_0)}{2(1+\omega_*^{-1}(\delta_0))}\right)} \right]_+ \right\rceil, \left\lceil \frac{2\delta_0}{\omega(1/6)} \right\rceil \right\},$$

which improves the bound (4.31). Combining this relation and (4.32), we thus obtain the following new iteration complexity for finding an approximate solution of (1.1) with $g = 0$ by the aforementioned inexact DN method.

Theorem 4.5 *Let $x^0 \in \mathrm{dom}(F)$ and $\epsilon > 0$ be given. Then the inexact DN method ([40, Algorithm 1]) for solving problem (1.1) with $g = 0$ requires at most*

$$\min\left\{ \left\lceil \left[\frac{\log(\frac{1}{2}\omega(1/6)) - \log\delta_0}{\log\left(1 - \frac{1-\omega_*^{-1}(\delta_0)}{2(1+\omega_*^{-1}(\delta_0))}\right)} \right]_+ \right\rceil, \left\lceil \frac{2\delta_0}{\omega(1/6)} \right\rceil \right\} + \left\lceil \log_2\left(\frac{2\omega(1/6)}{\epsilon}\right)\right\rceil$$

iterations for finding some x^k satisfying $F(x^k) - F^* \leq \epsilon$, where $\delta_0 = F(x^0) - F^*$.

Dinh-Tran et al. recently proposed in [31, Algorithm 1] a proximal Newton method for solving problem (1.1) with general g. Akin to the aforementioned method [22, Section 4.1.5] for (1.1) with $g = 0$, this method also consists of two stages (or phases). The first stage performs the PDN iterations in the form of (4.23) for finding some x^{K_1} such that $\tilde{\lambda}(x^{K_1}) \leq \omega(0.2)$, while the second stage executes the proximal Newton iterations in the form of (4.24) starting at x^{K_1} and terminating at some x^{K_2} such that $\tilde{\lambda}(x^{K_2}) \leq \epsilon$. As shown in [31, Theorem 6], the second stage converges quadratically. The following relations are essentially established in [31, Theorem 7]:

$$K_1 \leq \lceil (F(x^0) - F^*)/\omega(0.2) \rceil, \tag{4.34}$$

$$K_2 - K_1 \leq \left\lceil 1.5 \log\log \frac{0.28}{\epsilon} \right\rceil. \tag{4.35}$$

Throughout the remainder of this subsection, suppose that Assumption 2 holds. Observe that the first stage of this method is just PDN, which is a special case of RBPDN with $n = 1$ and $\eta = 0$. It thus follows from Corollary 4.2 that the first stage converges linearly. In fact, it can be shown that

$$F(x^{k+1}) - F^* \leq \left[1 - \frac{\hat{c}^2(1-\omega_*^{-1}(\delta_0))}{(1+\hat{c}\omega_*^{-1}(\delta_0)}\right]^k (F(x^0) - F^*), \qquad \forall k \leq K_1, \tag{4.36}$$

where $\delta_0 = F(x^0) - F^*$, $\hat{c} = c_3\sqrt{\sigma_f}$, and σ_f and c_3 are given in (2.4) and Assumption 2, respectively. Indeed, by (3.11) and (3.34), one has $\|\tilde{d}(x^k)\| \leq \tilde{\lambda}(x^k)/\sqrt{\sigma_f}$. In addition, by Assumption 2, we have $\|\tilde{d}(x^k)\| \geq c_3\bar{\lambda}(x^k)$. It follows from these two relations that $\tilde{\lambda}(x^k) \geq \hat{c}\bar{\lambda}(x^k)$, which together with (4.18) yields $\tilde{\lambda}(x^k) \geq \hat{c}\omega_*^{-1}(F(x^k) - F^*)$. This and (3.12) imply that

$$F(x^{k+1}) \leq F(x^k) - \omega(\hat{c}\omega_*^{-1}(F(x^k) - F^*)), \qquad \forall k \leq K_1.$$

The relation (4.36) then follows from this and a similar argument as in the proof of Theorem 4.3. Let

$$\bar{K} = \left\lceil \left\lceil \frac{\log(\omega(0.2)) - \log\delta_0}{\log\left(1 - \frac{\hat{c}^2(1-\omega_*^{-1}(\delta_0))}{(1+\hat{c}\omega_*^{-1}(\delta_0)}\right)} \right\rceil_+ \right\rceil.$$

By (4.36), one can easily verify that $F(x^{\bar{K}}) - F^* \leq \omega(0.2)$, which along with (3.13) implies that $\tilde{\lambda}(x^{\bar{K}}) \leq 0.2$. By (4.26) and the definition of K_1, one can have $K_1 \leq \min\{\bar{K}, \lceil \delta_0/\omega(0.2) \rceil\}$, which sharpens the bound (4.34). Combining this relation and (4.35), we thus obtain the following new iteration complexity for finding an approximate solution of (1.1) by the aforementioned proximal Newton method.

Theorem 4.6 *Let $x^0 \in \mathrm{dom}(F)$ and $\epsilon > 0$ be given. Suppose that Assumption 2 holds. Then the proximal Newton method [31, Algorithm 1] for solving problem* (1.1) *requires at most*

$$\min\left\{ \left\lceil \left\lceil \frac{\log(\omega(0.2)) - \log\delta_0}{\log\left(1 - \frac{\hat{c}^2(1-\omega_*^{-1}(\delta_0))}{(1+\hat{c}\omega_*^{-1}(\delta_0)}\right)} \right\rceil_+ \right\rceil, \left\lceil \frac{\delta_0}{\omega(0.2)} \right\rceil \right\} + \left\lceil 1.5 \log\log \frac{0.28}{\epsilon} \right\rceil$$

iterations for finding some x^k satisfying $\tilde{\lambda}(x^k) \leq \epsilon$, where $\delta_0 = F(x^0) - F^*$, $\hat{c} = c_3\sqrt{\sigma_f}$, and σ_f and c_3 are given in (2.4) *and Assumption 2, respectively.*

Remark: Suppose that g is Lipschitz differentiable in $\mathcal{S}(x^0)$ with a Lipschitz constant $L_g \geq 0$. It follows from Proposition 4.1 that Assumption 2 holds with $c_3 = \sqrt{\sigma_f}/(L_f + L_g)$, where L_f is defined in (3.29), and thus Theorem 4.6 holds with $\hat{c} = \sigma_f/(L_f + L_g)$.

5 Numerical results

In this section we conduct numerical experiment to test the performance of RBPDN. In particular, we apply RBPDN to solve a regularized logistic regression (RLR) model and a sparse regularized logistic regression (SRLR) model. We also compare RBPDN with a randomized block accelerated proximal gradient (RBAPG) method proposed in [15] on these problems. All codes are written in MATLAB and all computations are performed on a MacBook Pro running with Mac OS X Lion 10.7.4 and 4GB memory.

For the RLR problem, our goal is to minimize a regularized empirical logistic loss function, particularly, to solve the problem:

$$L_\mu^* := \min_{x\in\Re^N} \left\{ L_\mu(x) := \frac{1}{m}\sum_{i=1}^m \log(1+\exp(-y_i\langle w^i, x\rangle)) + \frac{\mu}{2}\|x\|^2 \right\} \tag{5.1}$$

for some $\mu > 0$, where $w^i \in \Re^N$ is a sample of N features and $y_i \in \{-1, 1\}$ is a binary classification of this sample. This model has recently been considered in [40]. Similarly, for the SRLR problem, we aim to solve the problem:

$$L_{\gamma,\mu}^* := \min_{x\in\Re^N} \left\{ L_{\gamma,\mu}(x) := \frac{1}{m}\sum_{i=1}^m \log(1+\exp(-y_i\langle w^i, x\rangle)) + \frac{\mu}{2}\|x\|^2 + \gamma\|x\|_1 \right\} \tag{5.2}$$

for some $\mu, \gamma > 0$.

In our experiments below, we fix $m = 1000$ and set $N = 3000, 6000, \ldots, 30000$. For each pair (m, N), we randomly generate 10 copies of data $\{(w^i, y_i)\}_{i=1}^m$ independently. In each copy, the elements of w^i are generated according to the standard uniform distribution on the open interval $(0, 1)$ and y_i is generated according to the distribution $\mathbf{P}(\xi = -1) = \mathbf{P}(\xi = 1) = 1/2$. As in [40], we normalize the data so that $\|w^i\| = 1$ for all $i = 1, \ldots, m$, and set the regularization parameters $\mu = 10^{-5}$ and $\gamma = 10^{-4}$.

We now apply RBPDN and RBAPG to solve problem (5.1). For both methods, the decision variable $x \in \Re^N$ is divided into 10 blocks sequentially and equally. At each iteration k, they pick a block ι uniformly at random. For RBPDN, it needs to find a search direction $d_\iota(x^k)$ satisfying (2.2) and (2.3) with $f = L_\mu$ and $g = 0$, that is,

$$\nabla^2_{\iota\iota} L_\mu(x^k) d_\iota(x^k) + \nabla_\iota L_\mu(x^k) + v_\iota = 0, \tag{5.3}$$

$$\sqrt{\langle v_\iota, (\nabla^2_{\iota\iota} L_\mu(x^k))^{-1} v_\iota\rangle} \le \eta \sqrt{\langle d_\iota(x^k), \nabla^2_{\iota\iota} L_\mu(x^k) d_\iota(x^k)\rangle} \tag{5.4}$$

for some $\eta \in [0, 1/4]$. To obtain such a $d_\iota(x^k)$, we apply conjugate gradient method to solve the equation

$$\nabla^2_{\iota\iota} L_\mu(x^k) d_\iota = -\nabla_\iota L_\mu(x^k)$$

until an approximate solution d_ι satisfying

$$\|\nabla^2_{\iota\iota} L_\mu(x^k) d_\iota + \nabla_\iota L_\mu(x^k)\| \le \frac{1}{4}\sqrt{\mu\langle d_\iota, \nabla^2_{\iota\iota} L_\mu(x^k) d_\iota\rangle}. \tag{5.5}$$

is found and then set $d_\iota(x^k) = d_\iota$. Notice from (5.1) that $\nabla^2_{\iota\iota} L_\mu(x^k) \succeq \mu I$. In view of this, one can verify that such $d_\iota(x^k)$ satisfies (5.3) and (5.4) with $\eta = 1/4$. In addition, we choose $x^0 = 0$ for both methods and terminate them once the duality gap is below 10^{-3}. More specifically, one can easily derive a dual of problem (5.1) given by

$$\max_{s\in\Re^m} \left\{ D_\mu(s) := -\frac{1}{m}\sum_{i=1}^m \log(1-ms_i) - \frac{1}{2\mu}\left\|\sum_{i=1}^m s_i y_i w^i\right\|^2 - \sum_{i=1}^m s_i \log\left(\frac{ms_i}{1-ms_i}\right) \right\}.$$

Let $\{x^k\}$ be a sequence of approximate solutions to problem (5.1) generated by RBPDN or RBAPG and $s^k \in \Re^m$ the associated dual sequence defined as follows:

$$s_i^k = \frac{\exp(-y_i\langle w^i, x^k\rangle)}{m(1+\exp(-y_i\langle w^i, x^k\rangle))}, \qquad i = 1, \ldots, m. \tag{5.6}$$

We use $L_\mu(x^k) - D_\mu(s^k) \leq 10^{-3}$ as the termination criterion for RBPDN or RBAPG, which is checked once every 10 iterations.

The computational results averaged over the 10 copies of data generated above are presented in Table 1. In detail, the problem size N is listed in the first column. The average number of iterations (upon round off) for RBPDN and RBAPG are given in the next two columns. The average CPU time (in seconds) for these methods are presented in columns four and five, and the average objective function value of (5.1) obtained by them are given in the last two columns. One can observe that both methods are comparable in terms of objective values, but RBPDN substantially outperforms RBAPG in terms of CPU time.

In the next experiment, we apply RBPDN and RBAPG to solve problem (5.2). Same as above, the decision variable $x \in \Re^N$ is divided into 10 blocks sequentially and equally. At each iteration k, they pick a block ι uniformly at random. For RBPDN, it needs to compute a search direction $d_\iota(x^k)$ satisfying (2.2) and (2.3) with $f = L_{\gamma,\mu}$ and $g = \gamma\|\cdot\|_1$, that is,

$$-v_\iota \in \nabla^2_{\iota\iota} L_\mu(x^k) d_\iota(x^k) + \nabla_\iota L_\mu(x^k) + \gamma\partial(\|x_\iota^k + d_\iota(x^k)\|_1), \tag{5.7}$$

$$\sqrt{\langle v_\iota, (\nabla^2_{\iota\iota} L_\mu(x^k))^{-1} v_\iota\rangle} \leq \eta\sqrt{\langle d_\iota(x^k), \nabla^2_{\iota\iota} L_\mu(x^k) d_\iota(x^k)\rangle} \tag{5.8}$$

for some $\eta \in [0, 1/4]$. To obtain such a $d_\iota(x^k)$, we apply FISTA [1] to solve the problem

$$\min_{d_\iota} \left\{ \frac{1}{2}\langle d_\iota, \nabla^2_{\iota\iota} L_\mu(x^k) d_\iota\rangle + \langle \nabla_\iota L_\mu(x^k), d_\iota\rangle + \gamma\|x_\iota^k + d_\iota\|_1 \right\}$$

until an approximate solution d_ι satisfying (5.5) and (5.7) is found and then set $d_\iota(x^k) = d_\iota$. By the same argument as above, one can see that such $d_\iota(x^k)$ also satisfies (5.8) with $\eta = 1/4$. In addition, we choose $x^0 = 0$ for both methods and terminate them the duality gap is below 10^{-3}. More specifically, one can easily derive a dual of problem (5.2) as follows:

$$\max_{s\in\Re^m} \left\{ \begin{array}{l} D_{\gamma,\mu}(s) := -\frac{1}{m}\sum_{i=1}^m \log(1 - ms_i) + \frac{\mu}{2}\|h(s)\|^2 + \gamma\|\theta(s)\|_1 - \sum_{i=1}^m s_i \log\left(\frac{ms_i}{1-ms_i}\right) \\ \qquad -\langle \sum_{i=1}^m s_i y_i w^i, h(s)\rangle \end{array} \right\},$$

where

$$h(s) := \arg\min_{h\in\Re^n} \left\{ \frac{\mu}{2}\|h\|^2 - \langle \sum_{i=1}^m s_i y_i w^i, h\rangle + \gamma\|h\|_1 \right\}, \qquad \forall s \in \Re^m.$$

Let $\{x^k\}$ be a sequence of approximate solutions to problem (5.2) generated by RBPDN or RBAPG and $s^k \in \Re^m$ the associated dual sequence defined as in (5.6). We use $L_{\gamma,\mu}(x^k) - D_{\gamma,\mu}(s^k) \leq 10^{-3}$ as the termination criterion for RBPDN or RBAPG, which is checked once every 10 iterations.

The computational results averaged over the 10 copies of data generated above are presented in Table 2, which is similar to Table 1 except that it has two additional columns displaying the average cardinality (upon round off) of the solutions obtained by RBPDN and RBAPG. We can observe that both methods are comparable in terms of objective values, but RBPDN substantially outperforms RBAPG in terms of CPU time and the sparsity of solutions.

Table 1: Comparison on RBPDN and RBAPG for solving (5.1)

Problem	Iteration		CPU Time		Objective Value	
N	RBPDN	RBAPG	RBPDN	RBAPG	RBPDN	RBAPG
3000	111	2837	0.13	2.01	0.2300	0.2298
6000	53	2756	0.12	3.61	0.2142	0.2141
9000	56	2339	0.22	5.80	0.2092	0.2092
12000	52	2083	0.32	7.64	0.2079	0.2078
15000	48	2084	0.40	10.33	0.2069	0.2069
18000	59	1881	0.59	9.23	0.2058	0.2059
21000	46	1866	0.55	10.28	0.2050	0.2050
24000	53	1854	0.72	11.33	0.2050	0.2050
27000	54	1848	0.82	12.38	0.2045	0.2044
30000	51	1924	0.87	13.87	0.2043	0.2043

Table 2: Comparison on RBPDN and RBAPG for solving (5.2)

Problem	Iteration		CPU Time		Objective Value		Cardinality	
N	RBPDN	RBAPG	RBPDN	RBAPG	RBPDN	RBAPG	RBPDN	RBAPG
3000	2233	6126	5.44	3.19	0.5529	0.5532	749	1705
6000	1003	6239	3.82	4.74	0.5941	0.5943	840	2372
9000	626	6174	3.17	6.39	0.6210	0.6211	857	3000
12000	408	5985	2.63	7.70	0.6398	0.6400	852	3108
15000	294	5762	2.30	9.06	0.6521	0.6523	815	3340
18000	272	5476	2.50	10.26	0.6616	0.6618	748	3237
21000	208	5287	2.26	11.49	0.6693	0.6694	698	3173
24000	186	5146	2.31	12.76	0.6748	0.6748	650	3334
27000	180	5059	2.78	14.37	0.6790	0.6791	571	4157
30000	153	4942	2.74	15.80	0.6824	0.6824	527	4312

References

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

[2] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM J. Optim.*, 13(4):2037–2060, 2013.

[3] S. Becker and M. J. Fadili. A quasi-Newton proximal splitting method. In Proceedings of Neutral Information Processing Systems Foundation, 2012.

[4] R. H. Byrd, J. Nocedal, and S. Solntsev. An algorithm for quadratic ℓ_1-regularized optimization with a flexible active-set strategy. *Optim. Method Softw.*, 30(6):1213–1237, 2015.

[5] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. Coordinate descent method for large-scale ℓ_2-loss linear support vector machines. *J. Mach. Learn. Res.*, 9:1369–1398, 2008.

[6] A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.*, 30:56–66, 2008.

[7] O. Fercoq and P. Richtárik. Accelerated, parallel and proximal coordinate descent. *SIAM J. Optim.*, 25(4): 1997–2023, 2015.

[8] C. J. Hsieh, M.A. Sustik, I.S. Dhillon, and P. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in Neutral Information Processing Systems (NIPS)*, 24:1–18, 2011.

[9] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

[10] E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation applied to compressed sensing: Implementation and numerical experiments. *J. Comput. Math*, 28(2):170-194, 2010.

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

[12] J.D. Lee, Y. Sun, and M.A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM J. Optim.*, 24(3), 1420–1443, 2014.

[13] Y. T. Lee and A. Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In Proceedings of IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS), pages 147–156, Berkeley, CA, October 2013. Full version at arXiv:1305.1922.

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

[15] Q. Lin, Z. Lu, and L. Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM J. Optim.*, 25(4): 2244–2273, 2015.

[16] J. Liu, S. J. Wright, C. Re, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *JMLR W & CP*, 32(1):469–477, 2014.

[17] Z. Lu. Adaptive first-order methods for general sparse inverse covariance selection. *SIAM J. Matrix Anal. Appl.*, 31(4):2000–2016, 2010.

[18] Z. Lu and X. Chen. Generalized conjugate gradient methods for ℓ_1 regularized convex quadratic programming with finite convergence. arXiv:1511.07837, 2015.

[19] Z. Lu and L. Xiao. A randomized nonmonotone block proximal gradient method for a class of structured nonlinear programming. arXiv:1306.5918, 2013.

[20] Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Math. Program.*, 152(1-2):615–642, 2015.

[21] A. Milzarek and M. Ulbrich. A semismooth Newton method with multidimensional filter globalization for l_1-optimization. *SIAM J. Optim.*, 24: 298–333, 2014.

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

[23] Y. Nesterov. Gradient methods for minimizing composite objective function. *Math. Program.*, 140(1):125–161, 2013.

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

[25] Y. Nesterov and A. Nemirovski. Interior Point Polynomial Time Methods in Convex Programming. SIAM, Philadelphia, 1994.

[26] 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.

[27] Z. Qu, P. Richtárik, M. Takáč, and O. Fercoq. SDNA: stochastic dual Newton ascent for empirical risk minimization. *Proceedings of The 33rd International Conference on Machine Learning*, 1823–1832, 2016.

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

[29] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *J. Mach. Learn. Res.* 14:567–599, 2013.

[30] Q. Tran-Dinh, A. Kyrillidis, and V. Cevher. An inexact proximal path-following algorithm for constrained convex minimization. *SIAM J. Optim.*, 24(4): 1718–1745, 2014.

[31] Q. Tran-Dinh, A. Kyrillidis, and V. Cevher. Composite self-concordant minimization. *J. Mach. Learn. Res.*, 16: 371–416, 2015.

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

[33] 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.

[34] Z. Wen, D. Goldfarb, and K. Scheinberg. Block coordinate descent methods for semidefi nite programming. In M. F. Anjos and J. B. Lasserre, editors, Handbook on Semide finite, Cone and Polynomial Optimization: Theory, Algorithms, Software and Applications, volume 166, pages 533–564. Springer, 2012.

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

[36] S. J. Wright, R. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE T. Signal Proces.*, 57:2479–2493, 2009.

[37] L. Xiao and T. Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM J. Optim.*, 23(2):1062–1091, 2013.

[38] J. Yang and Y. Zhang. Alternating direction algorithm for ℓ_1-problems in compressive sensing. *SIAM J. Sci. Comput.*, 33:250–278, 2010.

[39] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35, 2007.

[40] Y. Zhang and L. Xiao. Communication-efficient distributed optimization of self-concordant empirical loss. arXiv:1501.00263, January 2015.

[41] Y. Zhang and L. Xiao. DiSCO: communication-efficient distributed optimization of self-concordant loss. International Conference on Machine Learning (ICML 2015).