SIAM J. OPTIM.
Vol. 26, No. 1, pp. 197–226

PARALLEL RANDOM COORDINATE DESCENT METHOD FOR COMPOSITE MINIMIZATION: CONVERGENCE ANALYSIS AND ERROR BOUNDS*

ION NECOARA† AND DRAGOS CLIPICI†

Abstract. In this paper we employ a parallel version of a randomized (block) coordinate descent method for minimizing the sum of a partially separable smooth convex function and a fully separable nonsmooth convex function. Under the assumption of Lipschitz continuity of the gradient of the smooth function, this method has a sublinear convergence rate. Linear convergence rate of the method is obtained for the newly introduced class of *generalized error bound functions*. We prove that the new class of generalized error bound functions encompasses both global/local error bound functions and smooth strongly convex functions. We also show that the theoretical estimates on the convergence rate depend on the number of blocks chosen randomly and a natural measure of separability of the smooth component of the objective function.

Key words. partially separable functions, composite minimization, parallel random coordinate descent algorithm, rate of convergence, generalized error bound condition

AMS subject classifications. 90C25, 90C06, 90C46, 65K05, 49M27

DOI. 10.1137/130950288

1. Introduction.

In recent years there has been an ever-increasing interest in the optimization community for algorithms suitable for solving convex optimization problems with a very large number of variables. These optimization problems, known as big data problems, have arisen from more recent fields, such as network control [16, 10] or machine learning [2]. An important property of these problems is that they are *partially separable*, which permits parallel and/or distributed computations in the optimization algorithms that are to be designed for them [10, 22]. This, together with the surge of multicore machines or clustered parallel computing technology in the past decade, has led to the widespread focus on coordinate descent methods.

State of the art. Coordinate descent schemes are methods in which a number of (block) coordinate updates of vector of variables are conducted at each iteration. The reasoning behind this is that coordinate updates for problems with a large number of variables are much simpler than computing a full update, requiring less memory and computational power, and that they can be done independently. These make coordinate descent methods more scalable and suitable for distributed and parallel computing hardware. Coordinate descent methods can be divided into two main categories: deterministic and random. In deterministic coordinate descent methods, the (block) coordinates which are to be updated at each iteration are chosen in a cyclic fashion or based on some greedy strategy. For cyclic coordinate search, estimates on the rate of convergence were given recently in [1, 5], while they were given for the greedy coordinate search in [28]. On the other hand, in random coordinate descent methods, the (block) coordinates which are to be updated are chosen randomly

*Received by the editors December 23, 2013; accepted for publication (in revised form) November 18, 2015; published electronically January 20, 2016. This work was supported by UEFISCDI Romania, PNII-RU-TE 2014, project MoCOBiDS, 176/01.10.2015.
http://www.siam.org/journals/siopt/26-1/95028.html
†Automatic Control and Systems Engineering Department, University Politehnica Bucharest, 060042, Bucharest, Romania (ion.necoara@acse.pub.ro, dragos.clipici@acse.pub.ro).

based on some probability distribution. In [18], Nesterov presents a random coordinate descent method for smooth convex problems, and under some assumption of Lipschitz gradient and strong convexity of the objective function, the algorithm was proved to have linear convergence in the expected values of the objective function. In [9, 14, 15], a 2-block random coordinate descent method is proposed to solve linearly constrained (composite) convex problems. The results in [18, 19] were combined in [24, 7] to solve composite convex problems. To our knowledge, the first results on the linear convergence of coordinate descent methods under more relaxed assumptions than smoothness and strong convexity were obtained in, e.g., [28, 8]. In particular, linear convergence of these methods is proved under some local error bound property, which is more general than the assumption of Lipschitz gradient and strong convexity as required in [18, 9, 14, 15, 24]. However, the authors in [28, 8] were able to show linear convergence only locally. Finally, very few results were known in the literature on distributed and parallel implementations of coordinate descent methods. Recently, a more thorough investigation regarding the separability of the objective function and ways in which the convergence can be accelerated through parallelization was undertaken in [14, 10, 23, 20, 12], where it is shown that speedup can be achieved through this approach. In particular, the authors in [23] developed a general framework to analyze this type of parallel methods. Several other papers on parallel coordinate descent methods have appeared around the time this paper was finalized [3, 9, 22].

Motivation. Despite widespread use of coordinate descent methods for solving large convex problems, there are some aspects that have not been fully studied. In particular, in applications, the assumption of Lipschitz gradient and strong convexity is restrictive and the main interest is in finding larger classes of functions for which we can still prove linear convergence. We are also interested in providing schemes based on parallel and/or distributed computations and analyzing in which manner the number of components to be updated enters into the convergence rate. Finally, the convergence analysis has been almost exclusively limited to centralized step size rules and local convergence results. These represent the main issues we pursue here.

Contribution. In this paper we consider a parallel version of the random (block) coordinate gradient descent method [18, 15, 24] for solving large optimization problems with a convex separable composite objective function, i.e., consisting of the sum of a partially separable smooth function and a fully separable nonsmooth function. Our approach allows us to analyze in the same framework several methods: full gradient, serial random coordinate descent, and any parallel random coordinate descent method in between. Analysis of coordinate descent methods based on updating in parallel more than one (block) component per iteration was given in, e.g., [14, 10, 23, 20]. We extend this analysis to more general problems, e.g., satisfying an error bound type property, and having more knowledge on the structure of the problem. In particular, we show that our parallel algorithm attains linear convergence for problems belonging to a general class, called *generalized error bound problems*. We establish that our class includes problems with global/local error bound objective functions and implicitly strongly convex functions with some Lipschitz gradient. We also show that the new class of problems that we define in this paper covers many applications in big data and networks. Finally, we establish that the theoretical estimates on the convergence rate depend on the number of blocks chosen randomly and a natural measure of separability of the objective function. In summary, the contributions of this paper include the following:

(i) We introduce a new class of generalized error bound problems for which

we show that it encompasses problems with global/local error bound functions and smooth strongly convex functions and that it covers many practical applications.

(ii) We present a parallel random coordinate descent method and derive, in the smooth case, sublinear convergence rate depending on the number of blocks chosen randomly and a natural measure of separability of the smooth component of the objective function. Under the generalized error bound property, we prove that the parallel algorithm has a global linear convergence rate. For this algorithm, the iterate updates can be done independently, and thus it is suitable for parallel and distributed computing architectures.

(iii) We also perform a theoretical identification of which categories of problems and objective functions satisfy our generalized error bound property introduced here.

2. Problem formulation. In many big data applications arising from, e.g., networks, control, or data ranking, we have a system formed from several entities, with a communication graph which indicates the interconnections between entities (e.g., sources and links in network optimization [21], website pages in data ranking [2], or subsystems in control [16]). We denote this *bipartite graph* as $G = ([N] \times [\bar{N}], E)$, where $[N] = \{1, \dots, N\}$, $[\bar{N}] = \{1, \dots, \bar{N}\}$, and $E \in \{0,1\}^{N \times \bar{N}}$ is an incidence matrix. We also introduce two sets of neighbors $\mathcal{N}_j$ and $\bar{\mathcal{N}}_i$ associated to the graph, defined as

$$\mathcal{N}_j = \{i \in [N] : \ E_{ij} = 1\} \quad \forall j \in [\bar{N}] \ \text{ and } \ \bar{\mathcal{N}}_i = \{j \in [\bar{N}] : \ E_{ij} = 1\} \quad \forall i \in [N].$$

The index sets $\mathcal{N}_j$ and $\bar{\mathcal{N}}_i$, which, e.g., in the context of network optimization may represent the set of sources which share the link $j \in [\bar{N}]$ and the set of links which are used by the source $i \in [N]$, respectively, describe the local information flow in the graph. We denote the entire vector of variables for the graph as $x \in \mathbb{R}^n$. The vector x can be partitioned accordingly in block components $x_i \in \mathbb{R}^{n_i}$, with $n = \sum_{i=1}^{N} n_i$. In order to easily extract subcomponents from the vector x, we consider a partition of the identity matrix $I_n = [U_1, \dots, U_N]$, with $U_i \in \mathbb{R}^{n \times n_i}$, such that $x_i = U_i^T x$ and matrices $U_{\mathcal{N}_i} \in \mathbb{R}^{n \times n_{\mathcal{N}_i}}$, such that $x_{\mathcal{N}_i} = U_{\mathcal{N}_i}^T x$, with $x_{\mathcal{N}_i}$ being the vector containing all the components x_j with $j \in \mathcal{N}_i$. In this paper we address problems arising from such systems, where the objective function can be written in a general form as

$$F^* = \min_{x \in \mathbb{R}^n} F(x) \quad \left(= \sum_{j=1}^{\bar{N}} f_j(x_{\mathcal{N}_j}) + \sum_{i=1}^{N} \Psi_i(x_i) \right), \tag{2.1}$$

where $f_j : \mathbb{R}^{n_{\mathcal{N}_j}} \to \mathbb{R}$ and $\psi_i : \mathbb{R}^{n_i} \to \mathbb{R}$. We denote $f(x) = \sum_{j=1}^{\bar{N}} f_j(x_{\mathcal{N}_j})$ and $\Psi(x) = \sum_{i=1}^{N} \Psi_i(x_i)$. The function $f(x)$ is a smooth *partially separable* convex function, while $\Psi(x)$ is a *fully separable* convex nonsmooth function. The local information structure imposed by the graph G should be considered as part of the problem formulation. We consider the following natural measure of separability of the objective function F:

$$(\omega, \ \bar{\omega}) = \left(\max_{j \in [\bar{N}]} |\mathcal{N}_j|, \ \max_{i \in [N]} |\bar{\mathcal{N}}_i| \right).$$

Note that $1 \le \omega \le N$, $1 \le \bar{\omega} \le \bar{N}$, and the definition of the measure of separability $(\omega, \ \bar{\omega})$ is more general than the one considered in [23] that is defined only in terms of ω. It is important to note that coordinate gradient descent type methods for solving problem (2.1) are appropriate where $\bar{\omega}$ is relatively small; otherwise incremental type

methods [21] should be considered for solving (2.1). Indeed, difficulties may arise when f is the sum of a large number of component functions and $\bar{\omega}$ is large since in that case exact computation of the components of gradient (i.e., $\nabla_i f(x) = \sum_{j \in \bar{\mathcal{N}}_i} \nabla_i f_j(x_{\mathcal{N}_j})$) can be either very expensive or impossible due to noise. In conclusion, we assume that the algorithm is employed for problems (2.1), with $(\bar{\omega}, \omega)$ relatively small, i.e., $\bar{\omega}, \omega \ll n$ (see section 2.1 and [11] for practical applications satisfying this condition).

By x^* we denote an optimal solution of problem (2.1) and by X^* the set of optimal solutions. We define the index and the set indicator functions as

$$\mathbf{1}_{\mathcal{N}_j}(i) = \begin{cases} 1 & \text{if } i \in \mathcal{N}_j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \mathbf{I}_X(x) = \begin{cases} 0 & \text{if } x \in X, \\ +\infty & \text{otherwise.} \end{cases}$$

Also, by $\|\cdot\|$ we denote the standard Euclidean norm and we introduce an additional norm $\|x\|_W^2 = x^T W x$, where $W \in \mathbb{R}^{n \times n}$ is a positive diagonal matrix. Considering these, we denote by $\Pi_X^W(x)$ the projection of a point x onto a set X in the norm $\|\cdot\|_W$, i.e., $\Pi_X^W(x) = \arg\min_{y \in X} \|y - x\|_W^2$. Furthermore, for simplicity of exposition, we denote by $\bar{x}$ the projection of a point x on the optimal set X^*, i.e., $\bar{x} = \Pi_{X^*}^W(x)$. In this paper we consider that the smooth component $f(x)$ of (2.1) satisfies the following assumption.

Assumption 1. We assume the functions $f_j(x_{\mathcal{N}_j})$ have $L_{\mathcal{N}_j}$-Lipschitz gradient:

$$\|\nabla f_j(x_{\mathcal{N}_j}) - \nabla f_j(y_{\mathcal{N}_j})\| \leq L_{\mathcal{N}_j} \|x_{\mathcal{N}_j} - y_{\mathcal{N}_j j}\| \quad \forall x_{\mathcal{N}_j}, y_{\mathcal{N}_j} \in \mathbb{R}^{n_{\mathcal{N}_j}}. \tag{2.2}$$

Note that our assumption is different from those in [18, 9, 15, 23, 7], where the authors consider that the gradient of the function f is coordinatewise Lipschitz continuous, which states the following: if we define the partial gradient $\nabla_i f(x) = U_i^T \nabla f(x)$, then there exist some constants $L_i > 0$ such that

$$\|\nabla_i f(x + U_i y_i) - \nabla_i f(x)\| \leq L_i \|y_i\| \quad \forall x \in \mathbb{R}^n,\ y_i \in \mathbb{R}^{n_i}. \tag{2.3}$$

As a consequence of Assumption 1 we have that [17]

$$f_j(x_{\mathcal{N}_j} + y_{\mathcal{N}_j}) \leq f_j(x_{\mathcal{N}_j}) + \langle \nabla f_j(x_{\mathcal{N}_j}), y_{\mathcal{N}_j} \rangle + \frac{L_{\mathcal{N}_j}}{2} \|y_{\mathcal{N}_j}\|^2. \tag{2.4}$$

From Assumption 1 we derive the following descent lemma, which is central in our derivation of a parallel coordinate descent method and proving its convergence rate.

LEMMA 2.1. *Under Assumption* 1 *the following holds for* $f(x) = \sum_{j=1}^N f_j(x_{\mathcal{N}_j})$*:*

$$f(x + y) \leq f(x) + \langle \nabla f(x), y \rangle + \frac{1}{2} \|y\|_W^2 \quad \forall x, y \in \mathbb{R}^n, \tag{2.5}$$

where $W \succ 0$ *is block-diagonal with its blocks* $W_{ii} \in \mathbb{R}^{n_i \times n_i}$, $W_{ii} = \sum_{j \in \bar{\mathcal{N}}_i} L_{\mathcal{N}_j} I_{n_i}$, $i \in [N]$.

Proof. If we sum up (2.4) for $j \in [\bar{N}]$ and use the definition of f, we have that

$$f(x + y) \leq f(x) + \sum_{j=1}^{\bar{N}} \left[\langle \nabla f_j(x_{\mathcal{N}_j}), y_{\mathcal{N}_j} \rangle + \frac{L_{\mathcal{N}_j}}{2} \|y_{\mathcal{N}_j}\|^2 \right]. \tag{2.6}$$

Given matrices $U_{\mathcal{N}_j}$, we can express the first term in the right-hand side as follows:

$$\sum_{j=1}^{\bar{N}} \left\langle \nabla f_j(x_{\mathcal{N}_j}), y_{\mathcal{N}_j} \right\rangle = \sum_{j=1}^{\bar{N}} \left\langle \nabla f_j(x_{\mathcal{N}_j}), U_{\mathcal{N}_j}^T y \right\rangle = \sum_{j=1}^{\bar{N}} \left\langle U_{\mathcal{N}_j} \nabla f_j(x_{\mathcal{N}_j}), y \right\rangle = \langle \nabla f(x), y \rangle.$$

Note that since W is a diagonal matrix we can express the norm $\|\cdot\|_W$ as $\|y\|_W^2 = \sum_{i=1}^N \left(\sum_{j\in\bar{\mathcal{N}}_i} L_{\mathcal{N}_j}\right)\|y_i\|^2$. From the definition of $\mathcal{N}_j$ and $\bar{\mathcal{N}}_i$, note that $\mathbf{1}_{\mathcal{N}_j}(i)$ is equivalent to $\mathbf{1}_{\bar{\mathcal{N}}_i}(j)$. Thus, for the final term of the right-hand side of (2.6) we have

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{\bar{N}} L_{\mathcal{N}_j}\|y_{\mathcal{N}_j}\|^2 &= \frac{1}{2}\sum_{j=1}^{\bar{N}} L_{\mathcal{N}_j}\sum_{i\in\mathcal{N}_j}\|y_i\|^2 = \frac{1}{2}\sum_{j=1}^{\bar{N}} L_{\mathcal{N}_j}\sum_{i=1}^{N}\|y_i\|^2\mathbf{1}_{\mathcal{N}_j}(i)\\ &= \frac{1}{2}\sum_{i=1}^{N}\|y_i\|^2\sum_{j=1}^{\bar{N}} L_{\mathcal{N}_j}\mathbf{1}_{\bar{\mathcal{N}}_i}(j) = \frac{1}{2}\sum_{i=1}^{N}\|y_i\|^2\sum_{j\in\bar{\mathcal{N}}_i} L_{\mathcal{N}_j} = \frac{1}{2}\|y\|_W^2,\end{aligned}$$

which proves the statement of the lemma. ∎

Note that the convergence results of this paper hold for any descent lemma in the form (2.5), and thus the expression of the matrix W above can be replaced with any other block-diagonal matrix $W \succ 0$ for which (2.5) is valid. Based on (2.3) an inequality similar to that in (2.5) can be derived, but the matrix W is replaced in this case with the matrix $\omega W' = \omega\mathrm{diag}(L_i I_{n_i}; i \in [N])$. These differences in the matrices will lead to different step sizes in the algorithms of our paper and of, e.g., [10, 23]. The following relation establishes Lipschitz continuity for ∇f but in the norm $\|\cdot\|_W$, whose proof can be derived using arguments similar to those in [17]:

$$\|\nabla f(x) - \nabla f(y)\|_{W^{-1}} \le \|x - y\|_W \quad \forall x, y \in \mathbb{R}^n. \tag{2.7}$$

2.1. Motivating practical applications. We now present important applications from which the interest for problems of type (2.1) stems. One application is found in data mining or machine learning [27], where we must solve a sparse problem:

$$\min_{x\in\mathbb{R}^n} f(x) + \lambda\|x\|_1,$$

where $\lambda > 0$, $\|\cdot\|_1$ denotes the 1-norm, and $f(x)$ is the loss function. For example, in the sparse logistic regression, the average logistic loss function is $f(x) = \sum_{j=1}^{\bar{N}} f_j(x_{\mathcal{N}_j}) = \frac{1}{\bar{N}}\sum_{j=1}^{\bar{N}} \log\left(1 + \exp\left(-b^j\langle a^j, x\rangle\right)\right)$, where the vectors $a^j \in \mathbb{R}^n$ represent $\bar{N}$ samples, and b^j represent the binary class labels with $b^j \in \{-1, +1\}$. Note that $\Psi(x) = \lambda\|x\|_1$ is the separable nonsmooth component which promotes the sparsity of the decision variable x. If we associate to this problem a bipartite graph G where the incidence matrix E is defined such that $E_{ij} = 1$ provided that $a_i^j \neq 0$, then the vectors a^j have a certain sparsity according to this graph; i.e., they only have nonzero components in $a^j_{\mathcal{N}_j}$. It can be easily proven that the objective function f in this case satisfies (2.2) with $L_{\mathcal{N}_j} = \sum_{l\in\mathcal{N}_j}\|a_l^j\|^2/4$ and (2.3) with $L_i = \sum_{j\in\bar{\mathcal{N}}_i}\|a_i^j\|^2/4$.

Another classical problem which implies functions f_j of type (2.2) is

$$\min_{x_i\in X_i\subseteq\mathbb{R}^{n_i}} F(x) \quad \left(= \frac{1}{2}\|Ax - b\|^2 + \sum_{i=1}^{N}\lambda_i\|x_i\|_1\right), \tag{2.8}$$

where $A \in \mathbb{R}^{\bar{N}\times n}$, the sets X_i are convex, $n = \sum_{i=1}^N n_i$, and $\lambda > 0$. This problem is also known as the constrained lasso problem and is widely used in, e.g., signal processing, fused or generalized lasso, and monotone curve estimation [4]. For example, in image restoration, incorporating a priori information (such as box constraints on x) can lead to substantial improvements in the restoration and reconstruction process (see [4] for more details). Note that this problem is a special case of problem

(2.1), with $\Psi(x) = \sum_{i=1}^{N}[\lambda_i\|x_i\|_1 + \mathbf{I}_{X_i}(x_i)]$ being block separable and with f_j defined as $f_j(x_{\mathcal{N}_j}) = \frac{1}{2}(a_{\mathcal{N}_j}^T x_{\mathcal{N}_j} - b_j)^2$, where $a_{\mathcal{N}_j}$ are the nonzero components of row j of A, corresponding to $\mathcal{N}_j$. In this application the functions f_j satisfy (2.2) with Lipschitz constants $L_{\mathcal{N}_j} = \|a_{\mathcal{N}_j}\|^2$. Given these constants, we find that f in this case satisfies (2.5) with the matrix $W = \operatorname{diag}\big(\sum_{j\in\bar{\mathcal{N}}_i}\|a_{\mathcal{N}_j}\|^2 I_{n_i};\ i \in [N]\big)$. Also, note that functions of type (2.8) satisfy Lipschitz continuity (2.3) with $L_i = \|A_i\|^2$, where $A_i \in \mathbb{R}^{\bar{N}\times n_i}$ denotes block column i of the matrix A.

A third problem which falls under the same category is derived from the primal

$$f^* = \min_{u\in\mathbb{R}^m} \sum_{j=1}^{\bar{N}} g_j(u_j) \qquad \text{subject to (s.t.)} \quad Au \le b, \tag{2.9}$$

where $A \in \mathbb{R}^{n\times m}$, $u_j \in \mathbb{R}^{m_j}$, and the functions g_j are strongly convex with convexity parameters σ_j. This type of problem is often found in network control [16], network optimization, or utility maximization [21]. We formulate the dual problem of (2.9) as

$$f^* = \max_{x\in\mathbb{R}^n} \sum_{j=1}^{\bar{N}} -g_j^*(-A_j^T x) - \langle x, b\rangle - \Psi(x), \tag{2.10}$$

where $A_j \in \mathbb{R}^{n\times m_j}$ is the jth block column of A, x denotes the Lagrange multiplier, $\Psi(x) = \mathbf{I}_{\mathbb{R}^n_+}(x)$ is the indicator function for the nonnegative orthant $\mathbb{R}^n_+$, and $g_j^*(z)$ is the convex conjugate of the function $g_j(u_j)$. Note that, given the strong convexity of $g_j(u_j)$, the functions $g_j^*(z)$ have Lipschitz continuous gradient in z of type (2.2) with constants $\frac{1}{\sigma_j}$ [17]. Now, if the matrix A has some sparsity induced by a graph, i.e., the blocks $A_{ij} = 0$ if the corresponding incidence matrix has $E_{ij} = 0$, which in turn implies that the block columns A_j are sparse according to some index set $\mathcal{N}_j$, then the matrix-vector products $A_j^T x$ depend only on $x_{\mathcal{N}_j}$, such that $f_j(x_{\mathcal{N}_j}) = -g_j^*\big(-A_{\mathcal{N}_j}^T x_{\mathcal{N}_j}\big) - \langle x_{\mathcal{N}_j}, \bar{b}_{\mathcal{N}_j}\rangle$, with $\sum_j \langle x_{\mathcal{N}_j}, \bar{b}_{\mathcal{N}_j}\rangle = \langle x, b\rangle$. Then, f_j has Lipschitz continuous gradient of type (2.2) with $L_{\mathcal{N}_j} = \frac{\|A_{\mathcal{N}_j}\|^2}{\sigma_j}$. For this problem we also have componentwise Lipschitz continuous gradient of type (2.3) with $L_i = \sum_{j\in\bar{\mathcal{N}}_i} \frac{\|A_{ij}\|^2}{\sigma_j}$. Note that there are many applications in the form (2.9) with matrix A given as a column-linked block-angular form for which $\omega = \bar{N}$ and $\bar{\omega} = 2$ is small (see the extended report [11] of this paper for concrete examples).

3. Parallel random coordinate descent method. In this section we employ a parallel version of the random coordinate descent method [18, 24, 7], which we call **P-RCD**. Analysis of coordinate descent methods based on updating in parallel more than one (block) component per iteration was given in, e.g., [14, 10, 23, 22, 20]. Before we discuss the method, however, we first need to introduce some concepts. For a function $F(x)$ as defined in (2.1), we introduce the following mapping in the norm $\|\cdot\|_W$:

$$t_{[N]}(x, y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\|y - x\|_W^2 + \Psi(y). \tag{3.1}$$

Note that the mapping $t_{[N]}(x, y)$ is a fully separable and strongly convex in y w.r.t. the norm $\|\cdot\|_W$ with the constant 1. We denote by $T_{[N]}(x)$ the proximal step for function $F(x)$, which is the optimal point of the mapping $t_{[N]}(x, y)$, i.e.,

$$T_{[N]}(x) = \arg\min_{y\in\mathbb{R}^n} t_{[N]}(x, y). \tag{3.2}$$

The proximal step $T_{[N]}(x)$ can also be defined via the proximal operator of Ψ:

$$\text{prox}_{\Psi}(x) = \arg\min_{u\in\mathbb{R}^n} \Psi(u) + \frac{1}{2}\|u-x\|_W^2.$$

We recall an important property of the proximal operator [26]:

$$\|\text{prox}_{\Psi}(x) - \text{prox}_{\Psi}(y)\|_W \leq \|x-y\|_W. \tag{3.3}$$

Based on this proximal operator, note that we can write

$$T_{[N]}(x) = \text{prox}_{\Psi}(x - W^{-1}\nabla f(x)). \tag{3.4}$$

Given that $\Psi(x)$ is generally not differentiable, we denote by $\partial_\Psi(x)$ a vector belonging to the set of subgradients of $\Psi(x)$. Evidently, in both definitions, the optimality conditions of the resulting problem from which we obtain $T_{[N]}(x)$ are the same, i.e.,

$$0 \in \nabla f(x) + W(T_{[N]}(x) - x) + \partial_\Psi(T_{[N]}(x)). \tag{3.5}$$

It will become evident further on that the optimal solution $T_{[N]}(x)$ will play a crucial role in the parallel random coordinate descent method. We now establish some properties which involve the function $F(x)$, the mapping $t_{[N]}(x,y)$, and the proximal step $T_{[N]}(x)$. Given that $t_{[N]}(x,y)$ is strongly convex in y and that $T_{[N]}(x)$ is an optimal point when minimizing over y, we have the following inequality:

$$F(x) - t_{[N]}(x, T_{[N]}(x)) = t_{[N]}(x,x) - t_{[N]}(x, T_{[N]}(x)) \geq \frac{1}{2}\|x - T_{[N]}(x)\|_W^2. \tag{3.6}$$

Further, given that f is convex and by the definition of $t_{[N]}(x,y)$, we get

$$t_{[N]}(x, T_{[N]}(x)) \leq \min_{y\in\mathbb{R}^n} F(y) + \frac{1}{2}\|y-x\|_W^2. \tag{3.7}$$

In the algorithm that we discuss, at a step k, the (block) components of the iterate x^k which are to be updated are dictated by a set of indices $J^k \subseteq [N]$ which is randomly chosen. Let us denote by $x_J \in \mathbb{R}^n$ the vector whose blocks x_i, with $i \in J \subseteq [N]$, are identical to those of x, while the remaining blocks are zeroed out, i.e., $x_J = \sum_{i\in J} U_i x_i$. Also, for the separable function $\Psi(x)$, we denote the partial sum $\Psi_J(x) = \sum_{i\in J} \Psi_i(x_i)$ and the vector $\partial_J\Psi(x) = [\partial\Psi(x)]_J \in \mathbb{R}^n$. A random variable J is uniquely characterized by the probability density function

$$P_{\hat{J}} = P(J = \hat{J}),$$

where $\hat{J} \subseteq [N]$.

For the random variable J, we also define the probability with which a subcomponent $i \in [N]$ can be found in J as $p_i = \mathbb{P}(i \in J)$. In our algorithm, we consider a uniform sampling of τ unique coordinates i, $1 \leq \tau \leq N$ that make up J, i.e., $|J| = \tau$. For a random variable J with $|J| = \tau$, we observe that we have a total number of $\binom{N}{\tau}$ possible values that J can take, and with the uniform sampling we have that $P_J = \frac{1}{\binom{N}{\tau}}$. Given that J is random, we can express the probability that $i \in J$ as $p_i = \sum_{J:\, i\in J} P_J$. For a single index i, note that we have a total number of $\binom{N-1}{\tau-1}$

possible sets that J can take which will include i, and therefore the probability that this index is included in J is

$$p_i = \frac{\binom{N-1}{\tau-1}}{\binom{N}{\tau}} = \frac{\tau}{N}. \tag{3.8}$$

We can also consider other ways in which J can be chosen; however, due to space limitations we restrict our presentation to uniform sampling. Having defined the proximal step as $T_{[N]}(x^k)$ in (3.2), in the algorithm that follows we generate randomly at step k an index set J^k of cardinality $1 \leq \tau \leq N$. We denote the vector $T_{J^k}(x^k) = [T_{[N]}(x^k)]_{J^k}$ which will be used to update x^{k+1}, i.e., in the sense that $[x^{k+1}]_{J^k} = T_{J^k}(x^k)$. Also, by $\bar{J}^k$ we denote the complement set of J^k, i.e., $\bar{J}^k = \{i \in [N] : i \notin J^k\}$. Thus, the parallel algorithm we propose below consists of the following steps (note that our algorithm is similar to those studied in, e.g., [10, 14, 23] but with different step sizes):

Distributed and parallel random coordinate descent method (P-RCD)

1. Consider an initial point $x^0 \in \mathbb{R}^n$ and $1 \leq \tau \leq N$. For $k \geq 0$:
2. Generate with uniform probability a random set of indices $J^k \subseteq [N]$, with $|J^k| = \tau$.
3. Compute the update: $x^{k+1}_{J^k} = T_{J^k}(x^k)$ and $x^{k+1}_{\bar{J}^k} = x^k_{\bar{J}^k}$.

Clearly, the optimization problem from which we compute the iterate of **(P-RCD)** is fully separable. Then, it follows that for updating component $i \in J^k$ of x^{k+1} we need the following data: $\Psi_i(x^k_i), W_{ii}$, and $\nabla_i f = \sum_{j \in \bar{N}_i} \nabla_i f_j$. Therefore, if algorithm **(P-RCD)** runs on a multicore machine or as a multithread process, it can be observed that component updates can be done in parallel by each core/thread using the communication graph G. Note that the iterate update of the **(P-RCD)** method can also be expressed in the following way:

$$\begin{cases} x^{k+1} = x^k + T_{J^k}(x^k) - x^k_{J^k} = \operatorname{prox}_{\Psi_{J^k}}\left(x^k - W^{-1}\nabla_{J^k} f(x^k)\right), \\ x^{k+1} = \arg\min_{y \in \mathbb{R}^n} \langle \nabla_{J^k} f(x^k), y - x^k \rangle + \frac{1}{2}\|y - x^k\|^2_W + \Psi_{J^k}(y). \end{cases} \tag{3.9}$$

Note that the right-hand sides of the last two equalities contain the same optimization problem, whose optimality conditions are

$$W[x^k - x^{k+1}]_{J^k} \in \nabla_{J^k} f(x^k) + \partial \Psi_{J^k}(x^{k+1}) \quad \text{and} \quad [x^{k+1}]_{\bar{J}^k} = [x^k]_{\bar{J}^k}. \tag{3.10}$$

We now establish that method **(P-RCD)** is a descent method, i.e., $F(x^{k+1}) \leq F(x^k)$ for all $k \geq 0$. From the convexity of $\Psi(\cdot)$ and (2.5) we obtain the following:

$$\begin{aligned} F(x^{k+1}) &\leq F(x^k) + \langle \nabla_{J^k} f(x^k) + \partial_{\Psi_{J^k}}(x^{k+1}), [x^{k+1} - x^k]_{J^k} \rangle + \frac{1}{2}\|x^{k+1} - x^k\|^2_W \\ &\overset{(3.10)}{=} F(x^k) + \langle W[x^k - x^{k+1}]_{J^k}, [x^{k+1} - x^k]_{J^k} \rangle + \frac{1}{2}\|x^{k+1} - x^k\|^2_W \\ &= F(x^k) - \frac{1}{2}\|x^{k+1} - x^k\|^2_W. \end{aligned} \tag{3.11}$$

With **(P-RCD)** being a descent method, we can now introduce the term

$$R_W(x^0) = \max_{x:\, F(x) \leq F(x^0)} \min_{x^* \in X^*} \|x - x^*\|_W \tag{3.12}$$

and assume it to be bounded. We also define the random variable comprising the whole history of previous events as

$$\eta^k = \{J^0, \ldots, J^k\}.$$

4. Sublinear convergence for smooth convex minimization. In this section we establish the sublinear convergence rate of method **(P-RCD)** for problems of type (2.1) with the objective function satisfying Assumption 1. First we recall a basic relation from probability theory proven in, e.g., [23, Lemma 3]: let there be some constants θ_i with $i = 1, \dots, N$ and a sampling J chosen as described above, and define the sum $\sum_{i\in J} \theta_i$; then the expected value of the sum satisfies

$$\mathbb{E}\left[\sum_{i\in J} \theta_i\right] = \sum_{i=1}^{N} p_i \theta_i. \tag{4.1}$$

For any vector $d \in \mathbb{R}^n$ we consider its counterpart d_J for a sampling J taken as described above. Given the previous relation and by taking into account the separability of the inner product and of the squared norm $\|\cdot\|_W^2$, it follows immediately that

$$\mathbb{E}\left[\langle x, d_J\rangle\right] = \frac{\tau}{N}\langle x, d\rangle \quad \text{and} \quad \mathbb{E}\left[\|d_J\|_W^2\right] = \frac{\tau}{N}\|d\|_W^2. \tag{4.2}$$

Based on relations (4.2), the separability of the function $\Psi(x)$, and the properties of the expectation operator, the following inequalities can be immediately derived [23]:

$$\mathbb{E}\left[\Psi(x + d_J)\right] = \frac{\tau}{N}\Psi(x + d) + \left(1 - \frac{\tau}{N}\right)\Psi(x), \tag{4.3}$$

$$\mathbb{E}\left[F(x + d_J)\right] \le \left(1 - \frac{\tau}{N}\right)F(x) + \frac{\tau}{N}t_{[N]}(x, d). \tag{4.4}$$

By the definition of the operator $t_{[N]}(x, y)$, the convexity of f and Ψ, and the optimality conditions (3.5), we have the following inequalities:

$$\begin{aligned} t_{[N]}(x, T_{[N]}(x)) &\le f(y) + \langle \nabla f(x), x - y\rangle + \langle \nabla f(x), T_{[N]}(x) - x\rangle \\ &\quad + \frac{1}{2}\|T_{[N]}(x) - x\|_W^2 + \Psi(y) + \left\langle \partial\Psi(T_{[N]}(x)), T_{[N]}(x) - y\right\rangle \\ &\overset{(3.5)}{\le} f(y) + \langle \nabla f(x), x - y\rangle + \langle \nabla f(x), T_{[N]}(x) - x\rangle + \frac{1}{2}\|T_{[N]}(x) - x\|_W^2 \\ &\quad + \Psi(y) + \left\langle -\nabla f(x) - W\left(T_{[N]}(x) - x\right), T_{[N]}(x) - y\right\rangle \\ &= F(y) - \left\langle W\left(T_{[N]}(x) - x\right), x - y\right\rangle - \frac{1}{2}\|T_{[N]}(x) - x\|_W^2. \end{aligned} \tag{4.5}$$

This property will prove useful in the following theorem, which provides the sublinear convergence rate for method **(P-RCD)**.

THEOREM 4.1. *If Assumption* 1 *holds and considering that* $R_W(x^0)$ *defined in* (3.12) *is bounded, then for the sequence* x^k *generated by algorithm* ***(P-RCD)*** *we have*

$$\mathbb{E}[F(x^k)] - F^* \le \frac{N\left(1/2(R_W(x^0))^2 + F(x^0) - F^*\right)}{\tau k + N}. \tag{4.6}$$

Proof. Our proof uses the tools developed above (see [23] for more general settings) and generalizes the proof of Theorem 1 in [7] from one component update per iterate to the case of τ component updates, based on uniform sampling and on Assumption 1, and consequently on a different descent lemma. Thus, by taking expectation in both sides of (3.11) w.r.t. J^k conditioned on η^{k-1} we arrive at

$$\mathbb{E}[F(x^{k+1})] \le F(x^k) - \frac{1}{2}\mathbb{E}\left[\|x^{k+1} - x^k\|_W^2\right] \le F(x^k). \tag{4.7}$$

Now, if we take $x = x^k$, $J = J^k$, and $d_{J^k} = T_{J^k}(x^k) - x^k_{J^k}$ in (4.4), we get

$$\mathbb{E}\left[F(x^{k+1})\right] \leq \left(1 - \frac{\tau}{N}\right) F(x^k) + \frac{\tau}{N} t_{[N]}\left(x^k, T_{[N]}(x^k)\right). \tag{4.8}$$

From this and (4.5) we obtain

$$\frac{\tau}{N}F(y) + \frac{N-\tau}{N}F(x^k) \geq \mathbb{E}\left[F(x^{k+1})\right] + \frac{\tau}{N}\left\langle W\left(T_{[N]}(x^k) - x^k\right), x^k - y\right\rangle + \frac{\tau}{2N}\|T_{[N]}(x^k) - x^k\|^2_W. \tag{4.9}$$

Denote $r^k = \|x^k - x^*\|_W$. From the definition of x^{k+1} we have that

$$(r^{k+1})^2 = (r^k)^2 + \sum_{i \in J^k} \left[2W_{ii}\langle T_i(x^k) - x^k_i, x^k_i - x^*_i\rangle + W_{ii}\|T_i(x^k) - x^k_i\|^2\right].$$

If we divide both sides of the above inequality by 2 and take expectation, we obtain

$$\mathbb{E}\left[\frac{1}{2}(r^{k+1})^2\right] = \frac{(r^k)^2}{2} + \frac{\tau}{N}\left\langle W\left(T_{[N]}(x^k) - x^k\right), x^k - x^*\right\rangle + \frac{\tau}{2N}\|T_{[N]}(x^k) - x^k\|^2_W.$$

Through this inequality and (4.9) we arrive at

$$\mathbb{E}\left[\frac{1}{2}(r^{k+1})^2\right] \leq \frac{(r^k)^2}{2} + \frac{\tau}{N}F^* + \frac{N-\tau}{N}F(x^k) - \mathbb{E}\left[F(x^{k+1})\right].$$

After some rearranging of terms we obtain the following inequality:

$$\mathbb{E}\left[\frac{1}{2}(r^{k+1})^2 + F(x^{k+1}) - F^*\right] \leq \left(\frac{(r^k)^2}{2} + F(x^k) - F^*\right) - \frac{\tau}{N}(F(x^k) - F^*).$$

By applying this inequality repeatedly, by taking expectation over η^{k-1}, and from the fact that $\mathbb{E}[F(x^k)]$ is decreasing from (4.7), we obtain the following:

$$\mathbb{E}\left[F(x^{k+1})\right] - F^* \leq \mathbb{E}\left[\frac{1}{2}(r^{k+1})^2 + F(x^{k+1}) - F^*\right] \leq \frac{(r^0)^2}{2} + F(x^0) - F^*$$

$$-\frac{\tau}{N}\sum_{j=0}^{k}\left(\mathbb{E}\left[F(x^j)\right] - F^*\right) \leq \frac{(r^0)^2}{2} + F(x^0) - F^* - \frac{\tau(k+1)}{N}\left(\mathbb{E}\left[F(x^{k+1})\right] - F^*\right).$$

Rearranging some items and since $(r^0)^2 \leq (R_W(x^0))^2$, we arrive at (4.6). ∎

We notice that the sublinear convergence rate (4.6) of order $\mathcal{O}(N/\tau k)$ depends linearly on the choice of $\tau = |J|$, so that if the algorithm is implemented on a cluster, then τ reflects the available number of cores. Furthermore, given a suboptimality level ϵ and a confidence level $0 < \rho < 1$, using standard arguments as in [18, 23] we can easily establish a total number of iterations k^ϵ_ρ which will ensure an ϵ-suboptimal solution with probability at least $1 - \rho$. More precisely, for the iterates generated by algorithm **(P-RCD)** and a k^ϵ_ρ that satisfies

$$k^\epsilon_\rho \geq \frac{c}{\epsilon}\left(1 + \log\left(\frac{N}{\tau}\frac{\left(R_W(x^0)\right)^2 + 2\left(F(x^0) - F^*\right)}{4c\rho}\right)\right) + 2 - N, \tag{4.10}$$

with $c = \frac{2N}{\tau} \max\left\{(R_W(x^0))^2, F(x^0) - F^*\right\}$, we get the following result in probability:

$$\mathbb{P}\left(F\left(x^{k^\epsilon_\rho}\right) - F^* \leq \epsilon\right) \geq 1 - \rho.$$

We notice that in the smooth case, given the choice of τ, we obtain different sublinear convergence results of order $\mathcal{O}(1/k)$. For example, for $\tau = 1$ we obtain a sublinear convergence rate simular to that of the random coordinate descent method in [9, 18, 7, 24], i.e., of the form $\mathcal{O}(NR_W^2/k)$, while for $\tau = N$ we get a convergence rate similar to that of the full composite gradient method of [19]. However, the distances are measured in different norms in these papers. For example, when $\tau = N$ the comparison of convergence rates in our paper and [19] is reduced to comparing the quantities $L_f R(x^0)^2$ of [19] with our $R_W(x^0)^2$, where L_f is the Lipschitz constant of the smooth component of the objective function, i.e., of f, while $R(x^0)$ is defined in a fashion similar to our $R_W(x^0)$ but in the Euclidean norm instead of the norm $\|\cdot\|_W$. Let us now consider the two extreme cases. First, consider the smooth component of the objective function as follows: $f(x) = \sum_{j=1}^{\bar{N}} f_j(x_j)$. In this case, it can be seen that $L_f = \max_{j\in[\bar{N}]} L_{\mathcal{N}_j}$. Thus considering the definition of the matrix $W = \operatorname{diag}(L_{\mathcal{N}_j};\ j \in [\bar{N}])$ in Lemma 2.1 we have that $L_f\|x^0 - x^*\|^2 \geq \|x^0 - x^*\|_W^2$; i.e., our convergence rate is usually better. On the other hand, if we have f defined as $f(x) = \sum_{j=1}^{\bar{N}} f_j(x)$, then it can be easily proven that $L_f = \sum_{j\in\bar{N}} L_{\mathcal{N}_j}$ and $W = L_f I_n$ and the quantities $L_f R(x^0)^2$ and $R_W(x^0)^2$ would be the same. Thus, we get better rates of convergence when $\bar{\omega} < \bar{N}$. Furthermore, our sublinear convergence results are also similar to those of [23], but they are obtained under further knowledge regarding the objective function and with a modified analysis. In particular, using a reasoning similar to that in [7], we can argue, based on our analysis, that the expected value type of convergence rate given in (4.6) has better constants than that in [23] under certain separability properties given below. For example, the convergence rate of the algorithm, apart from essentially being of order $\mathcal{O}\left(\frac{1}{k}\right)$, depends on the step sizes involved when computing the next iterate x^{k+1}; see (3.9). Thus, let us compare the weighted step sizes $W = \operatorname{diag}(W_{ii})$ in our algorithm **(P-RCD)** and those in [23]. To this purpose, let us consider the smooth component in (2.1) in the form $f(x) = \frac{1}{2}\|Ax - b\|^2$ and $n_i = 1$. Under these considerations, we observe from (4.11) below that our step sizes are better than those in [23] as τ increases and $\bar{\omega} \ll \omega$:

$$W_{ii} = \sum_{j:i\in\mathcal{N}_j} \sum_{t=1}^{n} A_{jt}^2 \quad \text{and} \quad W_{ii}^{[16]} = \sum_{j=1}^{\bar{N}} \beta A_{ji}^2, \tag{4.11}$$

where $\beta = 1 + \frac{(\omega-1)(\tau-1)}{\max\{1,n-1\}}$ or $\beta = \min(\omega, \tau)$ depending on whether or not monotonicity is enforced in the algorithm of [23].

5. Linear convergence for error bound convex minimization. In this section we prove that, for certain minimization problems, the sublinear convergence rate of **(P-RCD)** from the previous section can be improved to a linear convergence rate. In particular, we prove that under additional assumptions on the objective function, which are often satisfied in practical applications (e.g., the dual of a linearly constrained smooth convex problem, a control problem, or a constrained lasso problem), we have a *generalized error bound property* for our optimization problem. In these settings we are able to provide for the first time *global* linear convergence rate for algorithm **(P-RCD)**, as opposed to the results in [8, 28], where only *local* linear convergence was derived for deterministic descent methods, or the results in [29], where

global linear convergence is proved for gradient type methods but applied only to problems where Ψ is the set indicator function of a polyhedron. Therefore, we proceed by introducing the proximal gradient mapping of the objective function $F(x)$:

$$\nabla^+ F(x) = x - \text{prox}_\Psi \left(x - W^{-1}\nabla f(x)\right). \tag{5.1}$$

Clearly, a point x^* is an optimal solution of the original problem (2.1) if and only if $\nabla^+ F(x^*) = 0$. In the following definition we introduce the new concept of *generalized error bound property* for problem (2.1).

DEFINITION 5.1. *Problem* (2.1) *has the generalized error bound property* ((GEBP) property) *w.r.t. the norm* $\|\cdot\|_W$ *if there exist two nonnegative constants* κ_1 *and* κ_2 *such that the composite objective function* F *satisfies the following relation (we use* $\bar{x} = \Pi^W_{X^*}(x)$*):*

$$\|x - \bar{x}\|_W \leq \left(\kappa_1 + \kappa_2 \|x - \bar{x}\|_W^2\right) \|\nabla^+ F(x)\|_W \quad \forall x \in \mathbb{R}^n. \tag{5.2}$$

Note that the class of problems introduced in Definition 5.1 includes other known categories of problems, e.g., problems with objective functions F composed of a smooth strongly convex function f with a constant σ_W w.r.t. the norm $\|\cdot\|_W$ and a general convex function Ψ satisfying our definition (5.2) with $\kappa_1 = \frac{2}{\sigma_W}$ and $\kappa_2 = 0$; or problems satisfying the classical error bound property [8, 28, 29], i.e., $\|x - \bar{x}\|_W \leq \kappa \|\nabla^+ F(x)\|_W$ for all $x \in \mathbb{R}^n$, satisfying our definition (5.2) with $\kappa_1 = \kappa$ and $\kappa_2 = 0$ (see section 6 for more details and for other classes of problems (2.1) satisfying the (GEBP) property).

Next, we prove that on optimization problems having (GEBP) property (5.2) our algorithm (**P-RCD**) has global linear convergence. Our analysis will employ ideas from the convergence proof of deterministic descent methods in [28]. However, the random nature of our method and the nonsmooth property of the objective function require a new approach. For example, the typical proof for linear convergence of gradient descent type methods for solving convex problems with an error bound like property is based on deriving an inequality of the form $F(x^{k+1}) - F^* \leq c\|x^{k+1} - x^k\|$ (see, e.g., [8, 28, 29]). Under our settings, we cannot derive this type of inequality, but instead we obtain a weaker inequality, where we replace $\|x^{k+1} - x^k\|$ with another term, which still allows us to prove linear convergence. We start with the following lemma, which shows an important property of algorithm **(P-RCD)** when it is applied to problems (2.1) having generalized error bound objective functions.

LEMMA 5.2. *If problem* (2.1) *satisfies* (GEBP) *given in* (5.2)*, then a point* x^k *generated by algorithm* **(P-RCD)** *and its projection onto* X^**, denoted by* $\bar{x}^k$*, satisfy*

$$\|x^k - \bar{x}^k\|_W^2 \leq \left(\kappa_1 + \kappa_2 \|x^k - \bar{x}^k\|_W^2\right)^2 \frac{N}{\tau} \mathbb{E}\left[\|x^{k+1} - x^k\|_W^2\right]. \tag{5.3}$$

Proof. For the iteration defined by algorithm **(P-RCD)** we have

$$\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^k\|_W^2\right] &= \mathbb{E}\left[\|x^k + T_{J^k}(x^k) - x^k_{J^k} - x^k\|_W^2\right] \\
&= \mathbb{E}\left[\|x^k_{J^k} - T_{J^k}(x^k)\|_W^2\right] \overset{(4.2)}{=} \frac{\tau}{N}\|x^k - T_{[N]}(x^k)\|_W^2 \\
&= \frac{\tau}{N}\|x^k - \text{prox}_\Psi(x^k - W^{-1}\nabla f(x^k))\|_W^2 = \frac{\tau}{N}\|\nabla^+ F(x^k)\|_W^2.
\end{aligned}$$

Through this equality and (5.2) we have that

$$\begin{aligned}
\|x^k - \bar{x}^k\|_W^2 &\leq (\kappa_1 + \kappa_2\|x^k - \bar{x}^k\|_W^2)^2 \|\nabla F^+(x^k)\|_W^2 \\
&\leq (\kappa_1 + \kappa_2\|x^k - \bar{x}^k\|_W^2)^2 \frac{N}{\tau}\mathbb{E}\left[\|x^{k+1} - x^k\|_W^2\right],
\end{aligned} \tag{5.4}$$

and the proof is complete. ∎

Remark 1. Note that if the iterates of an algorithm satisfy $\|x^k - x^*\| \le \|x^0 - x^*\|$ for all $k \ge 0$ (see, e.g., the case of the full gradient method [17]), then by taking $\bar{\kappa}(x^0) = \left(\kappa_1 + \kappa_2 \|x^0 - x^*\|_W^2\right)^2$ we have the inequality

$$\|x^k - \bar{x}^k\|_W^2 \le \frac{\bar{\kappa}(x^0)N}{\tau}\mathbb{E}\left[\|x^{k+1} - x^k\|_W^2\right] \quad \forall k \ge 0. \tag{5.5}$$

On the other hand, if the iterates of an algorithm satisfy (3.12) with $R_W(x^0)$ bounded (see, e.g., the case of our algorithm **(P-RCD)**, which is a descent method, as proven in (3.11)), then (5.5) is satisfied with $\bar{\kappa}(x^0) = (\kappa_1 + \kappa_2(R_W(x^0))^2)^2$.

Let us now note that given the separability of function $\Psi : \mathbb{R}^n \to \mathbb{R}$, then for any vector $d \in \mathbb{R}^n$, if we consider their counterparts Ψ_J and d_J for a sampling J taken as described above, the expected value $\mathbb{E}[\Psi_J(d_J)]$ satisfies

$$\begin{aligned}\mathbb{E}[\Psi_J(d_J)] &= \sum_{J\subseteq[N]}\left(\sum_{i\in J}\Psi_i(d_i)\right)P_J = \sum_{J\subseteq[N]}\left(\sum_{i=1}^N \Psi_i(d_i)\mathbf{1}_J(i)\right)P_J \\ &= \sum_{i=1}^N \Psi_i(d_i)\sum_{J\subseteq[N]:i\in J} P_J = \sum_{i=1}^N p_i\Psi_i(d_i) \overset{(3.8)}{=} \frac{\tau}{N}\sum_{i=1}^N \Psi_i(d_i) = \frac{\tau}{N}\Psi(d).\end{aligned} \tag{5.6}$$

Furthermore, considering that $\bar{x}^k \in X^*$, then from (5.3) we obtain

$$\left\|x^k - \bar{x}^k\right\|_W \le c_\kappa(\tau)\sqrt{\mathbb{E}\left[\|x^{k+1} - x^k\|^2\right]}, \tag{5.7}$$

where $c_\kappa(\tau) = \left(\kappa_1 + \kappa_2(R_W(x^0))^2\right)\sqrt{\frac{N}{\tau}}$. We now need to express $\mathbb{E}[\Psi(x^{k+1})]$ explicitly, where x^{k+1} is generated by algorithm **(P-RCD)**. Note that $x^{k+1}_{\bar{J}^k} = x^k_{\bar{J}^k}$. As a result, we have

$$\begin{aligned}\mathbb{E}[\Psi(x^{k+1})] &= \mathbb{E}\left[\sum_{i\in J^k}\Psi_i\left([T_{J^k}(x^k)]_i\right) + \sum_{i\in \bar{J}^k}\Psi_i\left([x^k]_i\right)\right] \\ &\overset{(5.6)}{=} \frac{\tau}{N}\Psi\left(T_{[N]}(x^k)\right) + \frac{N-\tau}{N}\Psi(x^k).\end{aligned} \tag{5.8}$$

The following lemma establishes an important upper bound for $\mathbb{E}\left[F(x^{k+1}) - F(x^k)\right]$.

LEMMA 5.3. *If problem* (2.1) *satisfies Assumption* 1 *and* (GEBP) *property* (5.2)*, then the iterate* x^k *generated by the* **(P-RCD)** *method has the following property:*

$$\mathbb{E}\left[F(x^{k+1}) - F(x^k)\right] \le \mathbb{E}[\Lambda^k] \qquad \forall k \ge 0, \tag{5.9}$$

where $\Lambda^k = \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{1}{2}\|x^{k+1} - x^k\|_W^2 + \Psi(x^{k+1}) - \Psi(x^k)$.
Furthermore, we have that

$$\frac{1}{2}\|x^{k+1} - x^k\|_W^2 \le -\Lambda^k \qquad \forall k \ge 0. \tag{5.10}$$

Proof. Taking $x = x^k$ and $y = x^{k+1} - x^k$ in (2.5) we get

$$f(x^{k+1}) \le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{1}{2}\|x^{k+1} - x^k\|_W^2.$$

By adding $\Psi(x^{k+1})$ and subtracting $\Psi(x^k)$ in both sides of this inequality and by taking expectation in both sides, we obtain (5.9). Recall the iterate update (3.9) of our algorithm **(P-RCD)**:

$$x^{k+1} = \arg\min_{y\in\mathbb{R}^n} \langle \nabla_{J^k} f(x^k), y - x^k\rangle + \frac{1}{2}\|y - x^k\|_W^2 + \Psi_{J^k}(y).$$

Given that x^{k+1} is optimal for the problem above and if we take a vector $y = \alpha x^{k+1} + (1-\alpha)x^k$, with $\alpha \in [0,1]$, we have that

$$\begin{aligned}&\langle \nabla_{J^k} f(x^k), x^{k+1} - x^k\rangle + \frac{1}{2}\|x^{k+1} - x^k\|_W^2 + \Psi_{J^k}(x^{k+1})\\ &\le \alpha\langle \nabla_{J^k} f(x^k), x^{k+1} - x^k\rangle + \frac{\alpha^2}{2}\|x^{k+1} - x^k\|_W^2 + \Psi_{J^k}(\alpha x^{k+1} + (1-\alpha)x^k).\end{aligned}$$

Further, by rearranging the terms and through the convexity of Ψ_{J^k}, we obtain

$$(1-\alpha)\left[\langle \nabla_{J^k} f(x^k), x^{k+1} - x^k\rangle + \frac{1+\alpha}{2}\|x^{k+1} - x^k\|_W^2 + \Psi_{J^k}(x^{k+1}) - \Psi_{J^k}(x^k)\right] \le 0.$$

If we divide this inequality by $(1-\alpha)$ and let $\alpha \uparrow 1$, we have that

$$\langle \nabla_{J^k} f(x^k), x^{k+1} - x^k\rangle + (\Psi_{J^k}(x^{k+1}) - \Psi_{J^k}(x^k)) \le -\|x^{k+1} - x^k\|_W^2.$$

By adding $\frac{1}{2}\|x^{k+1} - x^k\|_W^2$ in both sides of this inequality and observing that

$$\begin{aligned}\langle \nabla_{J^k} f(x^k), x^{k+1} - x^k\rangle &= \langle \nabla f(x^k), x^{k+1} - x^k\rangle,\\ \Psi_{J^k}(x^{k+1}) - \Psi_{J^k}(x^k) &= \Psi(x^{k+1}) - \Psi(x^k),\end{aligned}$$

we obtain (5.10). ∎

Additionally, note that by applying expectation in J^k to Λ^k we get

$$\begin{aligned}\mathbb{E}[\Lambda^k] &\overset{(4.2)}{=} \frac{\tau}{N}\langle \nabla f(x^k), T_{[N]}(x^k) - x^k\rangle + \frac{1}{2}\mathbb{E}\left[\|x^{k+1} - x^k\|_W^2\right]\\ &\quad + \mathbb{E}\left[\Psi(x^{k+1})\right] - \Psi(x^k)\\ &\overset{(5.8)}{=} \frac{\tau}{N}\langle \nabla f(x^k), T_{[N]}(x^k) - x^k\rangle + \frac{1}{2}\mathbb{E}\left[\|x^{k+1} - x^k\|_W^2\right]\\ &\quad + \frac{\tau}{N}\left(\Psi(T_{[N]}(x)) - \Psi(x^k)\right).\end{aligned} \tag{5.11}$$

The following theorem, which is the main result of this section, proves the linear convergence rate for the algorithm **(P-RCD)** on optimization problems having (GEBP) property (5.2).

THEOREM 5.4. *On optimization problems* (2.1) *with the objective function satisfying Assumption* 1 *and* (GEBP) *property* (5.2)*, the algorithm* **(P-RCD)** *has the following global linear convergence rate for the expected values of the objective function:*

$$\mathbb{E}\left[F(x^k) - F^*\right] \le \theta^k(F(x^0) - F^*) \qquad \forall k \ge 0, \tag{5.12}$$

where $\theta < 1$ *is a constant depending on* $N, \tau, \kappa_1, \kappa_2$*, and* $R_W(x^0)$*.*

Proof. We first need to establish an upper bound for $\mathbb{E}[F(x^{k+1})] - F(\bar{x}^k)$. By the definition of F and its convexity, we have that

$$\begin{aligned}
&F(x^{k+1}) - F(\bar{x}^k) = f(x^{k+1}) - f(\bar{x}^k) + \Psi(x^{k+1}) - \Psi(\bar{x}^k)\\
&\le \langle \nabla f(x^{k+1}), x^{k+1} - \bar{x}^k\rangle + \Psi(x^{k+1}) - \Psi(\bar{x}^k)\\
&= \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - \bar{x}^k\rangle + \langle \nabla f(x^k), x^{k+1} - \bar{x}^k\rangle + \Psi(x^{k+1}) - \Psi(\bar{x}^k)\\
&\le \|\nabla f(x^{k+1}) - \nabla f(x^k)\|_{W^{-1}} \|x^{k+1} - \bar{x}^k\|_W + \langle \nabla f(x^k), x^{k+1} - \bar{x}^k\rangle + \Psi(x^{k+1}) - \Psi(\bar{x}^k)\\
&\overset{(2.7)}{\le} \|x^{k+1} - x^k\|_W \|x^{k+1} - \bar{x}^k\|_W + \langle \nabla f(x^k), x^{k+1} - \bar{x}^k\rangle + \Psi(x^{k+1}) - \Psi(\bar{x}^k)\\
&\le \|x^{k+1} - x^k\|_W^2 + \|x^{k+1} - x^k\|_W \|x^k - \bar{x}^k\|_W + \langle \nabla f(x^k), x^{k+1} - \bar{x}^k\rangle + \Psi(x^{k+1}) - \Psi(\bar{x}^k).
\end{aligned}$$

By taking expectation in both sides of the previous inequality, we have

$$\begin{aligned}
\mathbb{E}[F(x^{k+1})] - F(\bar{x}^k) &\le \mathbb{E}[\|x^{k+1} - x^k\|_W^2] + \mathbb{E}\left[\|x^{k+1} - x^k\|_W \|x^k - \bar{x}^k\|_W\right]\\
&\quad + \mathbb{E}\left[\langle \nabla f(x^k), x^{k+1} - \bar{x}^k\rangle + \Psi(x^{k+1})\right] - \Psi(\bar{x}^k).
\end{aligned} \tag{5.13}$$

From (3.12) we have that $\|x^k - \bar{x}^k\| \le R_W(x^0)$ and derive the following:

$$\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^k\|_W \|x^k - \bar{x}^k\|_W\right] &= \|x^k - \bar{x}^k\|_W \mathbb{E}\left[\|x^{k+1} - x^k\|_W\right]\\
&\overset{(5.7)}{\le} c_\kappa(\tau) \sqrt{\mathbb{E}\left[\|x^{k+1} - x^k\|_W^2\right]} \sqrt{\left(\mathbb{E}\left[\|x^{k+1} - x^k\|_W\right]\right)^2}\\
&\le c_\kappa(\tau)\, \mathbb{E}\left[\|x^{k+1} - x^k\|_W^2\right],
\end{aligned}$$

where the last step comes from Jensen's inequality. Thus, (5.13) becomes

$$\begin{aligned}
\mathbb{E}[F(x^{k+1})] - F(\bar{x}^k) &\le c_1(\tau)\mathbb{E}[\|x^{k+1} - x^k\|_W^2] + \mathbb{E}\left[\left\langle \nabla f(x^k), x^{k+1} - \bar{x}^k\right\rangle\right]\\
&\quad + \mathbb{E}[\Psi(x^{k+1})] - \Psi(\bar{x}^k),
\end{aligned} \tag{5.14}$$

where $c_1(\tau) = (1 + c_\kappa(\tau))$. We now explicitly express the second term in the right-hand side of the above inequality:

$$\begin{aligned}
\mathbb{E}\left[\langle \nabla f(x^k), x^{k+1} - \bar{x}^k\rangle\right] &\overset{(3.9)}{=} \mathbb{E}\left[\langle \nabla f(x^k), x^k + T_{J^k}(x^k) - x^k_{J^k} - \bar{x}^k\rangle\right]\\
&= \langle \nabla f(x^k), x^k - \bar{x}^k\rangle + \mathbb{E}\left[\langle \nabla f(x^k), T_{J^k}(x^k) - x^k_{J^k}\rangle\right]\\
&\overset{(4.2)}{=} \langle \nabla f(x^k), x^k - \bar{x}^k\rangle + \frac{\tau}{N}\langle \nabla f(x^k), T_{[N]}(x^k) - x^k\rangle\\
&= \left(1 - \frac{\tau}{N}\right)\langle \nabla f(x^k), x^k - \bar{x}^k\rangle + \frac{\tau}{N}\langle \nabla f(x^k), T_{[N]}(x^k) - \bar{x}^k\rangle.
\end{aligned}$$

So, by replacing it in (5.14) and through (5.8), we get

$$\begin{aligned}
\mathbb{E}[F(x^{k+1})] - F(\bar{x}^k) &\le c_1(\tau)\mathbb{E}[\|x^{k+1} - x^k\|_W^2] + \frac{\tau}{N}\langle \nabla f(x^k), T_{[N]}(x^k) - \bar{x}^k\rangle\\
&\quad + \left(1 - \frac{\tau}{N}\right)\langle \nabla f(x^k), x^k - \bar{x}^k\rangle + \frac{\tau}{N}\Psi(T_{[N]}(x^k)) + \left(1 - \frac{\tau}{N}\right)\Psi(x^k) - \Psi(\bar{x}^k).
\end{aligned} \tag{5.15}$$

By taking $y = \bar{x}^k$ and $x = x^k$ in (2.5), we obtain

$$f(\bar{x}^k) \le f(x^k) + \langle \nabla f(x^k), \bar{x}^k - x^k\rangle + \frac{1}{2}\|\bar{x}^k - x^k\|_W^2.$$

Through this and by rearranging terms in (5.15), we obtain

$$\mathbb{E}[F(x^{k+1})]-F(\bar{x}^k) \leq c_1(\tau)\mathbb{E}[\|x^{k+1}-x^k\|_W^2] + \frac{1}{2}\left(1-\frac{\tau}{N}\right)\|x^k-\bar{x}^k\|_W^2$$
$$+\left(1-\frac{\tau}{N}\right)\left(F(x^k)-F(\bar{x}^k)\right)+\frac{\tau}{N}\left(\Psi(T_{[N]}(x^k))+\langle\nabla f(x^k),T_{[N]}(x^k)-\bar{x}^k\rangle-\Psi(\bar{x}^k)\right).$$

Furthermore, from (5.7) we obtain

(5.16)
$$\mathbb{E}[F(x^{k+1})]-F(\bar{x}^k) \leq \left(c_1(\tau)+\frac{1}{2}\left(1-\frac{\tau}{N}\right)c_\kappa(\tau)^2\right)\mathbb{E}[\|x^{k+1}-x^k\|_W^2]$$
$$+\left(1-\frac{\tau}{N}\right)\left(F(x^k)-F(\bar{x}^k)\right)+\frac{\tau}{N}\left(\Psi(T_{[N]}(x^k))+\langle\nabla f(x^k),T_{[N]}(x^k)-\bar{x}^k\rangle-\Psi(\bar{x}^k)\right).$$

Through the convexity of $\Psi(x)$, we have

$$\Psi(T_{[N]}(x^k))-\Psi(\bar{x}^k) \leq \left\langle \partial\Psi(T_{[N]}(x^k)), T_{[N]}(x^k)-\bar{x}^k\right\rangle.$$

Combining the previous relation with the optimality condition (3.5) and replacing the resulting expression in (5.16), we obtain

$$\mathbb{E}[F(x^{k+1})]-F(\bar{x}^k) \leq \left(c_1(\tau)+\frac{1}{2}\left(1-\frac{\tau}{N}\right)c_\kappa(\tau)^2\right)\mathbb{E}[\|x^{k+1}-x^k\|_W^2]$$
$$+\left(1-\frac{\tau}{N}\right)\left(F(x^k)-F(\bar{x}^k)\right)+\frac{\tau}{N}\langle -W\left(T_{[N]}(x^k)-x^k\right), T_{[N]}(x^k)-\bar{x}^k\rangle. \tag{5.17}$$

By rearranging some terms and through the Cauchy–Schwarz inequality, we get

$$\left\langle -W\left(T_{[N]}(x^k)-x^k\right), T_{[N]}(x^k)-\bar{x}^k\right\rangle = \left\langle -W\left(T_{[N]}(x^k)-x^k\right), T_{[N]}(x^k)-x^k+x^k-\bar{x}^k\right\rangle$$
$$\leq \left\langle W\left(T_{[N]}(x^k)-x^k\right), \bar{x}^k-x^k\right\rangle \leq \|W\left(T_{[N]}(x^k)-x^k\right)\|_{W^{-1}}\|\bar{x}^k-x^k\|_W$$
$$= \|T_{[N]}(x^k)-x^k\|_W\|\bar{x}^k-x^k\|_W.$$

Now, recall that

$$\mathbb{E}\left[\|x^{k+1}-x^k\|_W^2\right] = \frac{\tau}{N}\|x^k-T_{[N]}(x^k)\|_W^2.$$

Thus, from this and (5.7) we get

$$\frac{\tau}{N}\|T_{[N]}(x^k)-x^k\|_W\|\bar{x}^k-\bar{x}^k\|_W \leq c_\kappa(\tau)\sqrt{\frac{\tau}{N}}\mathbb{E}\left[\|x^{k+1}-x^k\|_W^2\right].$$

By replacing this in (5.17), we obtain

$$\mathbb{E}[F(x^{k+1})]-F(\bar{x}^k) \leq \underbrace{\left(c_1(\tau)+\frac{1}{2}\left(1-\frac{\tau}{N}\right)c_\kappa(\tau)^2+c_\kappa(\tau)\sqrt{\frac{\tau}{N}}\right)}_{c_2(\tau)}\mathbb{E}[\|x^{k+1}-x^k\|_W^2]$$
$$+\left(1-\frac{\tau}{N}\right)\left(F(x^k)-F(\bar{x}^k)\right)$$
$$= c_2(\tau)\mathbb{E}[\|x^{k+1}-x^k\|_W^2]+\left(1-\frac{\tau}{N}\right)\left(F(x^k)-F(\bar{x}^k)\right). \tag{5.18}$$

From (5.10) we have $\mathbb{E}[\|x^{k+1}-x^k\|_W^2] \leq -2\mathbb{E}[\Lambda^k]$. Now, through this and by rearranging some terms in (5.18), we obtain

$$\frac{\tau}{N}\left(\mathbb{E}[F(x^{k+1})]-F(\bar{x}^k)\right) \leq -2c_2(\tau)\mathbb{E}[\Lambda^k] + \left(1-\frac{\tau}{N}\right)\left(F(x^k)-\mathbb{E}[F(x^{k+1})]\right).$$

Furthermore, from (5.9) we obtain

$$\begin{aligned}\mathbb{E}[F(x^{k+1})]-F(\bar{x}^k) &\leq \underbrace{\frac{N}{\tau}\left(2c_2(\tau)+\left(1-\frac{\tau}{N}\right)\right)}_{c_3(\tau)}\left(F(x^k)-\mathbb{E}[F(x^{k+1})]\right)\\ &= c_3(\tau)\left(F(x^k)-\mathbb{E}[F(x^{k+1})]\right).\end{aligned}$$

By rearranging this inequality, we obtain

$$\mathbb{E}[F(x^{k+1})]-F(\bar{x}^k) \leq \frac{c_3(\tau)}{1+c_3(\tau)}\left(F(x^k)-F(\bar{x}^k)\right). \tag{5.19}$$

We denote $\theta = \frac{c_3(\tau)}{1+c_3(\tau)} < 1$ and define $\delta^k = F(x^{k+1}) - F(\bar{x}^k)$. By taking expectation over η^{k-1} in (5.19), we arrive at

$$\mathbb{E}[\delta^k] \leq \theta\mathbb{E}[\delta^{k-1}] \leq \cdots \leq \theta^k\mathbb{E}[\delta^0],$$

and linear convergence is proved. ∎

Finally, we establish the number of iterations k_ρ^ϵ which will ensure an ϵ-suboptimal solution with probability at least $1-\rho$. We first recall the following well-known inequality: for constants $\epsilon > 0$ and $\gamma \in (0,1)$ such that $\delta^0 > \epsilon > 0$ and $k \geq \frac{1}{\gamma}\log\left(\frac{\delta^0}{\epsilon}\right)$ we have

$$(1-\gamma)^k\delta^0 = \left(1-\frac{1}{1/\gamma}\right)^{(1/\gamma)(\gamma k)}\delta^0 \leq \exp(-\gamma k)\delta^0 \leq \exp\left(-\log(\delta^0/\epsilon)\right)\delta^0 = \epsilon. \tag{5.20}$$

Now, for problem (2.1) satisfying Assumption 1 and (GEBP) property (5.2), consider a probability level $\rho \in (0,1)$, suboptimality $0 < \epsilon < \delta_0$, and an iteration counter:

$$k_\rho^\epsilon \geq \frac{1}{1-\theta}\log\left(\frac{\delta^0}{\epsilon\rho}\right),$$

where $\delta^0 = F(x^0) - F^*$ and θ is defined in Theorem 5.4. Then, from Markov's inequality and (5.20) we have that the iterate $x^{k_\rho^\epsilon}$ generated by **(P-RCD)** satisfies

$$\mathbb{P}(F(x^{k_\rho^\epsilon}) - F^* \leq \epsilon) \geq 1-\rho.$$

Note that we have obtained global linear convergence for **(P-RCD)** on the general class of problems satisfying (GEBP) property (5.2), as opposed to the results in [28, 8], where the authors show only local linear convergence for deterministic coordinate descent methods applied to local error bound problems, i.e., for all $k \geq k_0 > 1$, where k_0 is an iterate after which some error bound condition of the form $\|x^k - \bar{x}^k\| \leq \bar{\kappa}\|\nabla^+F(x^k)\|$ is implicitly satisfied. In [29], global linear convergence is also proved for the full gradient method but applied only to problems with Ψ as indicator function of a polyhedron and having an error bound property. Further, our convergence results are also more general than those in [18, 10, 15, 23, 7], in the sense that we can

show linear convergence of algorithm **(P-RCD)** for larger classes of problems than in these papers, where linear convergence is proved for the more restricted class of problems having smooth and strongly convex objective functions. For example, to our knowledge, the best global convergence rate results known for gradient type methods for solving constrained lasso (2.8) or dual formulation of a linearly constrained convex problem (2.10) were of the sublinear form $\mathcal{O}\left(\frac{1}{k^2}\right)$ [19, 13]. In this paper we prove *global* linear convergence rate for random coordinate gradient descent methods for solving problems of type (2.8) or (2.10). Note that for the particular case of least-square problems $\min_{x\in\mathbb{R}^n} \|Ax-b\|^2$ the authors in [6], using also an error bound like property, were able to show linear convergence for a random coordinate gradient descent method. Our results can be viewed as a generalization of the results from [6] to more general optimization problems (2.1). Moreover, our proof for linear convergence is different from those in [18, 10, 23, 7]. Finally, our approach allows us to analyze in the same framework several methods: full gradient, serial coordinate descent, and any parallel coordinate descent method in between.

6. Conditions for generalized error bound functions. In this section we investigate under which conditions an objective function F of (2.1) satisfying Assumption 1 has the (GEBP) property given in Definition 5.1.

6.1. Case 1: f strongly convex and Ψ convex. In the first case we consider f satisfying Assumption 1 and also strong convexity, while Ψ is a general convex function. Then, F has the (GEBP) property defined in (5.2). Indeed, let us consider f to be σ_W-strongly convex w.r.t. the norm $\|\cdot\|_W$, i.e.,

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \frac{\sigma_W}{2}\|y-x\|_W^2. \tag{6.1}$$

Then, we can prove the following error bound property (see, e.g., [11]):

$$\|x-\bar{x}\|_W \leq \frac{2}{\sigma_W}\|\nabla^+ F(x)\|_W \qquad \forall x\in\mathbb{R}^n;$$

i.e., we have $\kappa_1 = \frac{2}{\sigma_W}$ and $\kappa_2 = 0$ in our Definition 5.1 of the (GEBP) property. It follows that objective functions of (2.1), written as the sum between a nonsmooth and a strongly convex function, are included in our class of generalized error bound problems (5.2). Combining (2.5) with (6.1) we get $\sigma_W \leq 1$. In this case we have the following linear convergence rate for algorithm **(P-RCD)** (see [11]):

$$\mathbb{E}\left[F(x^{k+1}) - F^*\right] \leq (1-\gamma_{sc}^{eb})^k \left(F(x^0) - F^*\right), \tag{6.2}$$

where $\gamma_{sc}^{eb} = \frac{\tau\sigma_W}{N}$. We notice that, given the choice of τ, we obtain different linear convergence results of order $\mathcal{O}(\theta^k)$. For example, for $\tau = 1$ we obtain a linear convergence rate similar to that of the random coordinate descent method in [9, 18, 7, 24], i.e., $\gamma_{sc}^{eb} = \mathcal{O}(\sigma_W/N)$, while for $\tau = N$ we get a convergence rate similar to that of the full composite gradient method of [19], i.e., $\gamma_{sc}^{eb} = \mathcal{O}(\sigma_W)$. Finally, if we consider f to be $\sigma_{W'}$-strongly convex in the norm $\|\cdot\|_{W'}$, where the matrix $W' = \mathrm{diag}\left(L_i I_{n_i}; i\in[N]\right)$, then algorithm (PCDM1) in [23] has a convergence rate as above with $\gamma_{sc} = \frac{\tau\sigma_{W'}}{N+\omega\tau}$. However, the distances are measured in different norms in all these papers.

6.2. Case 2: Ψ indicator function of a polyhedral set. Another important category of optimization problems (2.1) that we consider has the following form:

$$\min_{x\in\mathbb{R}^n} F(x) \quad \left(= \tilde{f}(Px) + c^T x + \mathbf{I}_X(x)\right), \tag{6.3}$$

where $f(x) = \tilde{f}(Px) + c^T x$ is a smooth convex function, $P \in \mathbb{R}^{p\times n} \setminus \{0\}$ is a constant matrix upon which we make no assumptions, and $\Psi(x) = \mathbf{I}_X(x)$ is the indicator function of the polyhedral set X. Note that an objective function F with the structure (6.3) appears in many applications; see, e.g., the dual problem (2.10) obtained from the primal formulation (2.9) given in section 2.1. Now, for proving (GEBP) property (5.2), we require that f satisfy the following assumption.

Assumption 2. We consider that $f(x) = \tilde{f}(Px) + c^T x$ satisfies Assumption 1. We also assume that $\tilde{f}(z)$ is σ-strongly convex in z, the set of optimal solutions X^* for problem (2.1) is bounded, and $P \neq 0$.

For problem (6.3), functions f under which the set X^* is bounded include, e.g., continuously differentiable coercive functions [26]. Also, if (6.3) is the dual formulation of the primal problem (2.9) for which the Slater condition holds, then by Gauvin's theorem we have that the set of optimal Lagrange multipliers, i.e., X^* in this case, is compact [26]. Note that for the nonsmooth component $\Psi(x) = \mathbf{I}_X(x)$ we only assume that X is a polyhedron (possibly unbounded).

Our generalized error bound property for problem (6.3) is in a way similar to that in [8, 28, 29]. However, our results are more general, in the sense that they hold globally, while in [8, 28] the authors prove their results only locally and in the sense that we allow the constraints set X to be an unbounded polyhedron, and in [29] an error bound like property is proved only for bounded polyhedra or $\mathbb{R}^n$. This extension is very important since it allows us, e.g., to tackle the dual formulation of a primal problem (2.9), in which $X = \mathbb{R}^n_+$ (nonnegative orthant), appearing in many practical applications. Last but not least important is that our error bound definition is more general than that used in [8, 28, 29], as we can see from the following example.

Example 1. Let us consider the following quadratic problem: $\min_{x\in\mathbb{R}^2_+} 1/2(x_1 - x_2)^2 + x_1 + x_2$. We can easily see that $X^* = \{0\}$, and thus this example satisfies Assumption 2. Clearly, for this example (GEBP) property (5.2) holds with, e.g., $\kappa_1 = \kappa_2 = 1$. However, there is no finite constant κ satisfying the classical error bound property [8, 28, 29]: $\|x - \bar{x}\|_W \leq \kappa \|\nabla^+ F(x)\|_W$ for all $x \in \mathbb{R}^2_+$ (we can see this by taking $x_1 = x_2 \geq 1$ in the previous inequality).

Since $\Psi(x)$ is a set indicator function, the gradient mapping of F can be expressed as

$$\nabla^+ F(x) = x - \Pi_X^W \left(x - W^{-1}\nabla f(x)\right).$$

The next lemma establishes the Lipschitz continuity of the proximal gradient mapping.

LEMMA 6.1. *For composite function F of* (6.3) *satisfying Assumption* 1*, we have*

$$\|\nabla^+ F(x) - \nabla^+ F(y)\|_W \leq 3\|x - y\|_W \quad \forall x, y \in X. \tag{6.4}$$

Proof. By definition of $\nabla^+ F(x)$, we have that

$$\begin{aligned}
\|\nabla^+ F(x) - \nabla^+ F(y)\|_W &= \|x - y + T_{[N]}(y) - T_{[N]}(x)\|_W \\
&\overset{(3.4)}{\leq} \|x-y\|_W + \|\mathrm{prox}_\Psi(x - W^{-1}\nabla f(x)) - \mathrm{prox}_\Psi(y - W^{-1}\nabla f(y))\|_W \\
&\overset{(3.3)}{\leq} \|x-y\|_W + \|x - y + W^{-1}(\nabla f(y) - \nabla f(x))\|_W \\
&\leq 2\|x - y\|_W + \|\nabla f(x) - \nabla f(y)\|_{W^{-1}} \overset{(2.7)}{\leq} 3\|x - y\|_W,
\end{aligned}$$

and the proof is complete. ∎

The following lemma introduces an important property for projection operator Π_X^W.

LEMMA 6.2. *Given a convex set X, its projection operator Π_X^W satisfies*

$$\left\langle W\left(\Pi_X^W(x)-x\right),\Pi_X^W(x)-y\right\rangle \leq 0 \quad \forall y\in X. \tag{6.5}$$

Proof. Following the definition of Π_X^W, we have that

$$\|x-\Pi_X^W(x)\|_W^2 \leq \|x-d\|_W^2 \quad \forall d\in X. \tag{6.6}$$

Since X is a convex set, consider a point $d=\alpha y+(1-\alpha)\Pi_X^W(x)\in X$, with $y\in X$ and $\alpha\in[0,1]$, and by (6.6) we obtain

$$\|x-\Pi_X^W(x)\|_W^2 \leq \|x-(\alpha y+(1-\alpha)\Pi_X^W(x))\|_W^2.$$

If we elaborate the squared norms in the inequality above. we arrive at

$$0\leq \alpha\left\langle W\left(\Pi_X^W(x)-x\right),y-\Pi_X^W(x)\right\rangle+\frac{1}{2}\alpha^2\|y-\Pi_X^W(x)\|^2.$$

If we divide both sides by α and let $\alpha\downarrow 0$, we get (6.5). ∎

The next lemma establishes an important property between $\nabla f(x)$ and $\nabla^+F(x)$.

LEMMA 6.3. *Given a function f that satisfies* (2.7) *and a convex set X, then the following inequality holds:*

$$\langle\nabla f(x)-\nabla f(y),x-y\rangle \leq 2\|\nabla^+F(x)-\nabla^+F(y)\|_W\|x-y\|_W \quad \forall x,y\in X.$$

Proof. Denote $z=x-W^{-1}\nabla f(x)$; then by replacing $x=z$ and $y=\Pi_X^W(y-W^{-1}\nabla f(y))$ in Lemma 6.2 we obtain the following inequality:

$$\left\langle W\left(\Pi_X^W(z)-x\right)+\nabla f(x),\Pi_X^W(z)-\Pi_X^W\big(y-W^{-1}\nabla f(y)\big)\right\rangle \leq 0.$$

From the definition of the projected gradient map, this inequality can be written as

$$\left\langle\nabla f(x)-W\nabla^+F(x),x-\nabla^+F(x)-y+\nabla^+F(y)\right\rangle\leq 0.$$

If we further elaborate the inner product, we obtain

$$\begin{aligned}\langle\nabla f(x),x-y\rangle &\leq \langle W\nabla^+F(x),x-y\rangle\\ &\quad+\langle\nabla f(x),\nabla^+F(x)-\nabla^+F(y)\rangle-\langle W\nabla^+F(x),\nabla^+F(x)-\nabla^+F(y)\rangle.\end{aligned} \tag{6.7}$$

By adding two copies of (6.7) with x and y interchanged, we have the inequality

$$\begin{aligned}&\langle\nabla f(x)-\nabla f(y),x-y\rangle\\ &\leq\left\langle W(\nabla^+F(x)-\nabla^+F(y)),x-y\right\rangle+\left\langle\nabla f(x)-\nabla f(y),\nabla^+F(x)-\nabla^+F(y)\right\rangle.\end{aligned}$$

From this inequality, through Cauchy–Schwarz and (2.7) we arrive at

$$\begin{aligned}\langle\nabla f(x)-\nabla f(y),x-y\rangle &\leq \|\nabla^+F(x)-\nabla^+F(y)\|_W\left(\|x-y\|_W+\|\nabla f(x)-\nabla f(y)\|_W^{-1}\right)\\ &\leq 2\|\nabla^+F(x)-\nabla^+F(y)\|_W\|x-y\|_W,\end{aligned}$$

and the proof is complete. ∎

We now introduce the following lemma regarding the optimal set X^*.

LEMMA 6.4 (see [8]). *Under Assumption* 2, *there exists a unique* z^* *such that*

$$Px^* = z^* \quad \forall x^* \in X^* \quad \textit{and} \quad \nabla f(x) = P^T \nabla \tilde{f}(z^*) + c$$

for all $x \in Q = \{y \in X : \ Py = z^*\}$.

Consider now a point $x \in X$, and denote by $\mathbf{q} = \Pi_Q^W(x)$ the projection of the point x onto the set $Q = \{y \in X : \ Py = z^*\}$, as defined in Lemma 6.4, and by $\bar{\mathbf{q}}$ its projection onto the optimal set X^*, i.e., $\bar{\mathbf{q}} = \Pi_{X^*}^W(\mathbf{q})$. Given the set Q, the distance to the optimal set can be decomposed as

$$\|x - \bar{x}\|_W \leq \|x - \bar{\mathbf{q}}\|_W \leq \|x - \mathbf{q}\|_W + \|\mathbf{q} - \bar{\mathbf{q}}\|_W.$$

Given this inequality, the outline for proving (GEBP) property (5.2) in this case is to obtain appropriate upper bounds for $\|x - \mathbf{q}\|_W$ and $\|\mathbf{q} - \bar{\mathbf{q}}\|_W$. In what follows we introduce lemmas for establishing bounds for each of these two terms.

LEMMA 6.5. *Under Assumption* 2, *there exists a constant* γ_1 *such that*

$$\|x - \mathbf{q}\|_W^2 \leq \gamma_1^2 \frac{2}{\sigma} \|\nabla^+ F(x)\|_W \|x - \bar{x}\|_W \quad \forall x \in X.$$

Proof. Corollary 2.2 in [25] states that if we have two polyhedra,

$$Ay \leq b_1, \ Py = d_1 \quad \text{and} \quad Ay \leq b_2, \ Py = d_2, \tag{6.8}$$

then there exists a finite constant (so-called Hoffman constant) $\gamma_1 > 0$ such that for a point y_1 which satisfies the first set of constraints and a point y_2 which satisfies the second one we have

$$\|y_1 - y_2\|_W \leq \gamma_1 \left\| \begin{matrix} \Pi_{\mathbb{R}^+}(b_1 - b_2) \\ d_1 - d_2 \end{matrix} \right\|_W. \tag{6.9}$$

Furthermore, the constant γ_1 is only dependent on the matrices A and P. Given that X is a polyhedral set, we can express it as $X = \{x \in \mathbb{R}^n : Ax \leq b\}$. Thus, for any $x \in X$, we can take ($b_1 = b$, $d_1 = Px$) and ($b_2 = b$, $d_2 = z^*$) in (6.8) such that

$$Ay \leq b, \ Py = Px \quad \text{and} \quad Ay \leq b, \ Py = z^*. \tag{6.10}$$

Evidently, the point $x \in X$ is feasible for the first polyhedra in (6.10). Consider now a point y_2 feasible for the second polyhedra in (6.10). Therefore, from (6.9) there exists a Hoffman constant γ_1 such that

$$\|x - y_2\|_W \leq \gamma_1 \|Px - z^*\|_W \quad \forall x \in X.$$

Furthermore, from the definition of $\mathbf{q}$ we get

$$\|x - \mathbf{q}\|_W^2 \leq \|x - y_2\|_W^2 \leq \gamma_1^2 \|Px - z^*\|_W^2 \quad \forall x \in X. \tag{6.11}$$

From the strong convexity of $\tilde{f}(z)$ we have the following property:

$$\sigma \|Px - z^*\|_W^2 \leq \left\langle \nabla \tilde{f}(Px) - \nabla \tilde{f}(P\bar{x}), Px - P\bar{x} \right\rangle = \langle \nabla f(x) - \nabla f(\bar{x}), x - \bar{x} \rangle$$

for all $\bar{x} \in X^*$. From this inequality and Lemma 6.3 we obtain

$$\sigma \|Px - z^*\|_W^2 \leq 2 \|\nabla^+ F(x) - \nabla^+ F(\bar{x})\|_W \|x - \bar{x}\|_W.$$

Since $\bar{x} \in X^*$, then $\nabla^+ F(\bar{x}) = 0$. Thus, from the inequality above and (6.11) we get

$$\|x - \mathbf{q}\|_W^2 \leq \gamma_1^2 \frac{2}{\sigma} \|\nabla^+ F(x)\|_W \|x - \bar{x}\|_W,$$

and the proof is complete. ∎

Note that if in (6.3) we have $c = 0$, then by definition we have that $Q = X^*$, and thus we get $\|\mathbf{q} - \bar{\mathbf{q}}\|_W = 0$. In such a case, also note that $\mathbf{q} = \bar{x}$ and through the previous lemma, in which we established an upper bound for $\|x - \mathbf{q}\|_W$, we can prove outright (GEBP) property (5.2) with $\kappa_1 = \gamma_1^2 \frac{2}{\sigma}$ and $\kappa_2 = 0$. If $c \neq 0$, the following two lemmas are introduced to investigate the distance between a point and a solution set of a linear program and then to establish a bound for $\|\mathbf{q} - \bar{\mathbf{q}}\|_W$.

LEMMA 6.6. *Consider a linear program on a nonempty polyhedral set Y,*

$$\min_{y \in Y} b^T y, \tag{6.12}$$

and assume that the optimal set $Y^* \subseteq Y$ is nonempty, convex, and bounded. Let $\bar{y}$ be the projection of a point $y \in Y$ on the optimal set Y^*. For this problem we have that

$$\|y - \bar{y}\|_W \leq \gamma_2 \left(\|y - \bar{y}\|_W + \|b\|_{W^{-1}} \right) \|y - \Pi_Z^W (y - W^{-1} b)\|_W \quad \forall y \in Y, \tag{6.13}$$

where Z is any convex set satisfying $Y \subseteq Z$ and γ_2 is a constant depending on (Y, b).

Proof. Because the solution set Y^* is nonempty, convex, and bounded, the linear program (6.12) is equivalent to the problem $\min_{y \in Y^*} b^T y$, and as a result, (6.12) is solvable. By the duality theorem of linear programming, the dual problem of (6.12) is well defined and solvable, and strong duality holds for the dual:

$$\max_{\mu \in Y'} l(\mu), \tag{6.14}$$

where $Y' \subseteq \mathbb{R}^m$ is the dual feasible set. For any pair of primal-dual feasible points (y, μ) for problems (6.12) and (6.14), we have a corresponding pair of optimal solutions (y^*, μ^*). By the solvability of (6.12), we have from Theorem 2 of [25] that there exists a constant γ_2 depending on Y and b such that we have the bound

$$\left\| \begin{matrix} y - y^* \\ \mu - \mu^* \end{matrix} \right\|_{\mathrm{diag}(W, I_m)} \leq \gamma_2 |b^T y - l(\mu)|.$$

By strong duality, we have that $l(\mu^*) = b^T \bar{y}$. Thus, taking $\mu = \mu^*$ and through the optimality conditions of (6.12), we obtain $\|y - y^*\|_W \leq \gamma_2 \langle b, y - \bar{y} \rangle$. From this inequality and $\|y - \bar{y}\|_W \leq \|y - y^*\|_W$ we arrive at

$$\|y - \bar{y}\|_W \leq \gamma_2 \langle b, y - \bar{y} \rangle. \tag{6.15}$$

By Lemma 6.2, we have that

$$\left\langle W \left(\Pi_Z^W \left(y - W^{-1} b \right) - \left(y - W^{-1} b \right) \right), \Pi_Z^W (y - W^{-1} b) - \bar{y} \right\rangle \leq 0.$$

This inequality can be rewritten as

$$\begin{aligned} \langle b, y - \bar{y} \rangle &\leq \left\langle W \left(y - \Pi_Z^W (y - W^{-1} b) \right), y - \bar{y} + W^{-1} b + \Pi_Z^W (y - W^{-1} b) - y \right\rangle \\ &\leq \|y - \Pi_Z^W (y - W^{-1} b)\|_W \left(\|y - \bar{y}\|_W + \|b\|_{W^{-1}} \right). \end{aligned}$$

From this inequality and (6.15) we obtain

$$\|y-\bar{y}\|_W \le \gamma_2 \left(\|y-\bar{y}\|_W + \|b\|_{W^{-1}}\right) \|y - \Pi_Z^W (y - W^{-1}b)\|_W,$$

and the proof is complete. ∎

LEMMA 6.7. *If Assumption* 2 *holds for optimization problem* (6.3)*, then there exists a constant* $\gamma_2 > 0$ *such that*

$$\|\mathbf{q}-\bar{\mathbf{q}}\|_W \le \gamma_2 \left(\|\mathbf{q}-\bar{\mathbf{q}}\|_W + \|\nabla f(\bar{x})\|_{W^{-1}}\right) \|\nabla F^+(\mathbf{q})\|_W \quad \forall x \in X. \tag{6.16}$$

Proof. By Lemma 6.4, we have that $Px = z^*$ for all $x \in Q$. As a result, the optimization problem $\min_{x\in Q} \tilde{f}(z^*) + c^T x$ has the same solution set as problem (6.3), due to the fact that $X^* \subseteq Q \subseteq X$. Since z^* is a constant, we can formulate the equivalent problem:

$$\min_{x\in Q} \nabla f(\bar{x})^T x \qquad \left(= \nabla \tilde{f}(z^*)^T z^* + c^T x\right).$$

Note that $\nabla f(\bar{x}) = P^T \nabla \tilde{f}(z^*) + c$ is constant and under Assumption 2 we have that X^* is convex and bounded. Furthermore, since $\bar{x}, \mathbf{q} \in Q$, then $\nabla f(\bar{x}) = \nabla f(\mathbf{q})$. Considering these details, and by taking $Y = Q$, $Z = X$, $y = \mathbf{q}$, and $b = \nabla f(\bar{x})$ in Lemma 6.6 and applying it to the previous problem, we obtain (6.16). ∎

The next theorem establishes the generalized error bound property for optimization problems in the form (6.3) having objective functions satisfying Assumption 2.

THEOREM 6.8. *Under Assumption* 2*, the optimization problem* (6.3) *with* $F(x) = \tilde{f}(Px) + c^T x + \mathbf{I}_X(x)$ *satisfies the following global generalized error bound property:*

$$\|x-\bar{x}\|_W \le \left(\kappa_1 + \kappa_2 \|x-\bar{x}\|_W^2\right) \|\nabla^+ F(x)\|_W \quad \forall x \in X, \tag{6.17}$$

where κ_1 *and* κ_2 *are two nonnegative constants.*

Proof. Since $\bar{x} \in X^*$, then $\nabla^+ F(\bar{x}) = 0$ and by Lemma 6.1 we have

$$\|\nabla^+ F(x)\|_W = \|\nabla^+ F(x) - \nabla^+ F(\bar{x})\|_W \le 3\|x-\bar{x}\|_W.$$

From this inequality and by applying Lemma 6.1, we also have

$$\begin{aligned}
\|\nabla^+ F(\mathbf{q})\|_W^2 &\le \left(\|\nabla^+ F(x)\|_W + \|\nabla^+ F(\mathbf{q}) - \nabla^+ F(x)\|_W\right)^2 \\
&\le 2\|\nabla^+ F(x)\|_W^2 + 2\|\nabla^+ F(\mathbf{q}) - \nabla^+ F(x)\|_W^2 \\
&\le 6\left(\|\nabla^+ F(x)\|_W \|x-\bar{x}\|_W + 3\|\mathbf{q}-x\|^2\right).
\end{aligned}$$

From this and Lemma 6.7 we arrive at

$$\begin{aligned}
&\|\mathbf{q}-\bar{\mathbf{q}}\|_W^2 \\
&\le \gamma_2^2 \left(\|\mathbf{q}-\bar{\mathbf{q}}\|_W + \|\nabla f(\bar{x})\|_{W^{-1}}\right)^2 \|\nabla F^+(\mathbf{q})\|_W^2 \\
&\le 6\gamma_2^2 \left(\|\mathbf{q}-\bar{\mathbf{q}}\|_W + \|\nabla f(\bar{x})\|_{W^{-1}}\right)^2 \left(\|\nabla^+ F(x)\|_W \|x-\bar{x}\|_W + 3\|\mathbf{q}-x\|^2\right).
\end{aligned} \tag{6.18}$$

Note that since $\nabla f(\bar{x})$ is constant on $Q \supseteq X^*$, we define the bound

$$\|\nabla f(\bar{x})\|_{W^{-1}} = \beta \qquad \forall \bar{x} \in X^*.$$

Furthermore, $\bar{\mathbf{q}} \in Q$ since $X^* \subseteq Q$. From this and through the nonexpansive property of the projection operator we obtain

$$\begin{aligned}
\|\mathbf{q}-\bar{\mathbf{q}}\|_W &\le \|\mathbf{q}-\bar{x}\|_W + \|\bar{x}-\bar{\mathbf{q}}\|_W \le \|x-\bar{x}\|_W + \|\bar{x}-\bar{\mathbf{q}}\|_W \le \|x-\bar{x}\|_W + \|x-\mathbf{q}\|_W \\
&\le 2\|x-\bar{x}\|_W + \|\bar{x}-\mathbf{q}\|_W \le 3\|x-\bar{x}\|_W.
\end{aligned}$$

From this and (6.18) we get the following bound:

$$\begin{aligned}(6.19)\qquad &\|\mathbf{q}-\bar{\mathbf{q}}\|_W^2 \\ &\leq 6\gamma_2^2(3\|x-\bar{x}\|_W+\beta)^2\left(\|\nabla^+F(x)\|_W\|x-\bar{x}\|_W+3\|\mathbf{q}-x\|_W^2\right) \\ &\leq 6\gamma_2^2(18\|x-\bar{x}\|_W^2+2\beta^2)\left(\|\nabla^+F(x)\|_W\|x-\bar{x}\|_W+3\|\mathbf{q}-x\|_W^2\right).\end{aligned}$$

Given the definition of $\bar{x}$, we have that

$$\|x-\bar{x}\|_W^2 \leq \|x-\bar{\mathbf{q}}\|_W^2 \leq \left(\|x-\mathbf{q}\|_W+\|\mathbf{q}-\bar{\mathbf{q}}\|_W\right)^2 \leq 2\|x-\mathbf{q}\|_W^2+2\|\mathbf{q}-\bar{\mathbf{q}}\|_W^2.$$

From Lemma 6.5 and (6.19), we can establish an upper bound for the right-hand side of the above inequality:

$$(6.20)\qquad \|x-\bar{x}\|_W^2 \leq (\kappa_1+\kappa_2\|x-\bar{x}\|_W^2)\|\nabla^+F(x)\|_W\|x-\bar{x}\|_W,$$

where

$$\kappa_1 = 24\gamma_2^2\beta^2\left(1+\frac{6\gamma_1^2}{\sigma}\right)+\frac{4\gamma_1^2}{\sigma} \quad \text{and} \quad \kappa_2 = 256\gamma_2^2\left(1+\frac{6\gamma_1^2}{\sigma}\right).$$

If we divide both sides of (6.20) by $\|x-\bar{x}\|_W$, the proof is complete. ∎

6.3. Case 3: Ψ polyhedral function. We now consider general optimization problems of the form

$$(6.21)\qquad \min_{x\in\mathbb{R}^n} F(x) \qquad \left(= \tilde{f}(Px)+c^Tx+\Psi(x)\right),$$

where $\Psi(x)$ is a polyhedral function. A function $\Psi:\mathbb{R}^n\to\mathbb{R}$ is polyhedral if its epigraph, epi $\Psi=\{(x,\zeta):\Psi(x)\leq\zeta\}$, is a polyhedral set. There are numerous functions Ψ which are polyhedral, e.g., $\mathbf{I}_X(x)$ with X a polyhedral set, $\|x\|_1$, $\|x\|_\infty$, or combinations of these functions. Note that an objective function with the structure (6.21) appears in many applications (see, e.g., the constrained lasso problem (2.8) in section 2.1). Now, for proving the generalized error bound property, we require that the objective function F satisfy the following assumption.

Assumption 3. We consider that $f(x)=\tilde{f}(Px)+c^Tx$ satisfies Assumption 1. Further, we assume that $\tilde{f}(z)$ is σ-strongly convex in z and the optimal set X^* is bounded. We also assume that $\Psi(x)$ is bounded above on its domain, i.e., $\Psi(x)\leq\bar{\Psi}<\infty$ for all $x\in\operatorname{dom}\Psi$, and is L_Ψ-Lipschitz continuous w.r.t. norm $\|\cdot\|_W$.

The proof of the generalized error bound property under Assumption 3 is similar to that of [28], but it requires new proof ideas and is done under different assumptions, e.g., that $\Psi(x)$ is bounded above on its domain. Boundedness of Ψ is in practical applications usually not restrictive. Since $\Psi(x)\leq\bar{\Psi}$ is satisfied for any $x\in\operatorname{dom}\Psi$, then problem (6.21) is equivalent to the following one:

$$\min_{x\in\mathbb{R}^n} \quad f(x)+\Psi(x) \quad \text{s.t. } \Psi(x)\leq\bar{\Psi}.$$

Consider now an additional variable $\zeta\in\mathbb{R}$. Then, the previous problem is equivalent to the following problem:

$$(6.22)\qquad \min_{x\in\mathbb{R}^n,\zeta\in\mathbb{R}} \quad f(x)+\zeta \qquad \text{s.t. } \Psi(x)\leq\zeta,\ \Psi(x)\leq\bar{\Psi}.$$

Take an optimal pair (x^*, ζ^*) for problem (6.22). We now prove that $\zeta^* = \Psi(x^*)$. Consider that (x^*, ζ^*) is strictly feasible, i.e., $\Psi(x^*) < \zeta^*$. Then, we can imply that $(x^*, \Psi(x^*))$ is feasible for (6.22) and the following inequality holds:

$$f(x^*) + \Psi(x^*) < f(x^*) + \zeta^*,$$

which contradicts the fact that (x^*, ζ^*) is optimal. Thus, the only possibility that remains is that $\Psi(x^*) = \zeta^*$. Further, it can be easily proved that (6.22) is equivalent to the following problem:

$$\min_{x \in \mathbb{R}^n, \zeta \in \mathbb{R}} \quad f(x) + \zeta \quad \text{s.t. } \Psi(x) \le \zeta,\ \zeta \le \bar{\Psi}. \tag{6.23}$$

Now, if we denote $z = [x^T\ \zeta]^T$, then problem (6.23) can be rewritten as

$$\min_{z \in Z \subseteq \mathbb{R}^{n+1}} \quad \tilde{F}(z) \quad \left(= \tilde{f}(\tilde{P}z) + \tilde{c}^T z \right), \tag{6.24}$$

where $\tilde{P} = [P\ \ 0]$ and $\tilde{c} = [c^T\ 1]^T$. The constraint set for this problem is

$$Z = \left\{ z = [x^T\ \zeta]^T : z \in \text{epi } \Psi, \zeta \le \bar{\Psi} \right\}.$$

Recall that from Assumption 3 we have that epi Ψ is polyhedral, i.e., there exist a matrix C and a vector d such that we can express epi $\Psi = \left\{(x, \zeta) : C[x^T\ \zeta]^T \le d\right\}$. Thus, we can write the constraint set Z as

$$Z = \left\{ z = [x^T\ \zeta]^T : \begin{bmatrix} C \\ e_{n+1}^T \end{bmatrix} z \le \begin{bmatrix} d \\ \bar{\Psi} \end{bmatrix} \right\};$$

i.e., Z is polyhedral. Denote by Z^* the set of optimal points of problem (6.23). Then, from X^* being bounded in accordance with Assumption 3, and the fact that $\Psi(x^*) = \zeta^*$, with Ψ a continuous function, it can be observed that Z^* is also bounded. We now denote $\bar{z} = \Pi_{Z^*}^{\tilde{W}}(z)$, where $\tilde{W} = \text{diag}(W, 1)$. Since the problems (6.22) and (6.24) are equivalent, we can apply the theory of the previous subsection to problem (6.24). That is, we can find two nonnegative constants $\tilde{\kappa_1}$ and $\tilde{\kappa_2}$ such that

$$\|z - \bar{z}\|_{\tilde{W}} \le (\tilde{\kappa}_1 + \tilde{\kappa}_2 \|z - \bar{z}\|_{\tilde{W}^2}) \, \|\nabla^+ \tilde{F}(z)\|_{\tilde{W}} \quad \forall z \in Z. \tag{6.25}$$

The proximal gradient mapping in this case, $\nabla^+ \tilde{F}(z)$, is defined as

$$\nabla^+ \tilde{F}(z) = z - \Pi_Z^{\tilde{W}} \left(z - \tilde{W}^{-1} \nabla \tilde{F}(z) \right),$$

where the projection operator $\Pi_Z^{\tilde{W}}$ is defined in the same manner as Π_X^W. We now show that from the error bound inequality (6.25) we can derive an error bound inequality for problem (6.21). From the definitions of z, $\bar{z}$, and $\tilde{W}$, we derive the following lower bound for the term on the right-hand side:

$$\|z - \bar{z}\|_{\tilde{W}} = \left\| \begin{matrix} x - \bar{x} \\ \zeta - \bar{\zeta} \end{matrix} \right\|_{\tilde{W}} \ge \|x - \bar{x}\|_W. \tag{6.26}$$

Further, note that we can express

$$\|z - \bar{z}\|_{\tilde{W}}^2 = \|x - \bar{x}\|_W^2 + (\zeta - \bar{\zeta})^2 = \|x - \bar{x}\|_W^2 + |\zeta - \bar{\zeta}|^2. \tag{6.27}$$

Now, if $\zeta \leq \bar{\zeta}$, then from $\bar{\zeta} = \Psi(\bar{x})$ and the Lipschitz continuity of Ψ we have that

$$|\zeta - \bar{\zeta}| = \bar{\zeta} - \zeta \leq \Psi(\bar{x}) - \Psi(x) \leq L_\Psi \|x - \bar{x}\|_W.$$

Otherwise, if $\zeta > \bar{\zeta}$, we have that

$$|\zeta - \bar{\zeta}| = \zeta - \bar{\zeta} \leq \bar{\Psi} - \bar{\zeta} \leq |\bar{\Psi}| + |\bar{\zeta}| \triangleq \kappa_1'.$$

From these two inequalities we derive the following inequality for $|\zeta - \bar{\zeta}|^2$:

$$|\zeta - \bar{\zeta}|^2 \leq (\kappa_1' + L_\Psi \|x - \bar{x}\|_W)^2 \leq 2\kappa_1'^2 + 2L_\Psi^2 \|x - \bar{x}\|_W^2.$$

Therefore, the following upper bound for $\|z - \bar{z}\|_{\tilde{W}}^2$ is established:

$$\|z - \bar{z}\|_{\tilde{W}}^2 \leq 2\kappa_1'^2 + (2L_\Psi^2 + 1)\|x - \bar{x}\|_W^2. \tag{6.28}$$

We are now ready to present the main result of this section that shows a generalized error bound property for problems (6.21) under general polyhedral functions Ψ.

THEOREM 6.9. *Under Assumption* 3, *problem* (6.21) *with* $F(x) = \tilde{f}(Px) + c^T x + \Psi(x)$ *satisfies the following global generalized error bound property:*

$$\|x - \bar{x}\|_W \leq \left(\kappa_1 + \kappa_2 \|x - \bar{x}\|_W^2\right) \|\nabla^+ F(x)\|_W \quad \forall x \in \operatorname{dom} \Psi, \tag{6.29}$$

where $\kappa_1 = (\tilde{\kappa}_1 + 2\kappa_1'^2 \tilde{\kappa}_2)(2L_\Psi + 1)$ *and* $\kappa_2 = 2\tilde{\kappa}_2(2L_\Psi + 1)(2L_\Psi^2 + 1)$.

Proof. From the previous discussion, it remains to show that we can find an appropriate upper bound for $\|\nabla^+ \tilde{F}(z)\|_{\tilde{W}}$. Given a point $z = [x^T \ \zeta]^T$, it can be observed that the gradient of $\tilde{F}(z)$ is

$$\nabla \tilde{F}(z) = \begin{bmatrix} P^T \nabla \tilde{f}(Px) + c \\ 1 \end{bmatrix} = \begin{bmatrix} \nabla f(x) \\ 1 \end{bmatrix}.$$

Now, denote $z^+ = \Pi_Z^{\tilde{W}}\left(z - \tilde{W}^{-1} \nabla \tilde{F}(z)\right)$. Following the definitions of the projection operator and of $\nabla^+ \tilde{F}$, note that z^+ is expressed as

$$z^+ = \arg \min_{y \in \mathbb{R}^n, \zeta' \in \mathbb{R}} \frac{1}{2} \left\| \begin{matrix} y - \left(x - W^{-1} \nabla f(x)\right) \\ \zeta' - (\zeta - 1) \end{matrix} \right\|_{\tilde{W}}^2 \quad \text{s.t. } \Psi(y) \leq \zeta', \ \zeta' \leq \bar{\Psi}.$$

Furthermore, from the definition of $\|\cdot\|_{\tilde{W}}$, note that we can also express z^+ as

$$\begin{aligned} z^+ = \arg \quad & \min_{y \in \mathbb{R}^n, \zeta' \in \mathbb{R}} \langle \nabla f(x), y - x \rangle + \frac{1}{2}\|y - x\|_W^2 + \frac{1}{2}(\zeta' - \zeta + 1)^2 \\ & \text{s.t.} \quad \Psi(y) \leq \zeta', \ \zeta' \leq \bar{\Psi}. \end{aligned}$$

Also, given the structure of z, consider that $z^+ = [\tilde{T}_{[N]}(x)^T \ \zeta'']^T$. Now, by a simple change of variables, we can define a pair $(\tilde{T}_{[N]}(x), \tilde{\zeta})$ as follows:

$$\begin{aligned} (\tilde{T}_{[N]}(x), \tilde{\zeta}) = \arg \quad & \min_{y \in \mathbb{R}^n, \zeta' \in \mathbb{R}} \langle \nabla f(x), y - x \rangle + \frac{1}{2}\|y - x\|_W^2 + \frac{1}{2}(\zeta' + 1)^2 \\ & \text{s.t.} \quad \Psi(y) - \zeta \leq \zeta', \ \zeta' \leq \bar{\Psi} - \zeta. \end{aligned} \tag{6.30}$$

Note that $\tilde{\zeta} = \zeta'' - \zeta$ and that we can express $z^+ = [\tilde{T}_{[N]}(x)^T \ \ \tilde{\zeta} + \zeta]^T$ and

$$\|\nabla^+ \tilde{F}(z)\|_{\tilde{W}} = \left\| \begin{matrix} x - \tilde{T}_{[N]}(x) \\ -\tilde{\zeta} \end{matrix} \right\|_{\tilde{W}}.$$

From (3.4) and (5.1), we can write $\nabla^+ F(x) = x - T_{[N]}(x)$ and recall that $T_{[N]}(x)$ can be expressed as

$$T_{[N]}(x) = \arg\min_{y\in\mathbb{R}^n} \langle \nabla f(x), y - x\rangle + \frac{1}{2}\|y - x\|_W^2 + \Psi(y) - \Psi(x).$$

Thus, we can consider that $T_{[N]}(x)$ belongs to a pair $(T_{[N]}(x), \hat{\zeta})$ which is the optimal solution of the following problem:

$$(T_{[N]}(x), \hat{\zeta}) = \arg \min_{y\in\mathbb{R}^n, \zeta'\in\mathbb{R}} \langle \nabla f(x), y - x\rangle + \frac{1}{2}\|y - x\|_W^2 + \zeta' \quad \text{s.t.} \quad \Psi(y) - \Psi(x) \le \zeta'. \tag{6.31}$$

Following the same reasoning as in problem (6.22), note that $\hat{\zeta} = \Psi(T_{[N]}(x)) - \Psi(x)$. Through Fermat's rule [26] and problem (6.31), we establish that $(T_{[N]}(x), \hat{\zeta})$ can also be expressed as

$$(T_{[N]}(x), \hat{\zeta}) = \arg \min_{y\in\mathbb{R}^n, \zeta'} \langle \nabla f(x) + W(T_{[N]}(x) - x), y - x\rangle + \zeta' \quad \text{s.t.} \quad \Psi(y) - \Psi(x) \le \zeta'. \tag{6.32}$$

Therefore, since $(T_{[N]}(x), \hat{\zeta})$ is optimal for the problem above, we get the inequality

$$\begin{aligned} &\langle \nabla f(x) + W(T_{[N]}(x) - x), T_{[N]}(x) - x\rangle + \hat{\zeta} \\ &\quad \le \langle \nabla f(x) + W(T_{[N]}(x) - x), \tilde{T}_{[N]}(x) - x\rangle + \tilde{\zeta}. \end{aligned} \tag{6.33}$$

Furthermore, since the pair $(\tilde{T}_{[N]}(x),\ \tilde{\zeta})$ is optimal for problem (6.30), we can derive

$$\begin{aligned} &\langle \nabla f(x), \tilde{T}_{[N]}(x) - x\rangle + \frac{1}{2}\|\tilde{T}_{[N]}(x) - x\|_W^2 + \frac{1}{2}(\tilde{\zeta} + 1)^2 \\ &\le \langle \nabla f(x), T_{[N]}(x) - x\rangle + \frac{1}{2}\|T_{[N]}(x) - x\|_W^2 + \frac{1}{2}(\hat{\zeta} + 1)^2. \end{aligned} \tag{6.34}$$

By adding up (6.33) and (6.34), we get the following relation:

$$\begin{aligned} &\|T_{[N]}(x) - x\|_W^2 + \frac{1}{2}\|\tilde{T}_{[N]}(x) - x\|_W^2 + \frac{1}{2}(\tilde{\zeta} + 1)^2 + \hat{\zeta} \\ &\le \frac{1}{2}\|T_{[N]}(x) - x\|_W^2 + \langle W(T_{[N]}(x) - x), \tilde{T}_{[N]}(x) - x\rangle + \frac{1}{2}(\hat{\zeta} + 1)^2 + \tilde{\zeta}. \end{aligned}$$

If we further simplify this inequality, we obtain

$$\frac{1}{2}\|T_{[N]}(x) - x\|_W^2 + \frac{1}{2}\|\tilde{T}_{[N]}(x) - x\|_W^2 - \langle W(T_{[N]}(x) - x), \tilde{T}_{[N]}(x) - x\rangle + \frac{1}{2}\tilde{\zeta}^2 \le \frac{1}{2}\hat{\zeta}^2.$$

Combining the first three terms in the left-hand side under the norm and if we multiply both sides by 2, the inequality becomes

$$\left\|(T_{[N]}(x) - x) - (\tilde{T}_{[N]}(x) - x)\right\|_W^2 + \tilde{\zeta}^2 \le \hat{\zeta}^2.$$

From this, we derive the following two inequalities:

$$\tilde{\zeta}^2 \le \hat{\zeta}^2 \quad \text{and} \quad \left\|(T_{[N]}(x) - x) - (\tilde{T}_{[N]}(x) - x)\right\|_W^2 \le \hat{\zeta}^2.$$

If we take the square root in both of these inequalities and apply the triangle inequality to the second one, we obtain

$$|\tilde{\zeta}| \leq |\hat{\zeta}| \quad \text{and} \quad \left\|\tilde{T}_{[N]}(x) - x\right\|_W - \|T_{[N]}(x) - x\|_W \leq |\hat{\zeta}|. \tag{6.35}$$

Recall that $\hat{\zeta} = \Psi(T_{[N]}(x)) - \Psi(x)$, and through the Lipschitz continuity of Ψ we have from the first inequality of (6.35) that

$$|\tilde{\zeta}| \leq |\hat{\zeta}| = |\Psi(T_{[N]}(x)) - \Psi(x)| \leq L_\Psi \|T_{[N]}(x) - x\|_W.$$

Furthermore, from the second inequality of (6.35) we obtain

$$\left\|\tilde{T}_{[N]}(x) - x\right\|_W \leq (L_\Psi + 1)\|T_{[N]}(x) - x\|_W.$$

From these, we arrive at the following upper bound on $\|\nabla^+ \tilde{F}(z)\|$:

$$\begin{aligned} \|\nabla^+ \tilde{F}(z)\| = \left\| \begin{matrix} x - \tilde{T}_{[N]}(x) \\ -\tilde{\zeta} \end{matrix} \right\|_{\tilde{W}} &\leq \left\|\tilde{T}_{[N]}(x) - x\right\|_W + |\tilde{\zeta}| \\ &\leq (2L_\Psi + 1)\|T_{[N]}(x) - x\|_W = (2L_\Psi + 1)\|\nabla^+ F(x)\|. \end{aligned} \tag{6.36}$$

Finally, from (6.25), (6.28), and (6.36) we obtain the following error bound property for problem (6.21):

$$\|x - \bar{x}\|_W \leq \left(\kappa_1 + \kappa_2 \|x - \bar{x}\|^2\right) \|\nabla^+ F(x)\|,$$

where $\kappa_1 = (\tilde{\kappa}_1 + 2\kappa_1'^2 \tilde{\kappa}_2)(2L_\Psi + 1)$ and $\kappa_2 = 2\tilde{\kappa_2}(2L_\Psi + 1)(2L_\Psi^2 + 1)$. ∎

6.4. Case 4: Dual formulation. Consider now linearly constrained problems:

$$\min_{u \in \mathbb{R}^m} \; g(u) \quad \text{s.t.} \;\; Au \leq b, \tag{6.37}$$

where $A \in \mathbb{R}^{n \times m}$. In many applications, however, its dual formulation is used since the dual structure of the problem is easier; see, e.g., applications such as network optimization [21] or network control [16]. Now, for proving the generalized error bound property, we require that g satisfy the following assumption.

Assumption 4. We consider that g is σ_g-strongly convex and has L_g-Lipschitz continuous gradient w.r.t. the Euclidean norm and that there exists $\tilde{u}$ such that $A\tilde{u} < b$.

Denoting by g^* the convex conjugate of function g, then from Assumption 4 it follows that g^* is $\frac{1}{L_g}$-strongly convex and has $\frac{1}{\sigma_g}$-Lipschitz gradient [26]. Moreover, from $A\tilde{u} < b$ it follows that the set of optimal Lagrange multipliers is compact [26]. In conclusion, the primal problem (6.37) is equivalent to the following dual problem:

$$\max_{x \in \mathbb{R}^n} -g^*(-A^T x) - \langle x, b \rangle - \Psi(x), \tag{6.38}$$

where $\Psi(x) = \mathbf{I}_{\mathbb{R}^n_+}(x)$ is the set indicator function for the nonnegative orthant $\mathbb{R}^n_+$. From section 6.2, for $P = -A^T$, it follows that the dual problem (6.38) satisfies our (GEBP) property from Definition 5.1.

7. Conclusion. In conclusion, in this paper we have analyzed the convergence properties of a parallel random coordinate descent method for solving a general class of composite minimization problems. Our convergence results are more general than those typically found in the literature, in the sense that we can show convergence rate of the algorithm for larger classes of problems. Finally, our approach allows us to analyze in the same framework several methods: full gradient, serial coordinate descent, and any parallel coordinate descent method in between. Due to space limitations, numerical simulations are not included, but detailed numerical tests will be presented in our future work and in the extended technical report of this paper [11].

REFERENCES

[1] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23 (2013), pp. 2037–2060.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, New York, 2006.

[3] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, *Parallel coordinate descent for l_1-regularized loss minimization*, in Proceedings of the 28th International Conference on Machine Learning, Washington, D.C., 2011.

[4] X. Chen, M. K. Ng, and C. Zhang, *Non-Lipschitz ℓ_p-regularization and box constrained model for image restoration*, IEEE Trans. Image Process., 21 (2012), pp. 4709–4721.

[5] M. Hong, X. Wang, M. Razaviyayn, and Z-Q. Luo, *Iteration Complexity Analysis of Block Coordinate Descent Methods*, Tech. report, University of Minnesota, Minneapolis, MN, 2013; available online from http://arxiv.org/abs/1310.6957.

[6] A. Lewis and D. Leventhal, *Randomized methods for linear constraints: Convergence rates and conditioning*, Math. Oper. Res., 35 (2010), pp. 641–654.

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

[8] Z. Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: A general approach*, Ann. Oper. Res., 46 (1993), pp. 157–178.

[9] I. Necoara, *Random coordinate descent algorithms for multi-agent convex optimization over networks*, IEEE Trans. Automat. Control, 58 (2013), pp. 2001–2012.

[10] I. Necoara and D. Clipici, *Efficient parallel coordinate descent algorithm for convex optimization problems with separable constraints: Application to distributed MPC*, J. Process Contr., 23 (2013), pp. 243–253.

[11] I. Necoara and D. Clipici, *Parallel Coordinate Descent Methods for Composite Minimization: Convergence Analysis and Error Bounds*, Tech. report, University Politehnica Bucharest, Bucharest, Romania, 2013; available online from http://arxiv.org/abs/1312.5302.

[12] I. Necoara and R. Findeisen, *Parallel and distributed random coordinate descent method for convex error bound minimization*, in Proceedings of the American Control Conference, 2015, pp. 527–532.

[13] I. Necoara and V. Nedelcu, *Rate analysis of inexact dual first order methods: Application to dual decomposition*, IEEE Trans. Automat. Control, 59 (2014), pp. 1232–1243.

[14] I. Necoara, Yu. Nesterov, and F. Glineur, *Random Block Coordinate Descent Methods for Linearly Constrained Optimization over Networks*, Tech. report, University Politehnica Bucharest, Bucharest, Romania, 2011; available online from http://arxiv.org/abs/1504.06340.

[15] I. Necoara and A. Patrascu, *A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints*, Comput. Optim. Appl., 57 (2014), pp. 307–337.

[16] I. Necoara and J. A. K. Suykens, *An interior-point Lagrangian decomposition method for separable convex optimization*, J. Optim. Theory Appl., 143 (2009), pp. 567–588.

[17] Yu. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer, Boston, MA, 2004.

[18] Yu. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM J. Optim., 22 (2012), pp. 341–362.

[19] Yu. Nesterov, *Gradient methods for minimizing composite objective functions*, Mathematical Programming, 140 (2013), pp. 125–161.

[20] Z. Peng, M. Yan, and W. Yin, *Parallel and Distributed Sparse Optimization*, Tech. report, Rice University, Houston, TX, 2013.

[21] S. S. Ram, A. Nedić, and V. V. Veeravalli, *Incremental stochastic subgradient algorithms for convex optimization*, SIAM J. Optim., 20 (2009), pp. 691–717.
[22] P. Richtárik and M. Takáč, *Distributed Coordinate Descent Method for Learning with Big Data*, Tech. report, University of Edinburgh, Edinburgh, Scotland, 2013.
[23] P. Richtárik and M. Takáč, *Parallel coordinate descent methods for big data optimization*, Math. Program., DOI:10.1007/s10107-0150901-6, 2015.
[24] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Math. Program., 144 (2014), pp. 1–38.
[25] S. M. Robinson, *Bounds for error in the solution set of a perturbed linear program*, Linear Algebra Appl., 6 (1973), pp. 69–81.
[26] R. T. Rockafellar and R. J. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[27] S. Ryali, K. Supekar, D. A. Abrams, and V. Menone, *Sparse logistic regression for whole-brain classification of FMRI data*, NeuroImage, 51 (2010), pp. 752–764.
[28] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423.
[29] P. W. Wang and C. J. Lin, *Iteration complexity of feasible descent methods for convex optimization*, J. Mach. Learn. Res., 14 (2014), pp. 1523–1548.