On the Complexity of Parallel Coordinate Descent

Optimization Methods and Software

ISSN: 1055-6788 (Print) 1029-4937 (Online) Journal homepage: http://www.tandfonline.com/loi/goms20

On the complexity of parallel coordinate descent

Rachael Tappenden, Martin Takáč & Peter Richtárik

To cite this article: Rachael Tappenden, Martin Takáč & Peter Richtárik (2017): On the complexity of parallel coordinate descent, Optimization Methods and Software, DOI: 10.1080/10556788.2017.1392517

To link to this article: http://dx.doi.org/10.1080/10556788.2017.1392517

Published online: 03 Nov 2017.

Optimization Methods & Software, 2017
https://doi.org/10.1080/10556788.2017.1392517

On the complexity of parallel coordinate descent

Rachael Tappenden[a*], Martin Takáč[b] and Peter Richtárik[c]

[a]*School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand;* [b]*Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA;* [c]*School of Mathematics, University of Edinburgh, Edinburgh, UK*

(*Received 2 November 2016; accepted 5 October 2017*)

In this work we study the parallel coordinate descent method (PCDM) proposed by Richtárik and Takáč [*Parallel coordinate descent methods for big data optimization*, Math. Program. Ser. A (2015), pp. 1–52] for minimizing a regularized convex function. We adopt elements from the work of Lu and Xiao [*On the complexity analysis of randomized block-coordinate descent methods*, Math. Program. Ser. A 152(1–2) (2015), pp. 615–642], and combine them with several new insights, to obtain sharper iteration complexity results for PCDM than those presented in [Richtárik and Takáč, *Parallel coordinate descent methods for big data optimization*, Math. Program. Ser. A (2015), pp. 1–52]. Moreover, we show that PCDM is monotonic in expectation, which was not confirmed in [Richtárik and Takáč, *Parallel coordinate descent methods for big data optimization*, Math. Program. Ser. A (2015), pp. 1–52], and we also derive the first high probability iteration complexity result where the initial levelset is unbounded.

Keywords: block coordinate descent; parallelization; iteration complexity; composite minimization; convex optimization; rate of convergence; unbounded levelset; monotonic algorithm

Mathematics Subject Classification: 65K05; 90C05; 90C06; 90C25

1. Introduction

Block coordinate descent methods are being thrust into the optimization spotlight because of a dramatic increase in the size of real world problems, and because of the 'Big data' phenomenon. It is little wonder, when these seemingly simple methods, with low iteration costs and low memory requirements, can solve problems where the dimension is more than one billion, in a matter of hours [24].

There is an abundance of coordinate descent variants arising in the literature including: [2,6,9,11,12,15,16,22,26–28,31–38]. The main differences between these methods is the way in which the block of coordinates to update is chosen, and also how the subproblem to determine the update to apply a block of variables is to be solved. One of the current, state-of-the-art block coordinate descent method is the Parallel (block) Coordinate Descent Method (PCDM) of Richtárik and Takáč [24]. This method selects the coordinates to update *randomly* and the update is determined by *minimizing an overapproximation of the objective function* at the current point (see

*Corresponding author. Email: rachael.tappenden@canterbury.ac.nz

Section 3 for a detailed description). PCDM can be applied to a problem with a general convex composite objective, it is supported by iteration complexity results to guarantee the method's convergence, and it has been tested numerically on a wide range of problems to demonstrate its practical capabilities.

In this work we are interested in the following convex composite/regularized optimization problem

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

where we assume that $f(x)$ is a continuously differentiable convex function, and $\Psi(x)$ is assumed to be a (possibly non-smooth) block separable convex regularizer.

The Expected Separable Overapproximation (ESO) assumption introduced in [24] enabled the development of a unified theoretical framework that guarantees convergence of a serial [23], parallel [24] and even distributed [4,14,25] version of PCDM. To benefit from the ESO abstraction, we derive all the results in this paper based on the assumption that f admits an ESO with respect to a uniform block sampling $\hat{S}$. This concept will be precisely defined in Section 3.2. For now it is enough to say that updating a random set of τ coordinates (selected uniformly at random) is one particular uniform sampling and the ESO enables us to overapproximate the expected value of the function at the next iteration by a separable function, which is easy to minimize in parallel.

1.1 *Brief literature review*

Nesterov [17] provided some of the earliest iteration complexity results for a serial Randomized Coordinate Descent Method (RCDM) for problems of the form (1), where $\Psi \equiv 0$, or is the indicator function for simple bound constraints. Later, Richtárik and Takáč generalized this work to optimization problems with a composite objective of the form (1), where the function Ψ is any (possibly non-smooth) convex block separable function [23]; their algorithm is called the Uniform Coordinate Descent for Composite functions (UCDC) algorithm. Xiao and Lu [39] have combined and extended some of the ideas from [17] and [23] to tighten the complexity results for UCDC. In particular, they adopted the idea of a gradient mapping (developed in Nesterov [17]) and married this with the technical analysis in [23] resulting in an improved constant in the iteration complexity result for UCDC.

One of the main advantages of randomized coordinate descent methods is that each iteration is extremely cheap, and can require as little as a few multiplications in some cases [22]. However, a large number of iterations may be required to obtain a sufficiently accurate solution, and for this reason, parallelization of coordinate descent methods is essential.

The SHOTGUN algorithm presented in [1] represents a naïve way of parallelizing RCDM, applied to functions of the form (1) where $\Psi \equiv \|\cdot\|_1$. They also present theoretical results to show that parallelization can lead to algorithm speedup. Unfortunately, their results show that only a small number of coordinates should be updated in parallel at each iteration, otherwise there is no guarantee of algorithm speedup.

The first true complexity analysis of Parallel RCDM (PCDM) was provided in [24] after the authors developed the concept of an ESO assumption, which was central to their convergence analysis. The ESO gives an upper bound on the expected value of the objective function after a parallel update of PCDM has been performed, and depends on both the objective function, and the particular 'sampling' (way that the coordinates are chosen) that was used. Moreover, several distributed PCDMs were considered in [4,14,25] and their convergence was proved simply by deriving the ESO parameters for particular distributed samplings.

In [3,10] the accelerated PCDM was presented and its efficient distributed implementation was considered in [4]. Recently, there has also been a focus on PCDMs that use an arbitrary sampling of coordinates [19–21,26].

One of the goals of this work is to incorporate some of the ideas from [17,39] and extend them from a serial, to a *parallel* setting. We will provide improved iteration complexity results for PCDM [24], through the development of a smaller constant in the complexity bound, which reinforces the strength of PCDM.

1.2 *Summary of contributions*

In this section we summarize the main contributions of this paper (not in order of significance).

(1) *No need to enforce 'monotonicity'.* PCDM in [24] was analysed (for a general convex composite function of the form (1)) under a monotonicity assumption; if, at any iteration of PCDM, an update was computed that would lead to a higher objective value than the objective value at the current point, then that update is rejected. Hence, PCDM presented in [24] included a step to force monotonicity of the function values at each iteration. In this paper we confirm that the monotonicity test is redundant, and can be removed from the algorithm. This is crucial because computing function values can be prohibitively expensive in a large-scale context, and our new monotonicity result means that *no function values need be computed* in PCDM. Clearly, while this monotonicity result is interesting from a theoretical perspective, it also leads to this significant practical implication as well.

(2) *First high-probability results for PCDM without levelset information.* Currently, the high probability iteration complexity results for coordinate descent type methods require the levelset to be bounded. In this paper we derive the first high-probability result which *does not rely on the size of the levelset*.[1] In particular, the analysis of PCDM in [24] assumes that the levelset $\{x \in \mathbf{R}^N : F(x) \leq F(x_0)\}$ is bounded for the initial point x_0, and under this assumption, convergence is guaranteed. However, in this paper we show that PCDM will converge in expectation to the optimal solution even if the levelset is unbounded (see Section 5).

(3) *Sharper iteration complexity results.* In this work we obtain sharper iteration complexity results for PCDM than those presented in [24], and Table 1 summarizes our findings. A thorough discussion of the results can be found in Section 6.2. We briefly describe the variables used in the table (all will be properly defined in later sections.) Variable c is a constant, k is the iteration counter, $\alpha \in [0, 1]$ is the expected proportion of coordinates updated at each iteration, $\xi_0 = F(x_0) - F_*$, and v is a (vector) parameter of the method. Also, μ_f and μ_Ψ are the (strong) convexity constants of f and Ψ respectively (both with respect to $\|\cdot\|_v$ for some v) and ϵ and ρ are the desired accuracy and confidence level respectively. (C = Convex, SC = Strongly Convex.)

(4) *Improved convergence rates for PCDM.* In this work we show that PCDM converges at a faster rate than that given in [24] in both the convex and strongly convex cases. Table 2 provides a summary of our results and a thorough discussion can be found in Section 6.1.

1.3 *Paper outline*

The remainder of this paper is structured as follows. In Section 2 we introduce the notation and assumptions that will be used throughout the paper. Section 3 describes PCDM of Richtárik and Takáč [24] in detail. We also present a new convergence rate result for PCDM, which is sharper than that presented in [24]. The proof of the result is given in Section 4 along with several necessary technical lemmas.

In Section 5 we present several iteration complexity results, which show that PCDM will converge to an ϵ-optimal solution with high probability. In Section 5.1 we provide the first iteration complexity result for PCDM that does not require the assumption of a bounded levelset. The results shows that PCDM requires $\mathcal{O}(1/\rho)$ iterations, so we have devised a 'multiple run

Table 1. Comparison of the iteration complexity results for PCDM obtained in [24] and in this paper.

F	Richtárik and Takáč [24]	This paper	Theorem
C	$\frac{2c}{\alpha\epsilon}\left(1+\log\left(\frac{1}{\rho}\right)\right)+2-\frac{2c}{\alpha\xi_0}$	$\frac{2c}{\alpha\epsilon}\left(1+\log\left(\frac{\frac{1}{2}\|\|x_0-x^*\|\|_v^2+\xi_0}{2c\rho}\right)\right)+2-\frac{1}{\alpha}$	13(i)
SC	$\frac{1+\mu_\Psi}{\alpha(\mu_f+\mu_\Psi)}\log\left(\frac{\xi_0}{\epsilon\rho}\right)$	$\frac{1+\mu_f+2\mu_\Psi}{2\alpha(\mu_f+\mu_\Psi)}\log\left(\frac{\frac{1+\mu_\Psi}{2}\|\|x_0-x^*\|\|_v^2+\xi_0}{\epsilon\rho}\right)$	13(ii)

Note: The analysis used in this paper provides a sharper iteration complexity result in both the convex and strongly convex cases when ϵ and/or ρ are small.

Table 2. Comparison of the convergence rates for PCDM obtained in [24] and in this paper. (C = Convex, SC = Strongly Convex).

F	Richtárik and Takáč [24]	This paper	Theorem
C	$\frac{2c\xi_0}{2c+\alpha k\xi_0}$	$\frac{1}{1+\alpha k}\left(\frac{1}{2}\|\|x_0-x^*\|\|_v^2+\xi_0\right)$	3(i)
SC	$\left(1-\alpha\frac{\mu_f+\mu_\Psi}{1+\mu_\Psi}\right)^k\xi_0$	$\left(1-\frac{2\alpha(\mu_f+\mu_\Psi)}{1+\mu_f+2\mu_\Psi}\right)^k\left(\frac{1+\mu_\Psi}{2}\|\|x_0-x^*\|\|_v^2+\xi_0\right)$	3(ii)

Note: The analysis used in this paper provides a better rate of convergence in both the convex and strongly convex cases when ϵ and/or ρ are small.

strategy' that achieves the classical $\mathcal{O}(\log(1/\rho))$ result. Moreover, in Section 5.1 we present a high probability iteration complexity result for PCDM, that assumes boundedness of the levelset, which is sharper than the result given in [24].

In Section 6 we give a comparison of the results derived in this work, with the results given in [24]. Then we present several numerical experiments in Section 7 to highlight the practical capabilities of PCDM under different ESO assumptions. The ESO assumptions are given in the appendix, where we also provide a new ESO for doubly uniform samplings (see Theorem A7).

2. Notation and assumptions

In this section we introduce block structure and associated objects such as norms and projections. The parallel (block) coordinate descent method will operate on blocks instead of coordinates.

2.1 *Block structure*

The problem under consideration is assumed to have block structure and this is modelled by decomposing the space $\mathbf{R}^N$ into n subspaces as follows. Let $U \in \mathbf{R}^{N\times N}$ be a column permutation of the $N \times N$ identity matrix and further let $U = [U_1, U_2, \ldots, U_n]$ be a decomposition of U into n submatrices, where U_i is $N \times N_i$ and $\sum_{i=1}^n N_i = N$. Note that $U_i^{\mathrm{T}}U_j = I_{N_i}$ when $i = j$ and $U_i^{\mathrm{T}}U_j = \mathbf{0}$ (where $\mathbf{0}$ is the $N_i \times N_j$ matrix of all zeros) when $i \neq j$. Subsequently, any vector $x \in \mathbf{R}^N$ can be written uniquely as

$$x = \sum_{i=1}^{n} U_i x^{(i)}, \tag{2}$$

where $x^{(i)} = U_i^{\mathrm{T}}x \in \mathbf{R}^{N_i}$. For simplicity we will write $x = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})^{\mathrm{T}}$.

In what follows let $\langle \cdot, \cdot \rangle$ denote the standard Euclidean inner product. Then we have

$$\langle x, y \rangle = \left\langle \sum_{i=1}^{n} U_i x^{(i)}, \sum_{j=1}^{n} U_j y^{(j)} \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle U_j^{\mathrm{T}} U_i x^{(i)}, y^{(j)} \rangle \equiv \sum_{i=1}^{n} \langle x^{(i)}, y^{(i)} \rangle. \tag{3}$$

Norms. Further we equip $\mathbf{R}^{N_i}$ with a pair of conjugate Euclidean norms:

$$\|h\|_{(i)} := \langle B_i h, h \rangle^{1/2}, \quad \|h\|_{(i)}^* = \langle B_i^{-1} h, h \rangle^{1/2}, \quad h \in \mathbf{R}^{N_i}, \tag{4}$$

where $B_i \in \mathbf{R}^{N_i \times N_i}$ is a positive definite matrix. For fixed positive scalars $v_1, v_2, \ldots, v_n$, let $v = (v_1, \ldots, v_n)^{\mathrm{T}}$ and define a pair of conjugate norms in $\mathbf{R}^N$ by

$$\|x\|_v^2 := \sum_{i=1}^{n} v_i \|x^{(i)}\|_{(i)}^2, \quad (\|y\|_v^*)^2 := \max_{\|x\|_v \le 1} \langle y, x \rangle^2 = \sum_{i=1}^{n} \frac{1}{v_i} (\|y^{(i)}\|_{(i)}^*)^2. \tag{5}$$

Projection onto a set of blocks Let $\emptyset \neq S \subseteq \{1, 2, \ldots, n\}$. Then for $x \in \mathbf{R}^N$ we write

$$x_{[S]} := \sum_{i \in S} U_i x^{(i)} \tag{6}$$

and we define $x_{[\emptyset]} \equiv 0$. That is, given $x \in \mathbf{R}^N$, $x_{[S]}$ is the vector in $\mathbf{R}^N$ whose blocks $i \in S$ are identical to those of x, but whose other blocks are zeroed out.

2.2 *Assumptions and strong convexity*

Throughout this paper we make the following assumption regarding the block separability of the function Ψ.

ASSUMPTION 1 (Block separability) *The non-smooth function $\Psi : \mathbf{R}^N \to \mathbf{R} \cup \{+\infty\}$ is assumed to be block separable, that is, it can be decomposed as:*

$$\Psi(x) = \sum_{i=1}^{n} \Psi_i(x^{(i)}), \tag{7}$$

where the functions $\Psi_i : \mathbf{R}^{N_i} \to \mathbf{R} \cup \{+\infty\}$ are proper, closed and convex.

In some of the results presented in this work F is assumed to be strongly convex and we use $\mu_F > 0$ to denote the (strong) convexity parameter of F, with respect to the norm $\|\cdot\|_v$ for some $v \in \mathbf{R}_{++}^n$. A function $\phi : \mathbf{R}^N \to \mathbf{R} \cup \{+\infty\}$ is strongly convex with respect to the norm $\|\cdot\|_v$ with convexity parameter $\mu_\phi \ge 0$ if for all $x, y \in \operatorname{dom} \phi$,

$$\phi(y) \ge \phi(x) + \langle \phi'(x), y - x \rangle + \frac{\mu_\phi}{2} \|y - x\|_v^2, \tag{8}$$

where ϕ' is any subgradient of ϕ at x. The case with $\mu_\phi = 0$ reduces to convexity.

Strong convexity of F may come from f or Ψ or both and we will write μ_f (resp. μ_Ψ) for the strong convexity parameter of f (resp. Ψ). Following from (8)

$$\mu_F \ge \mu_f + \mu_\Psi. \tag{9}$$

From the first order optimality conditions for (1) we obtain $\langle F'(x_*), x - x_* \rangle \ge 0$ for all $x \in \operatorname{dom} F$. Combining this with (8) used with $y = x$ and $x = x_*$, yields the standard inequality

$$F(x) - F_* \ge \frac{\mu_F}{2} \|x - x_*\|_v^2, \quad x \in \operatorname{dom} F. \tag{10}$$

3. Parallel coordinate descent method

In this section we describe the PCDM (Algorithm 1) of Richtárik and Takáč [24]. We now present the algorithm, and a detailed discussion will follow.

Algorithm 1 PCDM: Parallel Coordinate Descent Method [24]

1: choose initial point $x_0 \in \mathbf{R}^N$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: randomly choose set of blocks $S_k \subseteq \{1, \ldots, n\}$
4: **for** $i \in S_k$ (**in parallel**) **do**
5: compute $h(x_k)^{(i)} = \arg\min_{t \in \mathbf{R}^{N_i}} \left\{ \langle (\nabla f(x_k))^{(i)}, t \rangle + \frac{v_i}{2} \|t\|^2_{(i)} + \Psi_i(x_k^{(i)} + t) \right\}$
6: **end for**
7: apply the update: $x_{k+1} \leftarrow x_k + \sum_{i \in S_k} U_i h(x_k)^{(i)}$
8: **end for**

The algorithm can be described as follows. At iteration k of Algorithm 1, a set of blocks S_k is chosen, corresponding to the (blocks of) coordinates that are to be updated. The set of blocks is selected via a *sampling*, which is described in detail in Section 3.1. Then, in Steps 4–6, the updates $h(x_k)^{(i)}$, for all $i \in S_k$, are computed *in parallel*, via a small/low dimensional minimization subproblem. (In Section 3.2, we describe the origin of this subproblem via an ESO.) Finally, in Step 7, the updates $h(x_k)^{(i)}$ are applied to the current point x_k, to give the new point x_{k+1}. Notice that Algorithm 1 *does not require knowledge of objective function values.*

We now describe the key steps of Algorithm 1 (Steps 3 and 4–6) in more detail.

3.1 *Step 3: Sampling*

At the kth iteration of Algorithm 1, a set of indices $S_k \subseteq \{1, \ldots, n\}$ (corresponding to the blocks of x_k to be updated) is selected. Here we briefly explain several schemes for choosing the set of indices S_k; a thorough description can be found in [24]. Formally, S_k is a realization of a *random set-valued mapping* $\hat{S}$ with values in $2^{\{1,\ldots,n\}}$. Richtárik and Takáč [24] have coined the term *sampling* in reference to $\hat{S}$.

In what follows, we will assume that all samplings are *proper*. That is, we assume that $p_i > 0$ for all blocks i, where p_i is the probability that the ith block of x is updated.

We state several sampling schemes now.

(1) *Uniform*: A sampling $\hat{S}$ is uniform if all blocks have the same probability of being updated.
(2) *Doubly uniform*: A doubly uniform sampling is one that generates all sets of equal cardinality with equal probability. That is $\mathbf{P}(S') = \mathbf{P}(S'')$ whenever $|S'| = |S''|$.
(3) *Non-overlapping uniform*: A non-overlapping uniform sampling is one that is uniform and assigns positive probabilities only to sets forming a partition of $\{1, \ldots, n\}$.

In fact, doubly uniform and non-overlapping uniform samplings are special cases of uniform samplings, so in this work all results are proved for uniform samplings. Other samplings, which are also special cases of uniform samplings, are presented in [24], but we omit details of all, except a τ-nice sampling, for brevity. We say that a sampling $\hat{S}$ is τ-nice, if for any

$S \subseteq \{1, 2, \ldots, n\}$ we have

$$\mathbf{P}(\hat{S} = S) = \begin{cases} 0, & \text{if } |S| \neq \tau, \\ \dfrac{\tau!(n-\tau)!}{n!}, & \text{otherwise.} \end{cases} \tag{11}$$

3.2 *Step 3: Computing the step-length*

The block update $h(x_k)^{(i)}$ is chosen in such a way that an upper bound on the expected function value at the next iterate is minimized, with respect to the particular sampling $\hat{S}$ that is used. The construction of the expected upper bound should be (block) separable to ensure efficient parallelizability. Before focusing on how to construct the expected upper-bound on F we will state a definition of an ESO.

DEFINITION 2 (ESO; Definition 5 in [24]) *Let $v \in \mathbf{R}^n_{++}$ and $\hat{S}$ be a proper uniform sampling. We say that $f : \mathbf{R}^N \to \mathbf{R}$ admits an ESO with respect to the sampling $\hat{S}$ with parameter v, if, for all $x, h \in \mathbf{R}^N$ the following inequality holds:*

$$\mathbf{E}[f(x + h_{[\hat{S}]})] \leq f(x) + \frac{\mathbf{E}[|\hat{S}|]}{n}\left(\langle \nabla f(x), h\rangle + \frac{1}{2}\|h\|_v^2\right). \tag{12}$$

We say that the ESO is monotonic if $\forall S \in \hat{S}$ such that $\mathbf{P}(S = \hat{S}) > 0$ the following holds:

$$f(x + h_{[S]}) \leq f(x).$$

In the appendix, a review of different smoothness assumptions on f and corresponding ESO parameters v for a doubly uniform sampling, is given. In all that follows, we assume that f admits an ESO with ESO parameter v, and $\hat{S}$ is a proper uniform sampling. Then

$$\begin{aligned} \mathbf{E}[F(x + h_{[\hat{S}]})] &\overset{D_F}{=} \mathbf{E}[f(x + h_{[\hat{S}]})] + \mathbf{E}[\Psi(x + h_{[\hat{S}]})] \\ &\overset{(12)(25)}{\leq} f(x) + \tfrac{\mathbf{E}[|\hat{S}|]}{n}\left(\langle \nabla f(x), h\rangle + \frac{1}{2}\|h\|_v^2\right) + \left(1 - \frac{\mathbf{E}[|\hat{S}|]}{n}\right)\Psi(x) \\ &\quad + \frac{\mathbf{E}[|\hat{S}|]}{n}\Psi(x + h), \end{aligned} \tag{13}$$

where we have used that fact that Ψ is block separable and that $\hat{S}$ is a proper uniform sampling (see [24, Theorem 4]).

Now, it is easy to see that minimizing the right-hand side of (13) in h is the same as minimizing the function $\mathcal{H}_v$ in h, where $\mathcal{H}_v$ is defined to be

$$\mathcal{H}_v(x, h) := f(x) + \langle \nabla f(x), h\rangle + \tfrac{1}{2}\|h\|_v^2 + \Psi(x + h). \tag{14}$$

In view of (2), (5), and (7), we can write

$$\mathcal{H}_v(x, h) := f(x) + \sum_{i=1}^{n}\left\{\langle (\nabla f(x))^{(i)}, h^{(i)}\rangle + \frac{v_i}{2}\|h^{(i)}\|_{(i)}^2 + \Psi_i(x^{(i)} + h^{(i)})\right\}.$$

Further, we define

$$h(x) := \arg\min_{h \in \mathbf{R}^N} \mathcal{H}_v(x, h), \tag{15}$$

which is the update used in Algorithm 1. Notice that the algorithm *never evaluates function values.*

3.3 *Complexity of PCDM*

We are now ready to present one of our main results, which is an improved iteration complexity result for PCDM. Our result holds in the *parallel* case, so it is a generalization of Theorem 1 in [39], which only applies in the *serial* case. The result shows that PCDM converges in expectation and provides a sharper convergence rate than that given in [24]. The proof is provided in Section 4. Let us mention that a similar result was given independently[2] in [15], but that result *only holds for the particular ESO described in Theorem A.9*. However, even for that ESO, our result (Theorem 3) is still much better because it depends on $\|x_0 - x_*\|_v$ and not on the size of the initial levelset (see (34) and (40)), which could even be unbounded. We state our result now.

THEOREM 3 *Let F^* be the optimal value of problem* (1)*, and let $\{x_k\}_{k\geq 0}$ be the sequence of iterates generated by PCDM using a uniform sampling $\hat{S}$. Let $\alpha = \mathbf{E}[|\hat{S}|]/n$ and suppose that f admits an ESO with respect to the sampling $\hat{S}$ with parameter v. Then for any $k \geq 0$,*

(i) *the iterate x_k satisfies*

$$\mathbf{E}[F(x_k) - F_*] \leq \frac{1}{1+\alpha k}\left(\frac{1}{2}\|x_0 - x_*\|_v^2 + F(x_0) - F_*\right), \tag{16}$$

(ii) *if $\mu_f + \mu_\Psi > 0$, then the iterate x_k satisfies*

$$\mathbf{E}[F(x_k) - F_*] \leq \left(1 - \frac{2\alpha(\mu_f + \mu_\Psi)}{1 + \mu_f + 2\mu_\Psi}\right)^k \left(\frac{1+\mu_\Psi}{2}\|x_0 - x_*\|_v^2 + F(x_0) - F_*\right). \tag{17}$$

Remark 4 Notice that Theorem 3 is a *general* result, in the sense that *any ESO can be used for PCDM* and the result holds.

4. Proof of the main result

In this section we provide a proof of our main convergence rate result, Theorem 3. However, first we will present several preliminary results, including the idea of a composite gradient mapping, and other technical lemmas.

4.1 *Block composite gradient mapping*

We now define the concept of a *block composite gradient mapping* [18,39]. By the first-order optimality conditions for problem (15), there exists a subgradient $s^{(i)} \in \partial\Psi_i(x^{(i)} + (h(x))^{(i)})$ (where $\partial\Psi_i(\cdot)$ denotes the subdifferential of $\Psi_i(\cdot)$) such that

$$(\nabla f(x))^{(i)} + v_i B_i (h(x))^{(i)} + s^{(i)} = 0. \tag{18}$$

We define the block composite gradient mapping as

$$(g(x))^{(i)} := -v_i B_i (h(x))^{(i)}, \quad i = 1, \ldots, n. \tag{19}$$

From (18) and (19) we obtain

$$-(\nabla f(x))^{(i)} + (g(x))^{(i)} \in \partial\Psi_i(x^{(i)} + (h(x))^{(i)}), \quad i = 1, \ldots, n. \tag{20}$$

If we let $g(x) := \sum_{i=1}^n U_i(g(x))^{(i)}$ (compare (2) and (19)), then since Ψ is separable, (20) can be written as

$$-\nabla f(x) + g(x) \in \partial\Psi(x + h(x)). \tag{21}$$

Moreover

$$\|h(x)\|_v^2 \overset{(5)}{=} \sum_{i=1}^n v_i \|(h(x))^{(i)}\|_{(i)}^2 \overset{(19)}{=} \sum_{i=1}^n \frac{1}{v_i}\|B_i^{-1}(g(x))^{(i)}\|_{(i)}^2 \overset{(4)+(5)}{=} (\|g(x)\|_v^*)^2, \tag{22}$$

and

$$\langle g(x), h(x)\rangle \overset{(3)+(19)}{=} -\|h(x)\|_v^2 \overset{(22)}{=} -(\|g(x)\|_v^*)^2. \tag{23}$$

Finally, note that using (4), (5), (19) and (22), we get

$$\|x + h(x) - y\|_v^2 = \|x - y\|_v^2 + 2\langle g(x), y - x\rangle + (\|g(x)\|_v^*)^2. \tag{24}$$

4.2 *Main technical lemmas*

The following result concerns the expected value of a block-separable function when a random subset of coordinates is updated.

LEMMA 5 (Theorem 4 in [24]) *Suppose that* $\Psi(x) = \sum_{i=1}^n \Psi_i(x^{(i)})$. *For any* $x, h \in \mathbf{R}^N$, *if we choose a uniform sampling* $\hat{S}$, *then letting* $\alpha = \mathbf{E}[|\hat{S}|]/n$, *we have*

$$\mathbf{E}[\Psi(x + (h(x))_{[\hat{S}]})] = \alpha\Psi(x + h(x)) + (1 - \alpha)\Psi(x). \tag{25}$$

The following technical lemma plays a central role in our analysis. The result can be viewed as a generalization of Lemma 3 in [39], which considers the serial case ($\alpha = 1$), to the parallel setting.

LEMMA 6 *Let* $x \in \operatorname{dom} F$ *and* $x_+ = x + (h(x))_{[\hat{S}]}$, *where* $\hat{S}$ *is any uniform sampling. Then for any* $y \in \operatorname{dom} F$,

$$\mathbf{E}[F(x_+) + \tfrac{\mu_\Psi + 1}{2}\|x_+ - y\|_v^2] \le F(x) + \frac{\mu_\Psi + 1}{2}\|x - y\|_v^2 - \alpha\left(F(x) - F(y) + \frac{\mu_f + \mu_\Psi}{2}\|x - y\|_v^2\right). \tag{26}$$

Moreover,

(i)

$$\mathbf{E}[F(x_+)] \le F(x) - \frac{\alpha}{2}(\mu_\Psi + 1)\|h(x)\|_v^2 = F(x) - \frac{\alpha}{2}(\mu_\Psi + 1)(\|g(x)\|_v^*)^2, \tag{27}$$

(ii)

$$\mathbf{E}[F(x_+) + \tfrac{1}{2}\|x_+ - y\|_v^2] \le F(x) + \tfrac{1}{2}\|x - y\|_v^2 - \alpha(F(x) - F(y)). \tag{28}$$

Proof We first note that

$$\mathbf{E}[\|x_+ - y\|_v^2] = \alpha\|x + h(x) - y\|_v^2 + (1 - \alpha)\|x - y\|_v^2. \tag{29}$$

This is a special case of the identity $\mathbf{E}[\psi(u + h_{[\hat{S}]})] = \alpha\psi(u + h) + (1 - \alpha)\psi(u)$ (see Lemma 5, which holds for block separable functions ψ), with $\psi(u) = \|u\|_v^2$, $u = x - y$ and $h = h(x)$.

Further, for any h for which $x + h \in \text{dom}\,\Psi$, we have

$$\mathbf{E}[F(x + h_{[\hat{S}]})] \overset{(30)}{\leq} (1-\alpha)F(x) + \alpha\mathcal{H}_v(x,h). \tag{30}$$

This was established in [24, Section 5]. The claim now follows by combining (30), used with $h = h(x)$, and the following estimate of $\mathcal{H}_v(x, h(x))$:

$$\begin{aligned}
\mathcal{H}_v(x,h(x)) &\overset{(4)}{=} f(x) + \langle \nabla f(x), h(x)\rangle + \frac{1}{2}\|h(x)\|_v^2 + \Psi(x+h(x)) \\
&\overset{(8)+(21)}{\leq} f(y) + \langle \nabla f(x), x-y\rangle - \frac{\mu_f}{2}\|y-x\|_v^2 + \langle \nabla f(x), h(x)\rangle + \frac{1}{2}\|h(x)\|_v^2 \\
&\quad + \Psi(y) + \langle -\nabla f(x) + g(x), x + h(x) - y\rangle - \frac{\mu_\Psi}{2}\|x+h(x)-y\|_v^2 \\
&= F(y) + \langle g(x), x-y\rangle + \langle g(x), h(x)\rangle - \frac{\mu_f}{2}\|y-x\|_v^2 \\
&\quad - \frac{\mu_\Psi}{2}\|x+h(x)-y\|_v^2 + \frac{1}{2}\|h(x)\|_v^2 \\
&\overset{(23)}{=} F(y) + \langle g(x), x-y\rangle - \frac{\mu_f}{2}\|y-x\|_v^2 - \frac{\mu_\Psi}{2}\|x+h(x)-y\|_v^2 - \frac{1}{2}(\|g(x)\|_v^*)^2 \\
&\overset{(24)}{=} F(y) + \frac{1-\mu_f}{2}\|y-x\|_v^2 - \frac{\mu_\Psi+1}{2}\|x+h(x)-y\|_v^2 \\
&\overset{29}{=} F(y) + \frac{1-\mu_f}{2}\|y-x\|_v^2 - \frac{\mu_\Psi+1}{2\alpha}(\mathbf{E}[\|x_+ - y\|_v^2] - (1-\alpha)\|x-y\|_v^2).
\end{aligned}$$

Part (i) follows by letting $x = y$ and using (29) and (23). Part (ii) follows as a special case by choosing $\mu_f = \mu_\Psi = 0$. ■

Property (i) means that function values $F(x_k)$ of PCDM are monotonically decreasing in expectation when conditioned on the previous iteration.

4.3 *Proof of Theorem 3*

Proof Let x_* be an arbitrary optimal solution of (1). Let $r_k^2 = \|x_k - x_*\|_v^2$, $g_k = g(x_k)$, $h_k = h(x_k)$ and $F_k = F(x_k)$. Notice that $x_{k+1} = x_k + (h_k)_{[S_k]}$. By subtracting F_* from both sides of (28), we get

$$\mathbf{E}[\tfrac{1}{2}r_{k+1}^2 + F_{k+1} - F_* \,|\, x_k] \leq (\tfrac{1}{2}r_k^2 + F_k - F_*) - \alpha(F_k - F_*),$$

and taking expectations with respect to the whole history of realizations of $S_l, l \leq k$ gives us

$$\mathbf{E}[\tfrac{1}{2}r_{k+1}^2 + F_{k+1} - F_*] \leq \mathbf{E}[\tfrac{1}{2}r_k^2 + F_k - F_*] - \alpha\mathbf{E}[F_k - F_*].$$

Applying this inequality recursively and using the fact that $\mathbf{E}[F_j]$ is monotonically decreasing for $j = 0, 1, \ldots, k+1$ (27), we obtain

$$\begin{aligned}
\mathbf{E}[F_{k+1} - F_*] &\leq \mathbf{E}\left[\frac{1}{2}r_{k+1}^2 + F_{k+1} - F_*\right] \leq \frac{1}{2}r_0^2 + F_0 - F_* - \alpha\sum_{j=0}^{k}(\mathbf{E}[F_j] - F_*) \\
&\leq \frac{1}{2}r_0^2 + F_0 - F_* - \alpha(k+1)(\mathbf{E}[F_{k+1}] - F_*),
\end{aligned}$$

which leads to (16).

We now prove (17) under the strong convexity assumption $\mu_f + \mu_\Psi > 0$. From (26) we get

$$\mathbf{E}\left[\frac{1+\mu_\Psi}{2}r_{k+1}^2 + F_{k+1} - F_* \,|\, x_k\right] \le \left(\frac{1+\mu_\Psi}{2}r_k^2 + F_k - F_*\right) - \alpha\left(\frac{\mu_f+\mu_\Psi}{2}r_k^2 + F_k - F_*\right). \tag{31}$$

Notice that for any $0 \le \gamma \le 1$ we have

$$\begin{aligned}\frac{\mu_f+\mu_\Psi}{2}r_k^2 + F_k - F_* &= \gamma\left(\frac{\mu_f+\mu_\Psi}{2}r_k^2 + F_k - F_*\right) + (1-\gamma)\left(\frac{\mu_f+\mu_\Psi}{2}r_k^2 + F_k - F_*\right)\\ &\overset{(9)+(10)}{\ge} \gamma\left(\frac{\mu_f+\mu_\Psi}{2}r_k^2 + F_k - F_*\right) + (1-\gamma)(\mu_f+\mu_\Psi)r_k^2.\end{aligned}$$

Choosing

$$\gamma^* := \frac{2(\mu_f+\mu_\Psi)}{1+\mu_f+2\mu_\Psi} \in [0,1] \tag{32}$$

we obtain

$$\frac{\mu_f+\mu_\Psi}{2}r_k^2 + F_k - F_* \overset{(32)}{\ge} \gamma^*\left(\frac{1+\mu_\Psi}{2}r_k^2 + F_k - F_*\right).$$

Combining the inequality above with (31) gives

$$\mathbf{E}\left[\frac{1+\mu_\Psi}{2}r_{k+1}^2 + F_{k+1} - F_* \,|\, x_k\right] \le (1-\gamma^*\alpha)\left(\frac{1+\mu_\Psi}{2}r_k^2 + F_k - F_*\right). \tag{33}$$

It now only remains to take expectation in x_k on both sides of (33), and (17) follows. ■

5. High probability convergence result

Theorem 3 showed that Algorithm 1 converges to the optimal solution of (1) in expectation. In this section we derive iteration complexity bounds for PCDM for obtaining an ϵ-optimal solution with high probability. Let us mentioned that all existing [17,23,24,39] high-probability results for serial or parallel CD require a bounded levelset, that is, they assume that

$$\mathcal{L}(x_0) = \{x \in \mathbf{R}^N : F(x) \le F(x_0)\} \tag{34}$$

is bounded. In Section 5.1 we present the first high probability result in the case when the levelset can be unbounded (Corollaries 9 and 11). Then in Section 5.2 we derive a sharper high-probability result for PCDM [24] if a bounded levelset is assumed (i.e. $\mathcal{L}(x_0)$ is bounded).

5.1 *Case 1: Possibly unbounded levelset*

We begin by presenting Lemma 7, which will allow us to state *the first high-probability result* (Corollary 9) for a PCDM applied to a convex function that *does not require* the assumption of a *bounded* levelset.

LEMMA 7 *Let x_0 be fixed and $\{x_k\}_{k=0}^\infty$ be a sequence of random vectors in $\mathbf{R}^N$ such that the conditional distribution of x_{k+1} on x_k is the same as conditional distribution of x_{k+1} on the whole history $\{x_i\}_{i=0}^\infty$ (hence we have Markov sequence). Define $r_k = \phi_r(x_k)$ and $\xi_k = \phi_\xi(x_k)$ where*

$\phi_r, \phi_\xi : \mathbf{R}^N \to \mathbf{R}$ are non-negative functions. Further, assume that the following two inequalities hold for any k

$$\mathbf{E}[\tfrac{1}{2}r_{k+1} + \xi_{k+1} \,|\, x_k] \leq \tfrac{1}{2}r_k + (1-\zeta)\xi_k, \tag{35}$$

$$\mathbf{E}[\xi_{k+1}] \leq \xi_k \tag{36}$$

with some known $\zeta \in (0,1)$. For stopping tolerance $\epsilon > 0$ and confidence level $\rho \in (0,1)$, if

$$K \geq \frac{1}{\zeta}\left(\frac{\frac{1}{2}r_0 + \xi_0}{\rho\epsilon} - 1\right), \tag{37}$$

then

$$\mathbf{P}(\xi_K < \epsilon) \geq 1 - \rho.$$

Proof Using (35) we have

$$\mathbf{E}[\xi_k] \leq \mathbf{E}\left[\frac{1}{2}r_k + \xi_k\right] \leq \frac{1}{2}r_0 + \xi_0 - \zeta\sum_{j=0}^{k-1}\mathbf{E}[\xi_j] \overset{(36)}{\leq} \frac{1}{2}r_0 + \xi_0 - k\zeta\mathbf{E}[\xi_k].$$

Hence

$$\mathbf{E}[\xi_k] \leq \frac{\frac{1}{2}r_0 + \xi_0}{1 + k\zeta}. \tag{38}$$

Now, from the Markov inequality we have

$$\mathbf{P}(\xi_K \geq \epsilon) \leq \frac{\mathbf{E}[\xi_K]}{\epsilon} \overset{(38)}{\leq} \frac{1}{\epsilon}\frac{\frac{1}{2}r_0 + \xi_0}{1 + K\zeta} \overset{(37)}{\leq} \rho.$$

■

Naturally, the result $\mathcal{O}(1/\epsilon\rho)$ is very pessimistic and hence one may be concerned about tightness of the lemma. The following example shows that Lemma 7 is, indeed, tight, that is, the bound on K cannot be improved much. (We construct an example that, under the assumptions (35) and (36) (i.e. using the analysis of [39]), requires $\mathcal{O}(1/\epsilon\rho)$ iterations.)

Example 8 (Tightness of Lemma 7) Let $\zeta \in (0,1)$ and $\epsilon > 0$. Fix some small value of $\rho \in (0,1)$ and assume that (r_1, ξ_1) have following distribution:

$$(r_1, \xi_1) = \begin{cases} (0,0), & \text{with probability } 1-\rho \\ (2\vartheta, \epsilon), & \text{otherwise,} \end{cases}$$

where ϑ is chosen in such a way that (35) is satisfied. Then, we can chose it as follows

$$\rho(\vartheta + \epsilon) = \frac{1}{2}r_0 + (1-\zeta)\xi_0 \quad \Rightarrow \quad \vartheta = \frac{\frac{1}{2}r_0 + (1-\zeta)\xi_0}{\rho} - \epsilon.$$

Define, for $k = 1, 2, 3, \ldots$

$$(r_{k+1}, \xi_{k+1}) = \begin{cases} (r_k - 2\zeta\epsilon, \epsilon), & \text{if } r_k \geq 2\zeta\epsilon \\ (0,0), & \text{otherwise.} \end{cases}$$

If

$$K := \left\lfloor \frac{\vartheta}{\zeta\epsilon} \right\rfloor = \frac{1}{\zeta}\left\lfloor \frac{\frac{1}{2}r_0 + (1-\zeta)\xi_0}{\rho\epsilon} - 1 \right\rfloor,$$

then $\mathbf{P}(\xi_K \geq \epsilon) \leq \rho$.

COROLLARY 9 (High probability result without bounded levelset) *Using Lemma 7 with* $\frac{1}{2}r_k = \phi_r(x_k) = \frac{1}{2}\|x_k - x_*\|_v^2$, $\xi_k = \phi_\xi(x_k) = F(x_k) - F_*$ *and* $\zeta = \alpha = \mathbf{E}|\hat{S}|/n$ *then we obtain that if*

$$K \geq \frac{n}{\mathbf{E}|\hat{S}|}\left(\frac{\frac{1}{2}\|x_0 - x_*\|_v^2 + F(x_0) - F_*}{\rho\epsilon} - 1\right) \tag{39}$$

then $\mathbf{P}(F(x_K) - F_* < \epsilon) \geq 1 - \rho$.

The negative aspect of Corollary 9 is the fact that one needs $\mathcal{O}(1/\rho)$ iterations whereas classical results under the bounded levelset assumption require only $\mathcal{O}(\log(1/\rho))$ iterations.

Multiple run strategy Now we present a restarting strategy (which uses some of the ideas in [23]) that will give us a high probability result $\mathcal{O}(\log(1/\rho))$.

LEMMA 10 *Let* $\{x_k\}_{k=0}^\infty$, $\{r_k\}_{k=0}^\infty$ *and* $\{\xi_k\}_{k=0}^\infty$ *be the same as in Lemma 7, with* $\zeta \in (0,1)$, $\epsilon > 0$ *and* $\rho \in (0,1)$. *Assume that we observe* $r = \lceil \log(1/\rho) \rceil$ *different random and independent realizations of this sequence always starting from* x_0, *that is, for any k we have observed* $x_k^1, x_k^2, \ldots, x_k^r$. *Then if each realization continues for*

$$K \geq \frac{1}{\zeta}\left(\frac{\frac{1}{2}r_0 + \xi_0}{\epsilon(1/e)} - 1\right)$$

iterations, then

$$\mathbf{P}\left(\min_{l\in\{1,2,\ldots,r\}} \xi_K^l < \epsilon\right) \geq 1 - \rho.$$

Proof Because the realization are independent then for any $l \in \{1,2,\ldots,r\}$ we have from Lemma 7 that $\mathbf{P}(\xi_K^l \geq \epsilon) \leq 1/e$. Hence

$$\mathbf{P}\left(\min_{l\in\{1,2,\ldots,r\}} \xi_K^l \geq \epsilon\right) = \mathbf{P}\left(\xi_K^1 \geq \epsilon, \xi_K^2 \geq \epsilon, \ldots, \xi_K^r \geq \epsilon\right) = \prod_{l\in\{1,2,\ldots,r\}} \mathbf{P}(\xi_K^l \geq \epsilon) \leq \left(\frac{1}{e}\right)^r \leq \rho.$$

■

COROLLARY 11 *If we run PCDM* $r = \lceil \log(1/\rho) \rceil$ *many times for* $K \geq (n/\mathbf{E}[|\hat{S}|])((\frac{1}{2}\|x_0 - x_*\|_v^2 + F(x_0) - F_*)/(\epsilon(1/e) - 1))$ *each, then the best solution we get, indexed* $l \in \{1,2,\ldots,r\}$, *satisfies* $\mathbf{P}(F(x_K^l) - F_* < \epsilon) \geq 1 - \rho$. *Hence, in total we need* $\lceil (n/\mathbf{E}[|\hat{S}|])((\frac{1}{2}\|x_0 - x_*\|_v^2 + F(x_0) - F_*)/(\epsilon(1/e) - 1)) \rceil \lceil \log(1/\rho) \rceil \sim \mathcal{O}(\log(1/\rho))$ *iterations of PCDM.*

Remark 12 In Lemma 10 and Corollary 11, each run in the multiple restart strategy must begin from the same initial point x_0; this allows us to establish the $\mathcal{O}(\log(1/\rho))$ complexity rate for PCDM without a bounded levelset assumption. A natural question that arises is: 'Can we remove the assumption that all runs begin from the same initial point x_0?' It may be possible for one to establish such a result, but this is an open problem. One of the issues with removing the

assumption that all runs begin at x_0 is that the function values in PCDM reduce in *expectation*. Thus, it is possible that a run could end at a point with a higher function value than $F(x_0)$, and if a new run were to begin at this resulting point, the complication of beginning from a point with a higher function value would need to be overcome.

5.2 *Case 2: Bounded levelset*

The next result, Theorem 13, obtains the rate $\mathcal{O}(\log(1/\rho))$, under the assumption that the levelset is bounded. However, some results will hold only for a modified version of Algorithm 1. In particular, we now present Algorithm 2.

Algorithm 2 PCDM-M: Parallel Coordinate Descent Method [24]

1: choose initial point $x_0 \in \mathbf{R}^N$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: randomly choose set of blocks $S_k \subseteq \{1, \ldots, n\}$
4: **for** $i \in S_k$ (**in parallel**) **do**
5: compute $h(x_k)^{(i)} = \arg\min_{t \in \mathbf{R}^{N_i}} \left\{ \langle (\nabla f(x_k))^{(i)}, t \rangle + \frac{v_i}{2} \|t\|^2_{(i)} + \Psi_i(x_k^{(i)} + t) \right\}$
6: **end for**
7: **if** $F(x_k + \sum_{i \in S_k} U_i h(x_k)^{(i)}) \leq F(x_0)$ **then**
8: apply the update: $x_{k+1} \leftarrow x_k + \sum_{i \in S_k} U_i h(x_k)^{(i)}$
9: **else**
10: set $x_{k+1} \leftarrow x_k$
11: **end if**
12: **end for**

Notice that the first 6 steps of Algorithm 2 are exactly the same as those of Algorithm 1. However, Algorithm 2 forces the iterates to stay in $\mathcal{L}(x_0)$ (steps 7–11).

Distance to the optimal solution set. In order to obtain some of the results in this Section we need the distance to the optimal solution set, inside the levelset, to be finite, that is,

$$\mathcal{R}_{v,0} := \max_{x \in \mathcal{L}(x_0)} \left\{ \max_{x_* \in X^*} \|x - x_*\|_v \right\} < \infty. \tag{40}$$

Note that for any $x_* \in X^*$ (where X^* is a set of optimal solutions) it trivially holds that $\|x_0 - x_*\|_v \leq \mathcal{R}_{v,0}$. Moreover, for some problems the levelset can be unbounded, in which case $\mathcal{R}_{v,0}$ is infinite, whereas if $X^* \neq \emptyset$ then $\|x_0 - x_*\|$ is *always finite*.

THEOREM 13 *Let $\{x_k\}_{k \geq 0}$ be a sequence of iterates generated by*

- *PCDM* (*Algorithm* 1) *if F is strongly convex with $\mu_f(w) + \mu_\Psi(w) > 0$ or F is convex and a monotonic ESO is used,*
- *PCDM-M* (*Algorithm* 2) *if F is convex and a non-monotonic ESO is used.*

Let $0 < \epsilon < F(x_0) - F_*$ and $\rho \in (0, 1)$ be chosen arbitrarily. Define $\alpha = \mathbf{E}[|\hat{S}|]/n$, and let

$$c := \max\{\mathcal{R}^2_{v,0}, F(x_0) - F_*\}. \tag{41}$$

Then

(i) *if F is convex and we choose*

$$K \geq \frac{2c}{\alpha\epsilon}\left(1+\log\left(\frac{\frac{1}{2}\|x_0-x_*\|_v^2+F(x_0)-F_*}{2c\rho}\right)\right)+2-\frac{1}{\alpha}, \tag{42}$$

(ii) *or if F is strongly convex with $\mu_f+\mu_\Psi>0$ and we choose*

$$K \geq \frac{1+\mu_f+2\mu_\Psi}{2\alpha(\mu_f+\mu_\Psi)}\log\left(\frac{\frac{1+\mu_\Psi}{2}\|x_0-x_*\|_v^2+F(x_0)-F_*}{\epsilon\rho}\right) \tag{43}$$

then

$$\mathbf{P}(F(x_K)-F_*<\epsilon)\geq 1-\rho. \tag{44}$$

Proof The proof is similar to that of [23, Theorem 1], so is omitted for brevity. ■

In this section we have presented three new convergence results for PCDM. The first result shows that, combining the analysis techniques in [17,24,39], PCDM obtains a $\mathcal{O}(1/\rho)$ rate when the levelset is unbounded for a single run strategy. The second result shows that PCDM obtains a $\mathcal{O}(\log(1/\rho))$ rate for a restarting strategy.

On the other hand, if the levelset is bounded, we have shown that PCDM achieves a rate of $\mathcal{O}(\log(1/\rho))$. It is still an open problem to determine whether PCDM can achieve a rate of $\mathcal{O}(\log(1/\rho))$ for a single run strategy when the levelset is unbounded. Note that Example 8 shows that we cannot achieve a rate of $\mathcal{O}(\log(1/\rho))$ using the approach in Lemma 7, but it is not a general counterexample. Thus, it may be possible to prove a $\mathcal{O}(\log(1/\rho))$ rate for PCDM with an unbounded levelset using different arguments from that made in Lemma 7, but currently we are unsure how to show this.

6. Discussion

6.1 *Comparison of the convergence rate results*

We have the following remarks on comparing the results in Theorem 3 with those in [24].

6.1.1 *Comparison in the convex case.*

For problem (1), an expected-value type of convergence rate is not presented explicitly in [23], although it can be derived from the following relation (that is stated in [24] and proved in [23, Theorem 1]):

$$\mathbb{E}[F(x_{k+1})-F^*\,|\,x_k]\leq (F(x_k)-F^*)-\alpha\frac{(F(x_k)-F^*)^2}{2c},\quad \forall k\geq 0, \tag{45}$$

where c is defined in (41). Taking expectation on both sides of (45) and using a similar argument as that in [17], gives

$$\mathbb{E}[F(x_k)-F^*\,|\,x_{k-1}]\leq \frac{2c(F(x_0)-F^*)}{2c+\alpha k(F(x_0)-F^*)},\quad \forall k\geq 0. \tag{46}$$

Let a and b denote the right-hand side of (16) and (46) respectively. By the definition of c and the relation $\|x_0-x_*\|_v\leq \mathcal{R}_{v,0}$, we see that when k is sufficiently large,

$$\frac{b}{a}\approx\frac{4c}{\|x_0-x_*\|_v^2+2(F(x_0)-F^*)}\geq\frac{4}{3}. \tag{47}$$

6.1.2 *Comparison in the strongly convex case.*

For the special case of (1) where at least one of f and Ψ is strongly convex (i.e. $\mu_f + \mu_\Psi > 0$), the authors of [24] showed that for all $k \geq 0$,

$$\mathbb{E}[F(x_k) - F^* \,|\, x_{k-1}] \leq \left(1 - \alpha \frac{\mu_f + \mu_\Psi}{1 + \mu_\Psi}\right)(F(x_0) - F^*). \tag{48}$$

It is not hard to show that

$$\frac{2(\mu_f + \mu_\Psi)}{1 + \mu_f + 2\mu_\Psi} > \frac{\mu_f + \mu_\Psi}{1 + \mu_\Psi}. \tag{49}$$

Recall that γ is defined in (32). Then it follows that for sufficiently large k one has

$$(1 - \alpha\gamma)^k \left(\frac{1 + \mu_\Psi}{2} R_0^2 + F(x_0) - F^*\right) \overset{(9)}{\leq} (1 - \alpha\gamma)^k \left(\frac{1 + \mu_f + \mu_\Psi}{\mu_f + \mu_\Psi}\right)(F(x_0) - F^*).$$

6.2 *Comparison of the iteration complexity results*

Here we compare the results in Theorem 13 with those in [24].

Comparison in the convex case For any $0 < \epsilon < F(x_0) - F_*$ and $\rho \in (0, 1)$, Richtárik and Takáč [24] showed that (44) holds for all $k \geq \tilde{K}$ where

$$\tilde{K} := \frac{2c}{\alpha\epsilon}\left(1 + \log\left(\frac{1}{\rho}\right)\right) + 2 - \frac{2c}{\alpha(F(x_0) - F_*)}. \tag{50}$$

Using the definition of c and the fact that $\|x_0 - x_*\|_v \leq \mathcal{R}_{v,0}$ we observe that

$$\tau := \frac{\|x_0 - x_*\|_v^2 + 2\xi_0}{4c} \leq \frac{3}{4}. \tag{51}$$

By the definitions of K and $\tilde{K}$ we have that for sufficiently small $\epsilon > 0$,

$$K - \tilde{K} \approx \frac{2c \log \tau}{\alpha\epsilon} \leq -\frac{2c \log(4/3)}{\alpha\epsilon}. \tag{52}$$

In addition, $\|x_0 - x_*\|_v$ can be much smaller than $\mathcal{R}_{v,0}$ and thus τ can be very small. It follows from the above that K can be significantly smaller than $\tilde{K}$.

Comparison in the strongly convex case In the strongly convex case (i.e. $\mu_f(w) + \mu_\Psi(w) > 0$), [24] showed that (44) holds for all $k \geq \hat{K}$ where

$$\hat{K} := \frac{1}{\alpha} \frac{1 + \mu_\Psi(w)}{\mu_f(w) + \mu_\Psi(w)} \log\left(\frac{F(x_0) - F_*}{\epsilon\rho}\right).$$

We can see that for ρ or ϵ sufficiently small we have

$$\frac{K}{\hat{K}} \leq \frac{1 + \mu_f(w) + \mu_\Psi(w)}{2(1 + \mu_\Psi(w))} \leq 1, \tag{53}$$

because $\mu_f \leq 1$, which demonstrates that K is smaller than $\hat{K}$.

Table 3. Approaches used in the numerical experiments.

Name	v	Note
BKBG	$v_{BKBG} = L$	This is naïve approach, which was proposed in [1] and [22]. Note that this is not ESO.
RT-P	$v_{RT-P} = \left(1 + \frac{(\omega-1)(\tau-1)}{\max\{1, n-1\}}\right) L$	Theorem A6, originally derived in [24].
RT-D	$v_{RT-D} = \left(1 + \frac{(\sigma-1)(\tau-1)}{\max\{1, n-1\}}\right) L$	Derived in [25] as a special case for $C = 1$.
FR	$v_{FR} = \hat{L}$	Theorem A8, proposed in [3] and *generalized in this paper* (Theorem A.7).
NC	$v_{NC} = \tilde{L}$	Theorem A9, proposed in [15].

7. Numerical experiments

In this Section we present preliminary computational results. The purpose of these experiments is to provide a numerical comparison of the performance of PCDM, under the different ESOs summarized in Appendix A.2.

Least squares. Consider the following convex optimization problem

$$\min_{x \in \mathbf{R}^N} \frac{1}{2}\|Ax - b\|_2^2, \tag{54}$$

where $A \in \mathbf{R}^{8\cdot 10^3 \times 2\cdot 10^3}$. Each row has between 1 and $\omega = 20$ non-zero elements (uniformly at random). For simplicity, we normalize (in ℓ_2 norm) all the columns of A. The value of $\sigma = \lambda_{\max}(A^{\mathrm{T}}A) = 10.48$. We have compared five different approaches which are given in Table 3. Parameter $\tau = 512$ and hence $1 + (\omega-1)(\tau-1)/\max\{1, n-1\} = 5.856$ for RT-P (Richtárik-Takáč-Parallel [24]) and $1 + (\sigma-1)(\tau-1)/\max\{1, n-1\} = 3.424$ for RT-D (Richtárik-Takáč-Distributed [25]) approach. The distribution of vectors v can be found in Figure 1 (right). Figure 1 shows the evolution of $F(x_k) - F^*$ for all five methods. Note that the BKBG [1,22] did not converge. The speed of RT-P, RT-D and FR [3] is quite similar and NC [15] is approximately 3 times worse because $v_{NC} \approx 3.22 v_{FR}$.

SVM dual. In this experiment we compare 4 methods from Table 3 (we have excluded the naïve approach because it usually diverges for large τ) on a real-world data set *astro-ph*, which consists of data from papers in physics [30]. This data set has 29,882 training samples and a total of 99,757 features. This data set is very sparse. Indeed, each sample uses on average only 77.317 features and each sample belongs to one of two classes. Hence, one might be interested in finding a hyperplane that separates the samples into their corresponding classes. The optimization problem can be formulated as follows:

$$\min_w P(w) := \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{N} \max\{0, 1 - y^{(i)} a_{(i)}^{\mathrm{T}} w\}, \tag{55}$$

where $y^{(i)} \in \{-1, 1\}$ is the label of the class to which sample $a_{(i)} \in \mathbf{R}^m$ belongs.

While problem formulation (55) does not fit our framework (the non-smooth part is non-separable) the dual formulation (see [5,29,31]) does:

$$\max_{x \in [0,1]^N} D(x) := \frac{1}{N}\mathbf{1}^{\mathrm{T}}x - \frac{1}{2\lambda N^2} x^{\mathrm{T}} Q x, \tag{56}$$

where $Q \in \mathbf{R}^{N\times N}$, $Q_{i,j} = y^{(i)}y^{(j)}\langle a_{(i)}, a_{(j)}\rangle$. In particular, problem formulation (56) is the sum of a smooth term, and the restriction $x \in [0,1]^N$ can be formulated as a (block separable) indicator function. In this data set, each sample is normalized, hence $L = (1,\ldots,1)^{\mathrm{T}}$.

For any dual feasible point x we can obtain a primal feasible point $w(x) = (1/\lambda n)\sum_{i=1}^{N} x^{(i)}y^{(i)}a_{(i)}$. Moreover, from strong duality we know that if x^* is an optimal solution of (56), then $w^* = w(x^*)$ is optimal for problem (55). Therefore, we can associate a gap $G(x) = P(w(x)) - D(x)$ to each feasible point x, which measures the distance of the objective value from optimality. Clearly $G(x^*) = 0$.

Figure 2 (left) shows the evolution of $G(x_k)$ as the iterates progress, and the distribution of an ESO parameter v for different choice of $\tau \in \{32, 256\}$. Naturally, as τ increases, the distribution of $v, \hat{v}$ shifts to the right, whereas the distribution of $\tilde{v}$ is not influenced by changing τ. The value of important parameters for other methods are $\sigma = 287.273$ and $\omega = 29881$. For $\tau = 32$ we have $1 + (\omega - 1)(\tau - 1)/\max\{1, n-1\} = 31.998$ for RT-P and $1 + (\sigma - 1)(\tau - 1)/\max\{1, n-1\} = 1.296$ for RT-D, and for $\tau = 256$ we have $1 + (\omega - 1)(\tau - 1)/\max\{1, n-1\} = 255.991$ for RT-P and $1 + (\sigma - 1)(\tau - 1)/\max\{1, n-1\} = 3.443$ for RT-D. Again the best performance is given by RT-D which requires knowledge of σ. If we do not want to estimate the parameter σ then we should use FR. If was shown in [3] that for a quadratic objective function FR is always better than RT-P.

Underdetermined Least Squares. Here we perform a numerical experiment on an underdetermined least squares problem; such problems have the property that the *initial levelset is unbounded*. The work in Section 5.1 establishes iteration complexity results for PCDM applied to problems with an unbounded levelset, which motivates us to investigate them here. In particular, we consider a problem of the form (54), where the sparse matrix A has $m = 2^{12}$ rows and $n = 2^{13}$ columns, (so $m < n$), and there are approximately 20 non-zeros in each column. We also suppose that we are in possession of a vector x_t that is a solution of problem (54). That is, we form a vector $x_t \in \mathbf{R}^n$ and compute a (noiseless) data vector $b = Ax_t \in \mathbf{R}^m$. This ensures that the system of equations is consistent, so that $F^* = 0$ here, and we note that there are infinitely many solutions to this (convex) problem.

We investigate PCDM with a τ-nice sampling, so that the norm weighting vector v is as described in Row 2 of Table 3 (see also Theorem A6) with $L = (\|a_1\|_2^2, \ldots, \|a_n\|_2^2)^{\mathrm{T}}$, where $a_1, \ldots, a_n$ are the columns of A. We simulated a parallel set-up where we vary the number of

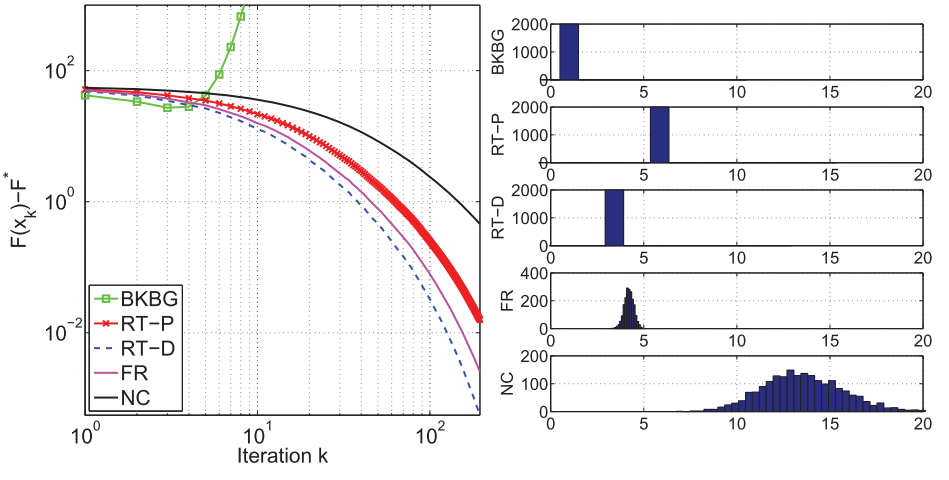

Figure 1. Evolution of $F(x_k) - F^*$ for 5 different methods (left) and distribution of v (right).

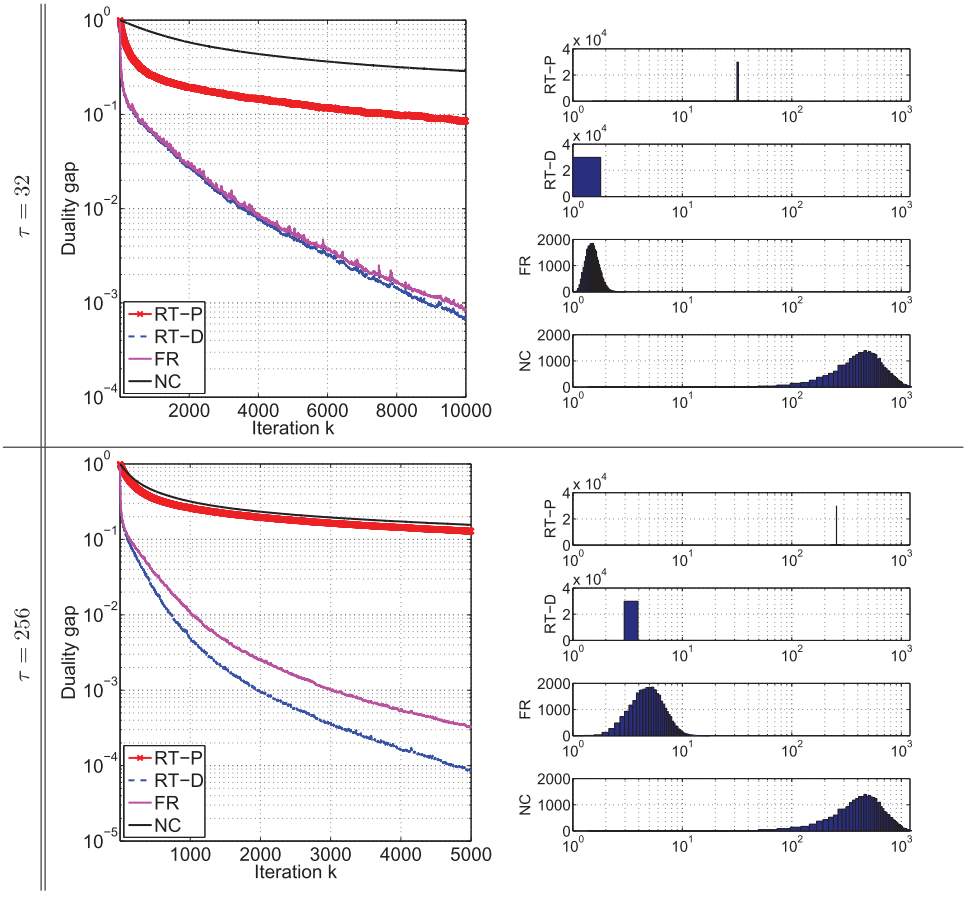

Figure 2. Comparison of evolution of $G(x_k)$ for various methods and the distribution of v.

processors available from $\tau = 512$ to $\tau = 1$. For a fixed number of τ processors, we form 100 problem instances (i.e. we form 100 matrices A and data vectors b) and we report the average results in Table 4. In each case we use a randomly generated starting point x_0.

The columns of Table 4 are as follows. In the first column we state the number of processors, and in the second column we report the average degree of separability ω over the 100 runs (see Definition A.1). In columns 3 and 4 we report the average values of $F(x_0)$ (recall that $F^* = 0$, so

Table 4. Average results for the experiment investigating problems where the initial levelset is unbounded.

τ	Average ω	Average F_0	Average $\|\|x_0 - x^*\|\|_v^2$	Average k	Theoretical K
512	38.12	4.84×10^2	1.60×10^3	1.66×10^3	3.34×10^8
256	37.84	4.97×10^2	1.06×10^3	2.23×10^3	4.99×10^8
128	38.09	4.87×10^2	7.65×10^2	3.34×10^3	8.02×10^8
64	38.07	5.09×10^2	6.51×10^2	5.87×10^3	1.48×10^9
32	37.75	5.02×10^2	5.72×10^2	9.83×10^3	2.75×10^9
16	38.13	4.94×10^2	5.28×10^2	1.94×10^4	5.24×10^9
8	37.97	4.96×10^2	5.10×10^2	3.73×10^4	1.03×10^{10}
4	38.09	5.03×10^2	5.09×10^2	7.10×10^4	2.01×10^{10}
2	37.89	4.98×10^2	4.99×10^2	1.44×10^5	4.09×10^{10}
1	37.96	5.01×10^2	5.02×10^2	4.85×10^5	8.22×10^{10}

$\xi_0 \equiv F(x_0) - F^* = F(x_0)$), and $\|x_0 - x^*\|_v^2$, respectively. Column 5 gives the average number of iterations k (over 100 runs) required by PCDM to converge to an $\epsilon = 10^{-4}$ accuracy solution. In the final column we give the number of iterations K predicted by the theoretical bound (39). In each run, a theoretical bound was computed using the values F_0 and $\|x_0 - x^*\|_v^2$ and the average theoretical bound over all 100 runs is reported. For simplicity we computed the bound using $\rho = 1$.

Table 4 shows that, while we now have a theoretical bound for the number of iterations needed by PCDM to obtain an ϵ-optimal solution (with probability exceeding $1 - \rho$), for this problem set-up, the bound was pessimistic, and the actual number of iterations required was much smaller than predicted.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work of this author was supported by the NSF Grants CCF-1618717 and CMMI-1663256. The work of all authors was supported by the EPSRC grant EP/I017127/1 (Mathematics for Vast Digital Resources) and by the Centre for Numerical Algorithms and Intelligent Software (funded by EPSRC grant EP/G036136/1 and the Scottish Funding Council).

Notes

1. A preprint of this work was ready in 2013, and was sent to the authors of [19] via a private communication. Their expectation result without levelset information was heavily influenced by the result in this work.
2. A preliminary version of this paper was ready in August 2013.

References

[1] J.K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, *Parallel coordinate descent for L1-regularized loss minimization*, 28th International Conference on Machine Learning, Bellevue, WA, 2011.
[2] O. Fercoq and P. Richtárik, *Smooth minimization of nonsmooth functions by parallel coordinate descent*, preprint (2013). Available at arXiv:1309.5885.
[3] O. Fercoq and P. Richtárik, *Accelerated, parallel and proximal coordinate descent*, SIAMJ. Optim. 25(4) (2015), pp. 1997–2023.
[4] O. Fercoq, Z. Qu, P. Richtárik, and M. Takáč, *Fast distributed coordinate descent for non-strongly convex losses*, IEEE Workshop on Machine Learning for Signal Processing, Reims, France, 2014.
[5] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. Sathiya Keerthi, and S. Sundararajan, *A dual coordinate descent method for large-scale linear svm*, ICML 2008, Helsinki, Finland, 2008, pp. 408–415.
[6] M. Jaggi, V. Smith, M. Takác, J. Terhorst, S. Krishnan, T. Hofmann, and M.I. Jordan, *Communication-efficient distributed dual coordinate ascent*, Advances in Neural Information Processing Systems, 2014, pp. 3068–3076. Available at https://papers.nips.cc/book/advances-in-neural-information-processing-systems-27-2014.
[7] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Advances in Neural Information Processing Systems, 2013, pp. 315–323. Available at https://papers.nips.cc/book/advances-in-neural-information-processing-systems-26-2013.
[8] J. Konečný and P. Richtárik, *Semi-stochastic gradient descent methods*, Frontiers Appl. Math. Stat. (2017). doi:10.3389/fams.2017.00009.
[9] Y. Li and S. Osher, *Coordinate descent optimization for l_1 minimization with application to compressed sensing; a greedy algorithm*, Inverse Probl. Imag. 3(3) (2009), pp. 487–503.
[10] Q. Lin, Z. Lu, and L. Xiao, *An accelerated proximal coordinate gradient method*, Advances in Neural Information Processing Systems, 2014, pp. 3059–3067. Available at https://papers.nips.cc/book/advances-in-neural-information-processing-systems-27-2014.
[11] J. Liu, S.J. Wright, C. Ré, and V. Bittorf, *An asynchronous parallel stochastic coordinate descent algorithm*, in *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32, E.P. Xing and T. Jebara, eds., JMLR W &CP, Beijing, China, 2014.

[12] C. Ma, V. Smith, M. Jaggi, M.I. Jordan, P. Richtárik, and M. Takáč, *Adding vs. averaging in distributed primal-dual optimization*, in *Proceedings of the 32nd International Conference on Machine Learning (accepted)*, Vol. 33, E.P. Xing and T. Jebara, eds., JMLR W &CP, Lille, Frace, 2015.

[13] M. Mahdavi, L. Zhang, and R. Jin, *Mixed optimization for smooth functions*, Advances in Neural Information Processing Systems, 2013, pp. 674–682. Available at https://papers.nips.cc/book/advances-in-neural-information-processing-systems-26-2013.

[14] J. Mareček, P. Richtárik, and M. Takáč, *Distributed block coordinate descent for minimizing partially separable functions*. in *Numerical Analysis and Optimization*, M. Al-Baali, L. Grandinetti, and A. Purnama, eds., *Springer Proceedings in Mathematics & Statistics*, Vol. 134, Muscat, Oman, 2015, pp. 261–288.

[15] I. Necoara and D. Clipici, *Parallel random coordinate descent method for composite minimization: Convergence analysis and error bounds*, SIAMJ. Optim. 26(1) (2016), pp. 197–226.

[16] I. Necoara, Y. Nesterov, and F. Glineur, *Efficiency of randomized coordinate descent methods on optimization problems with linearly coupled constraints*, Technical report, 2012.

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

[18] Yu. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program. Ser. B 140 (2013), pp. 125–161.

[19] Z. Qu and P. Richtárik, *Coordinate descent with arbitrary sampling I: Algorithms and complexity*, Optim. Methods Softw. 31(5) (2016), pp. 829–857.

[20] Z. Qu and P. Richtárik, *Coordinate descent with arbitrary sampling II: Expected separable overapproximation*, Optim. Methods Softw. 31(5) (2016), pp. 858–884.

[21] Z. Qu, P. Richtarik, and T. Zhang, *Quartz: Randomized dual coordinate ascent with arbitrary sampling*, in *Advances in Neural Information Processing Systems 28*, C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, eds., Curran Associates, Inc., 2015, pp. 865–873. Available at https://papers.nips.cc/book/advances-in-neural-information-processing-systems-28-2015.

[22] P. Richtárik and M. Takáč, *Efficient serial and parallel coordinate descent methods for huge-scale truss topology design*, Operations Research Proceedings 2011, Springer, Berlin, Heidelberg, 2012, pp. 27–32.

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

[24] P. Richtárik and M. Takáč, *Parallel coordinate descent methods for big data optimization*, Math. Program. Ser. A (2015), pp. 1–52.

[25] P. Richtárik and M. Takác, *Distributed coordinate descent method for learning with big data*, J. Mach. Learn. Res. 17 (2016), pp. 1–25.

[26] P. Richtárik and M. Takáč, *On optimal probabilities in stochastic coordinate descent methods*, Optim. Lett. 10(6) (2016), pp. 1233–1243.

[27] A. Saha and A. Tewari, *On the nonasymptotic convergence of cyclic coordinate descent methods*, SIAMJ. Optim. 23(1) (2013), pp. 576–601.

[28] J. Nutini, M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke, *Coordinate descent converges faster with the Gauss–Southwell rule than random selection*, ICML 2015, Lile, France, 2015.

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

[30] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, *Pegasos: Primal estimated sub-gradient SOlver for SVM*, Math. Program 127 (2011), pp. 3–30.

[31] M. Takáč, A. Bijral, P. Richtárik, and N. Srebro, *Mini-batch primal and dual methods for SVMs*, 30th International Conference on Machine Learning, Atlanta, USA, 2013.

[32] Q. Tao, K. Kong, D. Chu, and G. Wu, *Stochastic coordinate descent methods for regularized smooth and nonsmooth losses*, in *Machine Learning and Knowledge Discovery in Databases*, P.A. Flach, T. De Bie, and N. Cristianini, eds, *Lecture Notes in Computer Science*, Vol. 7523, Springer, Berlin, Heidelberg, 2012, pp. 537–552.

[33] R. Tappenden, P. Richtárik, and J. Gondzio, *Inexact coordinate descent: Complexity and preconditioning*, J. Optim. Theory Appl. 170(1) (2016), pp. 144–176.

[34] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, J. Optim. Theory Appl. 109 (2001), pp. 475–494.

[35] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program. Ser. B 117 (2009), pp. 387–423.

[36] Wright S.J., *Accelerated block-coordinate relaxation for regularized optimization*, SIAMJ. Optim. 22(1) (2012), pp. 159–186.

[37] S.J. Wright, *Coordinate descent algorithms*, Math. Program. 151(1) (2015), pp. 3–34.

[38] T.T. Wu and K. Lange, *Coordinate descent algorithms for lasso penalized regression*, Ann. Appl. Stat. 2(1) (2008), pp. 224–244.

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

A. Appendix. Expected separable overapproximation

A.1 *Smoothness assumptions*

In this work we assume that the function f is partially separable and smooth, and the purpose of this section is to define these two concepts. We begin with the definition of partial separability for a smooth convex function, introduced by Richtárik and Takáč [24].

DEFINITION A.1 (Partial separability [24]) *A smooth convex function $f : \mathbf{R}^N \to \mathbf{R}$ is partially separable of degree ω if there exists a collection $\mathcal{J}$ of subsets of $\{1, 2, \ldots, n\}$ such that*

$$f(x) = \sum_{J \in \mathcal{J}} f_J(x) \quad \text{and} \quad \max_{J \in \mathcal{J}} |J| \leq \omega, \tag{A1}$$

where for each J, f_J is a smooth convex function that depends on $x^{(i)}$ for $i \in J$ only.

Now we introduce different types of smoothness assumptions for the function f. Each smoothness type gives rise to a different ESO. Note that *all* of the following smoothness assumptions are *equivalent*. That is, if a given function satisfies one of the assumptions, then there exist constants such that the other assumptions also hold.

The first type of assumption is a classical assumption in the literature [22–24].

ASSUMPTION A.2 ((Block) Coordinate-wise Lipschitz continuous gradient) *The gradient of f is block Lipschitz, uniformly in x, with positive constants $L_1, \ldots, L_n$. That is, for all $x \in \mathbf{R}^N$, $i = 1, \ldots, n$ and $h \in \mathbf{R}^{N_i}$ we have*

$$\|(\nabla f(x + U_i h))^{(i)} - (\nabla f(x))^{(i)}\|^*_{(i)} \leq L_i \|h\|_{(i)}, \tag{A2}$$

where $\nabla f(x)$ denotes the gradient of f and

$$(\nabla f(x))^{(i)} = U_i^{\mathrm{T}} \nabla f(x) \in \mathbf{R}^{N_i}. \tag{A3}$$

The second type of assumption we make is that each function in the sum (A1) has a Lipschitz continuous gradient. Such an assumption is made, for example, in [7,8,13,15]. Moreover, we allow each function to have Lipschitz continuous gradient with a *different constant* (which was also assumed in [15]).

ASSUMPTION A.3 (Lipschitz continuous gradient of sub-functions) *The gradient of f_J, $J \in \mathcal{J}$ has a Lipschitz continuous gradient, uniformly in x, with positive constant $\tilde{L}_J$ with respect to some Euclidean norm $\|\cdot\|_{(\tilde{J})}$. That is, for all $x \in \mathbf{R}^N$, $J \in \mathcal{J}$ and $h \in \mathbf{R}^N$ we have*

$$\|\nabla f_J(x + h) - \nabla f_J(x)\|^*_{(\tilde{J})} \leq \tilde{L}_J \|h\|_{(\tilde{J})}. \tag{A4}$$

Note that this smoothness assumption is more general than that made in [15] because of the possibility of choosing general norms of the form $\|\cdot\|_{(\tilde{J})}$. Further, Assumption A.3 generalizes the smoothness assumptions imposed in [1,25].

The third type of assumption we make is that each function in the sum (A1) has coordinate-wise Lipschitz continuous gradient.

ASSUMPTION A.4 ((Block) Coordinate-wise Lipschitz continuous gradient of sub-functions) *The gradient of f_J, $J \in \mathcal{J}$ is block Lipschitz, uniformly in x, with non-negative constants $\hat{L}_{J,1}, \ldots, \hat{L}_{J,n}$. That is, for all $x \in \mathbf{R}^N$, $i = 1, \ldots, n$, $J \in \mathcal{J}$ and $h \in \mathbf{R}^{N_i}$ we have*

$$\|(\nabla f_J(x + U_i h))^{(i)} - (\nabla f_J(x))^{(i)}\|^*_{(i)} \leq \hat{L}_{J,i} \|h\|_{(i)}. \tag{A5}$$

One can think of Assumptions A.2 and A.3 as being 'opposite' to each other in the following sense. If we associate the block coordinates with the columns, and the functions with the rows, we see that Assumption A.2 captures the dependence columns-wise, while Assumption A.3 captures the dependence row-wise. Hence, Assumption A.4 can be thought of as an element-wise smoothness assumption.

To make this more concrete, we present an example that demonstrates how to compute the Lipschitz constants for a quadratic function, under each of the three smoothness assumptions stated above.

Example A.5 Let the function $f(x) = \frac{1}{2}\|Ax - b\|_2^2 = \frac{1}{2}\sum_{j=1}^m (b^{(i)} - \sum_{i=1}^n a_{j,i} x^{(i)})^2$, where $A \in \mathbf{R}^{m \times n}$ and $a_{j,i}$ is (j, i)th element of the matrix A. Let us fix all the norms $\|\cdot\|_{(\tilde{J})}$ from Assumption A.3 to be standard Euclidean norms. Then one can easily verify that Equations (A2), (A4) and (A5) are satisfied with the following choice of constants

$$L_i = \sum_{j=1}^m a_{j,i}^2, \quad \tilde{L}_j = \sum_{i=1}^n a_{j,i}^2, \quad \hat{L}_{j,i} = a_{j,i}^2.$$

In words, L_i is equal to square of the ℓ_2 norm of i th column, $\tilde{L}_j$ is equal to the square of the ℓ_2 norm of the j th row and $\hat{L}_{j,i}$ is simply the square of the (j, i) th element of the matrix A.

One could be misled into believing that Assumption A.4 is the best because it is the most restrictive. However, while this is true for the quadratic objective shown in Example A.5, for a general convex function, Assumption A.4 can give Lipschitz constants that lead to worse ESO bounds (see Example A.10 for further details).

A.2 *Expected separable overapproximation*

Now, it is clear that the update h in Algorithm 1 depends on the ESO parameter v. This shows that the ESO is not just a technical tool; the parameters are actually *used* in Algorithm 1. Therefore we must be able to obtain/compute these parameters easily. We now present the following three theorems, namely Theorems A6, A8 and A9, that explain how to obtain the v parameter for a τ-nice sampling, under different smoothness assumptions.

THEOREM A.6 (ESO for a τ-nice sampling, Theorem 14 in [24]) *Let Assumption* A.2 *hold with constants $L_1, \dots, L_n$ and let $\hat{S}$ be a τ-nice sampling. Then $f : \mathbf{R}^N \to \mathbf{R}$ admits an ESO with respect to the sampling $\hat{S}$ with parameter*

$$v = \left(1 + \frac{(\omega - 1)(\tau - 1)}{\max\{1, n - 1\}}\right) L,$$

where $L = (L_1, \dots, L_n)^{\mathrm{T}}$.

The obvious disadvantage of Theorem A.6 is the fact that v in the ESO, depends on ω. (When ω is large, so too is v.) One can imagine a situation in which ω is much larger than the average cardinality of $J \in \mathcal{J}$, resulting in a large v. For example, if $|J|$ for $J \in \mathcal{J}$ is small for all but one function.

With this in mind, we introduce a new theorem that shows how the ESO in Theorem A.6 can be modified if we know that Assumption A.4 holds. In this case, the role of ω is slightly suppressed.

THEOREM A.7 (ESO for a doubly uniform sampling) *Let Assumption* A.4 *hold with constants $\hat{L}_{J,i}$, $J \in \mathcal{J}$, $i \in \{1, \dots, n\}$ and let $\hat{S}$ be a doubly uniform sampling. Then $f : \mathbf{R}^N \to \mathbf{R}$ admits an ESO with respect to the sampling $\hat{S}$ with parameter*

$$\bar{v} = \sum_{J \in \mathcal{J}} \left(1 + \frac{\left(\frac{\mathbf{E}[|\hat{S}|^2]}{\mathbf{E}[|\hat{S}|]} - 1\right)(|J| - 1)}{\max\{1, n - 1\}}\right) (\hat{L}_{J,1}, \dots, \hat{L}_{J,n})^{\mathrm{T}}. \tag{A6}$$

Proof From Theorem 15 in [24] we know that for each function $f_J, J \in \mathcal{J}$ we have

$$(f_J, \hat{S}) \sim \mathrm{ESO}\left(1 + \frac{(\frac{\mathbf{E}[|\hat{S}|^2]}{\mathbf{E}[|\hat{S}|]} - 1)(|J| - 1)}{\max\{1, n - 1\}}, (\hat{L}_{J,1}, \dots, \hat{L}_{J,n})^{\mathrm{T}}\right).$$

Now, using Theorem 10 in [24], which deals with conic combinations of functions, we have

$$\left(\sum_{J \in \mathcal{J}} f_J, \hat{S}\right) \sim \mathrm{ESO}\left(1, \sum_{J \in \mathcal{J}} \left(1 + \frac{(\frac{\mathbf{E}[|\hat{S}|^2]}{\mathbf{E}[|\hat{S}|]} - 1)(|J| - 1)}{\max\{1, n - 1\}}\right) (\hat{L}_{J,1}, \dots, \hat{L}_{J,n})^{\mathrm{T}}\right).$$

■

The following Theorem is a special case of Theorem A.7 for a τ-nice sampling.

THEOREM A.8 (ESO for a τ-nice sampling, Theorem 1 in [3]) *Let Assumption* A.4 *hold with constants $\hat{L}_{J,i}$, $J \in \mathcal{J}$, $i \in \{1, \dots, n\}$ and let $\hat{S}$ be a τ-nice sampling. Then $f : \mathbf{R}^N \to \mathbf{R}$ admits an ESO with respect to the sampling $\hat{S}$ with parameter*

$$\hat{v} = \sum_{J \in \mathcal{J}} \left(1 + \frac{(\tau - 1)(|J| - 1)}{\max\{1, n - 1\}}\right) (\hat{L}_{J,1}, \dots, \hat{L}_{J,n})^{\mathrm{T}}. \tag{A7}$$

Proof Notice that, if $\hat{S}$ is τ-nice sampling, then $\mathbf{E}[|\hat{S}|] = \tau$ and $\mathbf{E}[|\hat{S}|^2] = \tau$ and the result follows from Theorem A.7. ■

The following theorem explains how to compute an ESO if Assumption A.3 holds. This ESO was proposed and proved in [15].

THEOREM A.9 (ESO for τ-nice sampling, Lemma 1 in [15]) *Let Assumption* A.3 *hold with constants* $\tilde{L}_J$, $J \in \mathcal{J}$ *and let* $\hat{S}$ *be a* τ*-nice sampling. Then* $f : \mathbf{R}^N \to \mathbf{R}$ *admits an ESO with respect to the sampling* $\hat{S}$ *with parameter*

$$\tilde{v} = \sum_{J \in \mathcal{J}} \tilde{L}_J e_{[J]},$$

where $e = (1, \ldots, 1)^{\mathrm{T}} \in \mathbf{R}^n$. *Moreover, this ESO is monotonic.*

As it is shown in Theorem 3 the speed of the algorithm (number of iterations needed to solve the problem) depends on ESO parameter v via the term $\|x_0 - x_*\|_v^2$. Moreover, for a given objective function and sampling $\hat{S}$, there may be more than one ESO that could be chosen. Suppose that we have two ESOs to choose from, characterized by two parameters v_a and v_b respectively, and let $v_a < v_b$. In this case, the ESO characterized by v_a will give us (theoretically) faster convergence, and so it is obvious that this ESO should be used. Furthermore, v_a is used as a parameter in Algorithm 1, and so, intuitively, this faster theoretical convergence, is expected to lead to fast practical performance.

In Section 6.1 in [3] it was shown that for a quadratic objective, the ESO in Theorem A.6 is always worse than the ESO from Theorem A.8. However, for a general objective the opposite can be true. The following simple example shows that the ESO from Theorems A.8 and A.9 can be m times worse than the ESO from Theorem A.6.

Example A.10 Consider the function

$$f(x) = \sum_{j=1}^{m} \underbrace{\log(1 + \mathrm{e}^{-x+\zeta j})}_{f_j(x)},$$

where ζ is large. It is clear that L_j is

$$L_j = \max_x (f_j(x))'' = \max_x \frac{\mathrm{e}^{x+\zeta j}}{(\mathrm{e}^{\zeta} + \mathrm{e}^{x})^2} = \frac{1}{4}.$$

Therefore, from Theorem A.9, we obtain $\tilde{v} = \tilde{L} = m/4$.

On the other hand, Theorem A.6 produces an ESO with

$$v = \max_x (f(x))'' \approx \frac{1}{4},$$

provided that ζ is large, for example, $\zeta = 100$. Hence, in this case, the ESO from Theorem A.6 will lead to an algorithm that is approximately m times faster than if the ESOs from Theorems A8 or A9 were used.

Remark A thorough discussion of the ESO is presented in [24, Section 4]. Moreover, [24, Section 5.5] presents a list of parameters v associated with a particular f and sampling scheme $\hat{S}$ that give rise to an ESO. Indeed, each of samplings described in Section 3.1 in this work gives rise to a v for which an ESO exists.