



# THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

### Coordinate Descent with Arbitrary Sampling I: Algorithms and Complexity

**Citation for published version:**

Qu, Z & Richtarik, P 2016, 'Coordinate Descent with Arbitrary Sampling I: Algorithms and Complexity' Optimization Methods and Software, vol. 31, no. 5, pp. 829-857. DOI: 10.1080/10556788.2016.1190360

**Digital Object Identifier (DOI):**

[10.1080/10556788.2016.1190360](https://doi.org/10.1080/10556788.2016.1190360)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Peer reviewed version

**Published In:**

Optimization Methods and Software

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# Coordinate Descent with Arbitrary Sampling I: Algorithms and Complexity\*

Zheng Qu <sup>†</sup>Peter Richtárik <sup>‡</sup>

June 16, 2015

## Abstract

We study the problem of minimizing the sum of a smooth convex function and a convex block-separable regularizer and propose a new randomized coordinate descent method, which we call ALPHA. Our method at every iteration updates a *random subset* of coordinates, following an *arbitrary distribution*. No coordinate descent methods capable to handle an arbitrary sampling have been studied in the literature before for this problem. ALPHA is a remarkably flexible algorithm: in special cases, it reduces to deterministic and randomized methods such as gradient descent, coordinate descent, parallel coordinate descent and distributed coordinate descent – both in nonaccelerated and accelerated variants. The variants with arbitrary (or importance) sampling are new. We provide a complexity analysis of ALPHA, from which we deduce as a direct corollary complexity bounds for its many variants, all matching or improving best known bounds.

## 1 Introduction

With the dawn of the big data age, there has been a growing interest in solving optimization problems of unprecedented sizes. It was soon realized that traditional approaches, which work extremely well for problems of moderate sizes and when solutions of high accuracy are required, are not efficient for modern problems of large enough size and for applications where only rough or moderate accuracy solutions are sufficient. The focus of the optimization, numerical analysis and machine learning communities, and of practitioners in the sciences and industry, shifted to first-order (gradient) algorithms [23].

However, once the size of problems becomes truly big, it is necessary to turn to methods which are able to output a reasonably good solutions after an amount of work roughly equivalent to reading the data describing the problem a few times. For this to be possible, methods need to be able to progress while reading only a small part of the data describing the problem, which often means that a single iteration needs to be based on less information than that contained in the gradient of the objective (loss) function. The most popular methods of this type are stochastic gradient methods

---

\*The authors acknowledge support from the EPSRC Grant EP/K02325X/1, *Accelerated Coordinate Descent Methods for Big Data Optimization*. Most of the material of this paper was obtained by the authors in Spring 2014, and was presented by PR in June 2014 at the “Khronos Days Summer School” focused on “High-Dimensional Learning and Optimization” in Grenoble, France [27]; <http://www.maths.ed.ac.uk/%7Eprichter/docs/cdm-talk.pdf>.

<sup>†</sup>School of Mathematics, The University of Edinburgh, United Kingdom (e-mail: zheng.qu@ed.ac.uk)

<sup>‡</sup>School of Mathematics, The University of Edinburgh, United Kingdom (e-mail: peter.richtarik@ed.ac.uk)

[44, 22, 34, 38, 46], randomized coordinate descent methods [4, 8, 31, 32, 38, 40, 39, 7, 28, 5, 46, 36, 37, 14, 26, 11] and semi-stochastic gradient descent methods [33, 43, 9, 12, 18, 19, 3, 42, 10, 11].

## 1.1 Randomized coordinate descent

In this paper we focus on randomized coordinate descent methods. After the seminal work of Nesterov [24], which provided an early theoretical justification of these methods for unconstrained convex minimization, the study has been successively extended to  $L1$ -regularized [35, 30], proximal [31, 17], parallel [32, 6], distributed [28, 5, 26] and primal-dual [38, 26] variants of coordinate descent. Accelerated coordinate descent—characterized by its  $O(1/k^2)$  complexity for non-strongly convex problems—was studied in [24, 17]. However, these methods are of theoretical nature only due to the fact that they rely on the need to perform full-dimensional vector operation at every iteration, which destroys the main advantage of coordinate descent – its ability to reduce the problem into subproblems of smaller sizes. A theoretically and practically efficient accelerated coordinate descent methods were proposed recently by Lee and Sidford [13] and Fercoq and Richtárik [6], the latter work (APPROX algorithm) combining acceleration with parallelism and proximal setup. An accelerated distributed coordinate descent algorithm [5] is obtained by specializing APPROX to a distributed sampling. All above mentioned papers only consider unconstrained or separably constrained problems. Some progress on linearly-coupled constraints has been made by Necoara et al in [20, 21]. Asynchronous variants of parallel coordinate descent methods were developed by Liu, Wright et al [16, 15].

Virtually all existing work in stochastic optimization deals with a *uniform sampling*. In the context of coordinate descent, this means that the random subset (sampling) of coordinates chosen and updated at every iteration has the property that each coordinate is chosen *equally likely*. The possibility to assign different selection probabilities to different coordinates—also known as *importance sampling*—was considered in [24, 31] and recently in [46, 45]. However, these works consider the serial case only: a single coordinate is updated in each iteration. Randomized coordinate descent methods updating a *subset of coordinates following an arbitrary distribution* (i.e., using an arbitrary sampling) were first investigated by Richtárik and Takáč [29] (NSync method) for strongly convex and smooth objective functions, and subsequently by Qu, Richtárik and Zhang [26] (QUARTZ method), for strongly convex and possibly nonsmooth functions, and in a primal-dual framework.

In this paper we give the *first fully unified analysis* of gradient type algorithms which contain randomized coordinate descent on one end of the spectrum and gradient (or accelerated gradient) descent on the other hand. All our complexity results match or improve on the state of the art in all cases where specialized algorithms for specific samplings already exist. Moreover, we managed to substantially simplify the analysis for the sake of making the material accessible to a wide community.

## 1.2 Problem Formulation

In this paper we consider the composite optimization problem

$$\begin{aligned} & \text{minimize} && F(x) \stackrel{\text{def}}{=} f(x) + \psi(x) \\ & \text{subject to} && x = (x^1, \dots, x^n) \in \mathbb{R}^{N_1} \times \dots \times \mathbb{R}^{N_n} = \mathbb{R}^N, \end{aligned} \tag{1}$$

where  $f : \mathbb{R}^N \rightarrow \mathbb{R}$  is convex and differentiable,  $\psi : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$  is block-separable:

$$\psi(x) = \sum_{i=1}^n \psi^i(x^i),$$

and each  $\psi^i : \mathbb{R}^{N_i} \rightarrow \mathbb{R} \cup \{+\infty\}$  is convex and closed.

### 1.3 Contributions

We now summarize the main contributions of this work.

**New algorithm.** We propose ALPHA (Algorithm 1) – a randomized gradient-type method for solving the convex composite optimization problem (1). In each iteration, ALPHA picks and updates a random subset of the blocks  $\{1, 2, \dots, n\}$ , using an *arbitrary sampling*. That is, we allow for the distribution of the random set-valued mapping to be arbitrary (and as explained further below, analyze the iteration complexity of the method).

To the best of our knowledge, there are only two methods in the literature with the “arbitrary sampling” property, the NSync method of Richtárik and Takáč [29] (focusing on the simple problem of minimizing a smooth strongly convex function) and the QUARTZ method of Qu, Richtárik and Zhang [26] (a primal-dual method; considering strongly convex but possibly nonsmooth functions appearing in machine learning). Hence, our work is complementary to this development.

**Complexity analysis.** We study the iteration complexity of ALPHA. That is, for an arbitrary (but “proper”) sampling, we provide bounds on the number of iterations needed to approximately solve the problem, in expectation. Our general bounds are formulated in Section 6: Theorem 6.1 covers the non-accelerated variant with  $O(1/k)$  rate and Theorem 6.2 covers the accelerated variant with  $O(1/k^2)$  rate, where  $k$  is the iteration counter. To the best of our knowledge, these are the first complexity results for a randomized coordinate descent methods utilizing an *arbitrary sampling* for problem (1).

**Expected separable overapproximation.** Besides the dependence of the complexity bound on the iteration counter  $k$ , it is important to study its dependence on the sampling  $\hat{S}$  and the objective function. Our results make this dependence explicit: they hold under the assumption that  $f$  admits an expected separable overapproximation (ESO) with respect to the sampling  $\hat{S}$  (Assumption 2.1). This is an inequality involving  $f$  and  $\hat{S}$  which determines certain important parameters  $v_1, \dots, v_n$  which are needed to run the method (they determine the stepsizes) and which also appear in the complexity bounds. In some cases it is possible to *design* a sampling which optimizes the complexity bound.

In the case of a *serial sampling*, which is by far the most common type of sampling studied in conjunction with randomized coordinate descent methods ( $\hat{S}$  is serial if  $|\hat{S}| = 1$  with probability 1), the parameter  $v_i$  can simply be set to the Lipschitz constant of the block-derivative of  $f$  corresponding to block  $i$ . In particular, if  $n = 1$ , then  $v_1$  is the Lipschitz constant of the gradient of  $f$  [24, 31]. The situation is more complicated in the case of a *parallel sampling* ( $\hat{S}$  is parallel if it is not serial; that is, if we allow for multiple blocks to be updated at every iteration) – and this why there is a need for the ESO inequality. Intuitively speaking, the parameters  $v_1, \dots, v_n$  capture

certain smoothness properties of the gradient of  $f$  in a random subspace spanned by the blocks selected by the sampling  $\hat{S}$ .

The ESO concept is of key importance in the design and analysis of randomized coordinate descent methods [32, 7, 28, 29, 6, 5, 26, 25]. We provide a *systematic study of ESO inequalities* in a companion paper [25].

**Simple complexity analysis in the smooth case.** In order to make the exposition more accessible, we first focus on ALPHA applied to problem (1) with  $\psi \equiv 0$  (we call this the “smooth case”). In this simpler setting, it is possible to provide a *simplified complexity analysis* – we do this in Section 3; see Theorem 3.1 (non-accelerated variant with  $O(1/k)$  rate) and Theorem 3.2 (accelerated variant with  $O(1/k^2)$  rate). For convenience, ALPHA specialized to the smooth case is formulated as Algorithm 2. Our analysis in this case is different from the one we give in Section 6, where we analyze the method in the general proximal setup.

**Flexibility.** ALPHA is a remarkably flexible<sup>1</sup> algorithm, encoding a number of classical, recent and new algorithms in special cases, depending on the choice of the parameters of the method: sampling  $\hat{S}$  and “stepsize sequence”  $\{\theta_k\}$ . We devote Section 4 to highlighting several of the many algorithms ALPHA reduces to in special cases, focusing on the smooth case for simplicity (special cases in the general proximal setting are discussed in the appendix). In particular, if  $\hat{S} = \{1, 2, \dots, n\}$  with probability 1, ALPHA reduces to a deterministic method: *gradient descent* (GD; Algorithm 3) or *accelerated gradient descent* (AGD; Algorithm 4), depending on the choice of the sequence  $\{\theta_k\}$ . For a non-deterministic sampling, we obtain *parallel coordinate descent* (PCD; Algorithm 5) and *accelerated parallel coordinate descent* (APCD; Algorithm 6) with *arbitrary sampling* – which is new. If a uniform sampling is used, PCD reduces to the PCDM algorithm [32]. If a distributed<sup>2</sup> sampling is used instead, PCD reduces to Hydra [28]. Similarly, if a uniform sampling is used, APCD reduces to APPROX [6] (in fact, our version of APPROX is a bit more flexible with respect to choice of  $\theta_0$ , which leads to a better complexity result). APCD specialized to a distributed sampling reduces to Hydra<sup>2</sup> [5].

**Robustness.** Since we establish a complexity result for an arbitrary sampling, one of the key contributions of this work is to show that coordinate descent methods are robust to the choice of the sampling  $\hat{S}$ . In many applications one is *forced* to sample the coordinates/blocks in a non-traditional way and up to this point the issue of whether the resulting algorithm would converge (let alone the issue of estimating its complexity) was open. For instance, in many metric learning / matrix problems one wishes to find a positive semidefinite matrix satisfying certain properties. It is often efficient to work with an algorithm which would in each iteration update all the elements in a certain row and the corresponding column of the matrix. If we think of the elements of this matrix as coordinates, then any sampling induced in this way puts more probability on the diagonal

---

<sup>1</sup>We have named the method ALPHA because of this flexibility: “ALPHA” as a single source from which one obtains diversity.

<sup>2</sup>Distributed sampling is a structured uniform sampling first introduced in [28] and further studied in [5, 26] and in the companion paper [25]. In a distributed sampling, the blocks  $\{1, 2, \dots, n\}$  are first partitioned into  $c$  sets of equal cardinality (it is useful to think of  $c$  to be equal to the number of compute nodes in a distributed computing environment). The sampling is constructed by letting each node choose a subset of a fixed size (say  $\tau$ ) of the blocks it owns, uniformly at random and independently from others, and then taking the union of these random sets. This union is a random subset of the set of blocks; and is called the  $(c, \tau)$ -distributed sampling.

elements of than on the off-diagonal elements. An algorithm of this type was not analyzed before. The complexity of such a method would follow as a special of our general results specialized to the corresponding sampling.

**Improved complexity results.** In all cases where ALPHA reduces to an existing method, our complexity bound either matches the best known bound for that method or improves upon the best known bound. For instance, while the complexity of PCDM [32] (which coincides with PCD specialized to a uniform sampling; Algorithm 5) depends on the size of a certain level-set of  $f$  (and in particular, requires the level set to be bounded), our bound does not involve this quantity (see Section A.3). Another example is APPROX [6] (which is closely related to APCD specialized to a uniform sampling): we obtain a more compact and improved result.

**Two in one.** We provide a *unified* complexity analysis covering the nonaccelerated and accelerated variants of ALPHA. This is achieved by establishing a certain key technical recursion (Lemma 6.4) for an *arbitrary* choice of the parameters  $\{\theta_k\}$ . Since the two variants of ALPHA differ in the choice of this sequence only, the analysis of both is identical up to this point. The recursion is then analyzed in two different ways, depending on the sequence  $\{\theta_k\}$ , which leads to the final complexity result.

**Efficient implementation.** As formulated in Algorithm 1, ALPHA seems to require that two vectors in  $\mathbb{R}^N$  be added at each iteration (unless the three sequences coincide, which happens in some important special cases). Motivated by [13, 6], in Section 5 we give an equivalent form of Algorithm 1, which under some structural assumptions on  $f$  (see (51)) does not require such full-dimensional operations. This is important as the efficiency of coordinate descent methods largely stems from their ability to decompose the problem into subproblems, in an iterative fashion, of much smaller size than is the size of the original problem.

## 1.4 Outline of the paper

The paper is organized as follows. In Section 2 we establish notation, describe the ALPHA algorithm and comment on the key assumption: Expected Separable Overapproximation (ESO). We defer the in-depth study ESO inequalities to a companion paper [25]. In Section 3 we give a simple complexity proof of ALPHA in the smooth case ( $\psi = 0$ ). Subsequently, in Section 4 we present four algorithms that ALPHA reduces to in special cases, and state the corresponding complexity results, which follow from our general result, in a simplified form. We do this for the benefit of the reader. In Section 5 we provide an equivalent form of writing ALPHA—one leading to an efficient implementation avoiding full dimensional operations. In Section 6 we state and prove the convergence result of ALPHA when applied to the general proximal minimization problem (1). Finally, in Section 7 we conclude and in the appendix we comment on several special cases ALPHA reduces to in the general proximal setup.

## 2 The Algorithm

In this section we first formalize the block structure of  $\mathbb{R}^N$  and establish necessary notation (Section 2.3), then proceed to describing the ALPHA algorithm (Section 2.2) and finally comment

in the key assumption needed for our complexity results (Section 2.3).

## 2.1 Preliminaries

**Blocks.** We first describe the block setup which has become standard in the analysis of block coordinate descent methods [24, 31, 32, 6]. The space  $\mathbb{R}^N$  is decomposed into  $n$  subspaces:  $\mathbb{R}^N = \mathbb{R}^{N_1} \times \dots \times \mathbb{R}^{N_n}$ . Let  $\mathbf{U}$  be the  $N \times N$  identity matrix and  $\mathbf{U} = [\mathbf{U}_1, \dots, \mathbf{U}_n]$  be its decomposition into column submatrices  $\mathbf{U}_i \in \mathbb{R}^{N \times N_i}$ . For  $x \in \mathbb{R}^N$ , let  $x^i \in \mathbb{R}^{N_i}$  be the block of coordinates corresponding to the columns of  $\mathbf{U}_i$ , i.e.,  $x^i = \mathbf{U}_i^\top x$ . For any  $h \in \mathbb{R}^N$  and  $S \subseteq [n] \stackrel{\text{def}}{=} \{1, \dots, n\}$  we define:

$$h_{[S]} \stackrel{\text{def}}{=} \sum_{i \in S} \mathbf{U}_i h^i. \quad (2)$$

For function  $f$ , we denote by  $\nabla f(x)$  the gradient of  $f$  at point  $x \in \mathbb{R}^N$  and by  $\nabla_i f(x) \in \mathbb{R}^{N_i}$  the block of partial derivatives  $\nabla_i f(x) = \mathbf{U}_i^\top \nabla f(x)$ .

**Norms.** The standard Euclidean inner product (with respect to the standard basis) in spaces  $\mathbb{R}^N$  and  $\mathbb{R}^{N_i}$ ,  $i \in [n]$ , will be denoted by  $\langle \cdot, \cdot \rangle$ . That is, for vectors  $x, y$  of equal size, we have  $\langle x, y \rangle = x^\top y$ . Each space  $\mathbb{R}^{N_i}$  is equipped with a Euclidean norm:

$$\|x^i\|_i^2 \stackrel{\text{def}}{=} \langle \mathbf{B}_i x^i, x^i \rangle = (x^i)^\top \mathbf{B}_i x^i, \quad (3)$$

where  $\mathbf{B}_i$  is an  $N_i$ -by- $N_i$  positive definite matrix. For  $w \in \mathbb{R}_{++}^n$  and  $x, y \in \mathbb{R}^N$  we further define

$$\langle x, y \rangle_w \stackrel{\text{def}}{=} \sum_{i=1}^n w_i \langle x^i, y^i \rangle \quad (4)$$

and

$$\|x\|_w^2 \stackrel{\text{def}}{=} \sum_{i=1}^n w_i \|x^i\|_i^2 \stackrel{(3)}{=} \sum_{i=1}^n w_i \langle \mathbf{B}_i x^i, x^i \rangle. \quad (5)$$

For  $x \in \mathbb{R}^N$ , by  $\mathbf{B}x$  we mean the vector  $\mathbf{B}x = \sum_{i=1}^n \mathbf{U}_i \mathbf{B}_i x^i$ . That is,  $\mathbf{B}x$  is the vector in  $\mathbb{R}^N$  whose  $i$ th block is equal to  $\mathbf{B}_i x^i$ . For vectors  $x, y \in \mathbb{R}^N$  we have

$$\|x + y\|_w^2 = \|x\|_w^2 + 2\langle \mathbf{B}x, y \rangle_w + \|y\|_w^2. \quad (6)$$

**Vectors.** For any two vectors  $x$  and  $y$  of the same size, we denote by  $x \circ y$  their Hadamard (i.e., elementwise) product. By abuse of notation, we denote by  $u^2$  the elementwise square of the vector  $u$ , by  $u^{-1}$  the elementwise inverse of vector  $u$  and by  $u^{-2}$  the elementwise square of  $u^{-1}$ . For vector  $v \in \mathbb{R}^n$  and  $x \in \mathbb{R}^N$  we will write

$$v \cdot x \stackrel{\text{def}}{=} \sum_{i=1}^n v_i \mathbf{U}_i x^i. \quad (7)$$

That is,  $v \cdot x$  is the vector obtained from  $x$  by multiplying its block  $i$  by  $v_i$  for each  $i \in [n]$ . If all blocks are of size one ( $N_i = 1$  for all  $i$ ), then  $v \cdot x = \text{Diag}(v)x$  where  $\text{Diag}(v)$  is the diagonal matrix with diagonal vector  $v$ .

## 2.2 ALPHA

In this section we describe ALPHA (Algorithm 1) – a randomized block coordinate descent method for solving (1).

We denote by  $\text{dom } \psi$  the domain of the proximal term  $\psi$ .

---

### Algorithm 1 ALPHA

---

- 1: **Parameters:** proper sampling  $\hat{S}$  with probability vector  $p = (p_1, \dots, p_n)$ ,  $v \in \mathbb{R}_{++}^n$ , sequence  $\{\theta_k\}_{k \geq 0}$
  - 2: **Initialization:** choose  $x_0 \in \text{dom } \psi$  and set  $z_0 = x_0$
  - 3: **for**  $k \geq 0$  **do**
  - 4:    $y_k = (1 - \theta_k)x_k + \theta_k z_k$
  - 5:   Generate a random set of blocks  $S_k \sim \hat{S}$
  - 6:    $z_{k+1} \leftarrow z_k$
  - 7:   **for**  $i \in S_k$  **do**
  - 8:      $z_{k+1}^i = \arg \min_{z \in \mathbb{R}^{N_i}} \{ \langle \nabla_i f(y_k), z \rangle + \frac{\theta_k v_i}{2p_i} \|z - z_k^i\|_i^2 + \psi^i(z) \}$
  - 9:   **end for**
  - 10:    $x_{k+1} = y_k + \theta_k p^{-1} \cdot (z_{k+1} - z_k)$
  - 11: **end for**
- 

To facilitate the presentation, we first recall some basic facts and terminology related to samplings (for a broader coverage see [32, 6, 26, 25]). A sampling  $\hat{S}$  is a random set valued mapping with values in  $2^{[n]}$ . We say that sampling  $\hat{S}$  is *nil* if  $\mathbb{P}(\hat{S} = \emptyset) = 1$ , *proper* if  $\mathbb{P}(i \in \hat{S}) > 0$  for all  $i \in [n]$ , *uniform* if  $\mathbb{P}(i \in \hat{S}) = \mathbb{P}(i' \in \hat{S})$  for all  $i, i' \in [n]$  and *serial* if  $\mathbb{P}(|\hat{S}| = 1) = 1$ . The probability that the block  $i$  is chosen is denoted by:

$$p_i \stackrel{\text{def}}{=} \mathbb{P}(i \in \hat{S}), \quad i \in [n]. \quad (8)$$

It is easy to see that if  $\hat{S}$  is uniform, then  $p_i = \mathbb{E}[|\hat{S}|]/n$ , for  $i \in [n]$ .

Algorithm 1 starts from an initial vector  $x_0 \in \mathbb{R}^N$  and generates three sequences  $\{x_k, y_k, z_k\}_{k \geq 0}$ . At iteration  $k$ , a random subset of blocks,  $S_k \subseteq [n]$ , is generated according to the distribution of sampling  $\hat{S}$  (a parameter to enter at the beginning of the algorithm). In order to guarantee that each block has a nonzero probability to be selected, it is necessary to assume that  $\hat{S}$  is *proper*.

To move from  $z_k$  to  $z_{k+1}$ , we only need to evaluate  $|S_k|$  partial derivatives of  $f$  at point  $y_k$  and update only the blocks of  $z_k$  belonging to  $S_k$  to the solutions of  $|S_k|$  proximal problems (Steps 6 to 8). The vector  $x_{k+1}$  is obtained from  $y_k$  by changing only the blocks of  $y_k$  belonging to  $S_k$  (Step 10). The vector  $y_k$  is a convex combination of  $x_k$  and  $z_k$  (Step 4) with coefficient  $\theta_k$ , a parameter to be chosen between  $(0, 1]$ . Note that unless  $\theta_k p_i = 1$  for all  $i \in [n]$  and  $k \in \mathbb{N}$ , in which case the three sequences  $\{x_k, y_k, z_k\}_{k \geq 0}$  reduce to one same sequence, the update in Step 4 requires a full-dimensional vector operation, as previously remarked in [24, 6]. In Section 5 we will provide an equivalent form of Algorithm 1, which avoids full-dimensional vector operations for special forms of  $f$ .

Let us extract the relations between the three sequences. Define

$$\tilde{z}_{k+1} \stackrel{\text{def}}{=} \arg \min_{z \in \mathbb{R}^N} \{ \langle \nabla f(y_k), z \rangle + \frac{\theta_k}{2} \|z - z_k\|_{p^{-1} \circ v}^2 + \psi(z) \}. \quad (9)$$



Then

$$z_{k+1}^i = \begin{cases} \tilde{z}_{k+1}^i & i \in S_k \\ z_k^i & i \notin S_k \end{cases}, \quad (10)$$

and hence  $z_{k+1} - z_k = (\tilde{z}_{k+1} - z_k)_{[S_k]}$  and

$$x_{k+1} = y_k + \theta_k p^{-1} \cdot (\tilde{z}_{k+1} - z_k)_{[S_k]}. \quad (11)$$

Note also that from the definition of  $y_k$  in Algorithm 1, we have:

$$\theta_k(y_k - z_k) = (1 - \theta_k)(x_k - y_k). \quad (12)$$

### 2.3 Expected separable overapproximation

To guarantee the convergence of Algorithm 1, we shall require that  $f$  admits an expected separable overapproximation (ESO) with respect to the sampling  $\hat{S}$  with parameter  $v \in \mathbb{R}_{++}^n$ :

**Assumption 2.1** (ESO assumption). *Let  $\hat{S}$  be a sampling and  $v \in \mathbb{R}_{++}^n$  a vector of positive weights. We say that the function  $f$  admits an ESO with respect to  $\hat{S}$  with parameter  $v$ , denoted as  $(f, \hat{S}) \sim ESO(v)$ , if the following inequality holds for all  $x, h \in \mathbb{R}^N$ ,*

$$\mathbb{E}[f(x + h_{[\hat{S}]})] \leq f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2} \|h\|_{v \circ p}^2, \quad (13)$$

where  $p = (p_1, \dots, p_n) \in \mathbb{R}^n$  is the vector of probabilities associated with  $\hat{S}$  defined in (8).

Looking behind the compact notation in which the assumption is formulated, observe that the upper bound is a quadratic function of  $h = (h^1, \dots, h^n)$ , separable in the blocks  $h^i$ :

$$f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2} \|h\|_{v \circ p}^2 \stackrel{(4)}{=} f(x) + \sum_{i=1}^n p_i (\langle \nabla_i f(x), h^i \rangle + v_i \langle \mathbf{B}_i h^i, h^i \rangle).$$

As a tool for the design and analysis of randomized coordinate descent methods, ESO was first formulated in [32] for the complexity study of parallel coordinate descent method (PCDM). It is a powerful technical tool which provides a generic approach to establishing the convergence of randomized coordinate descent methods of many flavours [32, 38, 7, 39, 29, 28, 5, 26]. As shown in the listed papers as well as our results which follow, the convergence of Algorithm 1 can be established for arbitrary sampling  $\hat{S}$  as long as the parameter vector  $v$  is chosen such that  $(f, \hat{S}) \sim ESO(v)$ . Moreover, the vector  $v$  appears in the convergence result and directly influences the complexity of the method.

Since [32], the problem of computing efficiently a vector  $v$  such that the ESO assumption 2.1 holds has been addressed in many papers for special uniform samplings relevant to practical implementation including serial sampling,  $\tau$ -nice sampling [32, 6] and distributed sampling [28, 5], and also for a particular example of nonuniform parallel sampling [29]. In this paper we focus on the complexity analysis of Algorithm 1 and refer the reader to the companion paper [25] for a systematic study of the computation of admissible vector parameter  $v$  for arbitrary sampling  $\hat{S}$ .

### 3 Simple complexity analysis in the smooth case

In this section we give a brief complexity analysis of ALPHA in the case when  $\psi \equiv 0$ ; that is, when applied to the following unconstrained smooth convex minimization problem:

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && x = (x^1, \dots, x^n) \in \mathbb{R}^{N_1} \times \dots \times \mathbb{R}^{N_n} = \mathbb{R}^N, \end{aligned} \quad (14)$$

While our general theory, which we develop in Section 6, covers also this special case, the analysis we present here is different and simpler. When applied to problem 14, Step 8 in ALPHA has an explicit solution and the method reduces to Algorithm 2.

---

**Algorithm 2** ALPHA specialized to the smooth minimization problem (14)

---

- 1: **Parameters:** proper sampling  $\hat{S}$  with probability vector  $p = (p_1, \dots, p_n)$ , vector  $v \in \mathbb{R}_{++}^n$ , sequence  $\{\theta_k\}_{k \geq 0}$
  - 2: **Initialization:** choose  $x_0 \in \mathbb{R}^N$ , set  $z_0 = x_0$
  - 3: **for**  $k \geq 0$  **do**
  - 4:    $y_k = (1 - \theta_k)x_k + \theta_k z_k$
  - 5:   Generate a random set of blocks  $S_k \sim \hat{S}$
  - 6:    $z_{k+1} \leftarrow z_k$
  - 7:   **for**  $i \in S_k$  **do**
  - 8:      $z_{k+1}^i = z_k^i - \frac{p_i}{v_i \theta_k} \mathbf{B}_i^{-1} \nabla_i f(y_k)$
  - 9:   **end for**
  - 10:    $x_{k+1} = y_k + \theta_k p^{-1} \cdot (z_{k+1} - z_k)$
  - 11: **end for**
- 

We now state the complexity result for ALPHA (Algorithm 2) in its nonaccelerated variant.

**Theorem 3.1** (ALPHA – smooth & nonaccelerated). *Let  $\hat{S}$  be an arbitrary proper sampling and  $v \in \mathbb{R}_{++}^n$  be such that  $(f, \hat{S}) \sim ESO(v)$ . Choose  $\theta_k = \theta_0 \in (0, 1]$  for all  $k \geq 0$ . Then for any  $y \in \mathbb{R}^N$ , the iterates  $\{x_k\}_{k \geq 1}$  of Algorithm 2 satisfy:*

$$\max \left\{ \mathbb{E}[f(\hat{x}_k)], \min_{l=1, \dots, k} \mathbb{E}[f(x_l)] \right\} - f(y) \leq \frac{C}{(k-1)\theta_0 + 1}, \forall k \geq 1 \quad (15)$$

where

$$\hat{x}_k = \frac{x_k + \theta_0 \sum_{l=1}^{k-1} x_l}{1 + (k-1)\theta_0}$$

and

$$C = (1 - \theta_0) (f(x_0) - f(y)) + \frac{\theta_0^2}{2} \|x_0 - y\|_{v \circ p^{-2}}^2.$$

In particular, if we choose  $\theta_0 = \min_i p_i$ , then for all  $k \geq 1$ ,

$$\max \left\{ \mathbb{E}[f(\hat{x}_k)], \min_{l=1, \dots, k} \mathbb{E}[f(x_l)] \right\} - f(y) \leq \frac{(1 - \min_i p_i) (f(x_0) - f(y)) + \frac{1}{2} \|x_0 - y\|_v^2}{(k-1) \min_i p_i + 1}. \quad (16)$$

The next result gives a complexity bound for ALPHA (Algorithm 2) in its accelerated variant.

**Theorem 3.2** (ALPHA – smooth and accelerated). *Let  $\hat{S}$  be an arbitrary proper sampling and  $v \in \mathbb{R}_{++}^n$  be such that  $(f, \hat{S}) \sim \text{ESO}(v)$ . Choose  $\theta_0 \in (0, 1]$  and define the sequence  $\{\theta_k\}_{k \geq 0}$  by*

$$\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2 - \theta_k^2}}{2}. \quad (17)$$

*Then for any  $y \in \mathbb{R}^N$  such that  $C \geq 0$ , the iterates  $\{x_k\}_{k \geq 1}$  of Algorithm 2 satisfy:*

$$\mathbb{E}[f(x_k)] - f(y) \leq \frac{4C}{((k-1)\theta_0 + 2)^2}, \quad (18)$$

where

$$C = (1 - \theta_0)(f(x_0) - f(y)) + \frac{\theta_0^2}{2} \|x_0 - y\|_{v \circ p^{-2}}^2.$$

*In particular, if we choose  $\theta_0 = 1$ , then for all  $k \geq 1$ ,*

$$\mathbb{E}[f(x_k)] - f(y) \leq \frac{2\|x_0 - y\|_{v \circ p^{-2}}^2}{(k+1)^2} = \frac{2 \sum_{i=1}^n \frac{v_i}{p_i^2} \|x_0^i - y^i\|^2}{(k+1)^2}. \quad (19)$$

In the rest of this section, we provide a short proof of Theorems 3.1 and 3.2. In Section 6 we shall present complexity bounds (Theorem 6.1 and 6.2) for ALPHA (Algorithm 1) as applied to the general regularized problem (1). The proof in the general case is more involved, which is why we prefer to present the smooth case first and also provide a separate briefer proof.

### 3.1 Two Lemmas

We first establish two lemmas and then proceed directly to the proofs of the theorems.

**Lemma 3.1.** *For any sampling  $\hat{S}$  and any  $x, a \in \mathbb{R}^N$  and  $w \in \mathbb{R}_{++}^n$ , the following identity holds:*

$$\|x\|_w^2 - \mathbb{E} \left[ \|x + a_{[\hat{S}]}\|_w^2 \right] = \|x\|_{w \circ p}^2 - \|x + a\|_{w \circ p}^2.$$

*Proof.* It is sufficient to notice that  $\mathbb{E} \left[ \|x + a_{[\hat{S}]}\|_w^2 \right] = \sum_{i=1}^n [(1-p_i)w_i \|x^i\|_i^2 + p_i w_i \|x^i + a^i\|_i^2]$ .  $\square$

**Lemma 3.2.** *Let  $\hat{S}$  be an arbitrary proper sampling and  $v \in \mathbb{R}_{++}^n$  be such that  $(f, \hat{S}) \sim \text{ESO}(v)$ . Let  $\{\theta_k\}_{k \geq 0}$  be an arbitrary sequence of positive numbers in  $(0, 1]$  and fix  $y \in \mathbb{R}^N$ . Then for the sequence of iterates produced by Algorithm 2 and all  $k \geq 0$ , the following recursion holds:*

$$\mathbb{E}_k \left[ f(x_{k+1}) + \frac{\theta_k^2}{2} \|z_{k+1} - y\|_{v \circ p^{-2}}^2 \right] \leq \left[ f(x_k) + \frac{\theta_k^2}{2} \|z_k - y\|_{v \circ p^{-2}}^2 \right] - \theta_k (f(x_k) - f(y)). \quad (20)$$

*Proof.* Based on line 8 of Algorithm 2, we can write

$$a \stackrel{\text{def}}{=} \tilde{z}_{k+1} - z_k = -\theta_k^{-1} (v^{-1} \circ p) \cdot \mathbf{B}^{-1} \nabla f(y_k), \quad (21)$$

or equivalently,  $-\nabla f(y_k) = \theta_k (v \circ p^{-1}) \cdot \mathbf{B}a$ . Using this notation, the update on line 10 of Algorithm 2 can be written as

$$x_{k+1} = y_k + \theta_k p^{-1} \cdot a_{[S_k]} = y_k + (\theta_k p^{-1} \cdot a)_{[S_k]}. \quad (22)$$

Letting  $b = \tilde{z}_{k+1} - y$  and  $t = \theta_k^2(v \circ p^{-1})$ , we apply the ESO assumption and rearrange the result:

$$\begin{aligned}
\mathbb{E}_k[f(x_{k+1})] &\stackrel{(13)+(22)}{\leq} f(y_k) + \langle \nabla f(y_k), \theta_k p^{-1} \cdot a \rangle_p + \frac{1}{2} \|\theta_k p^{-1} \cdot a\|_{v \circ p}^2 \\
&\stackrel{(4)+(5)+(21)}{=} f(y_k) - \frac{1}{2} \|a\|_t^2 \\
&\stackrel{(6)}{=} f(y_k) - \frac{1}{2} \|b\|_t^2 + \frac{1}{2} \|b - a\|_t^2 + \langle \mathbf{B}a, b - a \rangle_t.
\end{aligned} \tag{23}$$

Note that  $\|b\|_t^2 = \theta_k^2 \|\tilde{z}_{k+1} - y\|_{v \circ p^{-1}}^2$ ,  $\|b - a\|_t^2 = \theta_k^2 \|z_k - y\|_{v \circ p^{-1}}^2$  and

$$\begin{aligned}
\langle \mathbf{B}a, b - a \rangle_t &= \langle -\mathbf{B}a, a - b \rangle_t = \langle \theta_k^{-1}(v^{-1} \circ p) \cdot \nabla f(y_k), y - z_k \rangle_t \\
&= \theta_k \langle \nabla f(y_k), y - z_k \rangle \stackrel{(12)}{=} \theta_k \langle \nabla f(y_k), y - y_k \rangle + (1 - \theta_k) \langle \nabla f(y_k), x_k - y_k \rangle \\
&\leq \theta_k (f(y) - f(y_k)) + (1 - \theta_k) (f(x_k) - f(y_k)).
\end{aligned}$$

Substituting these expressions to (23), we obtain the recursion:

$$\mathbb{E}_k[f(x_{k+1})] \leq \theta_k f(y) + (1 - \theta_k) f(x_k) + \frac{\theta_k^2}{2} \|z_k - y\|_{v \circ p^{-1}}^2 - \frac{\theta_k^2}{2} \|\tilde{z}_{k+1} - y\|_{v \circ p^{-1}}^2. \tag{24}$$

It now only remains to apply Lemma 3.1 to the last two terms in (24), with  $x \leftarrow z_k - y$ ,  $w \leftarrow v \circ p^{-2}$  and  $\hat{S} \leftarrow S_k$ , and rearrange the resulting inequality.  $\square$

### 3.2 Proof of Theorem 3.1

Using the fact that  $\theta_k = \theta_0$ , for all  $k$  and taking expectation in both sides of (20), we obtain the recursion

$$\phi_{k+1} + \theta_0^2 r_{k+1} \leq (1 - \theta_0) \phi_k + \theta_0^2 r_k, \quad k \geq 0,$$

where  $\phi_k \stackrel{\text{def}}{=} \mathbb{E}[f(x_k)] - f(y)$  and  $r_k \stackrel{\text{def}}{=} \frac{1}{2} \mathbb{E}[\|z_k - y\|_{v \circ p^{-2}}^2]$ . Combining these inequalities, we get

$$(1 + \theta_0(k-1)) \min_{l=1, \dots, k} \phi_l \leq \phi_k + \theta_0 \sum_{l=1}^{k-1} \phi_l \leq (1 - \theta_0) \phi_0 + \theta_0^2 r_0. \tag{25}$$

Let  $\alpha_k = 1 + (k-1)\theta_0$ . By convexity,

$$f(\hat{x}_k) = f\left(\frac{x_k + \sum_{l=1}^{k-1} \theta_0 x_l}{\alpha_k}\right) \leq \frac{f(x_k) + \sum_{l=1}^{k-1} \theta_0 f(x_l)}{\alpha_k}.$$

Finally, subtracting  $f(y)$  from both sides and taking expectations, we obtain

$$\mathbb{E}[f(\hat{x}_k)] - f(y) \leq \frac{\phi_k + \sum_{l=1}^{k-1} \theta_0 \phi_l}{\alpha_k} \stackrel{(25)}{\leq} \frac{(1 - \theta_0) \phi_0 + \theta_0^2 r_0}{\alpha_k}.$$

### 3.3 Proof of Theorem 3.2

If  $\theta_0 \in (0, 1]$ , then the sequence  $\{\theta_k\}_{k \geq 0}$  has the following properties (see [41]):

$$0 < \theta_{k+1} \leq \theta_k \leq \frac{2}{k + 2/\theta_0} \leq 1, \quad (26)$$

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} = \frac{1}{\theta_k^2}. \quad (27)$$

After dividing both sides of (20) by  $\theta_k^2$ , using (27) and taking expectations, we obtain:

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} \phi_{k+1} + r_{k+1} \leq \frac{1 - \theta_k}{\theta_k^2} \phi_k + r_k \leq \frac{1 - \theta_0}{\theta_0^2} \phi_0 + r_0, \quad (28)$$

where  $\phi_k$  and  $r_k$  are as in the proof of Theorem 3.1. Finally,

$$\begin{aligned} \phi_k &\stackrel{(27)}{=} \frac{(1 - \theta_k)\theta_{k-1}^2}{\theta_k^2} \phi_k \leq \frac{(1 - \theta_k)\theta_{k-1}^2}{\theta_k^2} \phi_k + \theta_{k-1}^2 r_k \stackrel{(28)}{\leq} \frac{(1 - \theta_0)\theta_{k-1}^2}{\theta_0^2} \phi_0 + \theta_{k-1}^2 r_0 \\ &= \frac{\theta_{k-1}^2}{\theta_0^2} ((1 - \theta_0)\phi_0 + \theta_0^2 r_0) = \frac{\theta_{k-1}^2}{\theta_0^2} C \stackrel{(26)}{\leq} \frac{4C}{((k-1)\theta_0 + 2)^2}. \end{aligned}$$

Note that in the last inequality we used the assumption that  $C \geq 0$ .

## 4 The many variants of ALPHA (in the smooth case)

The purpose of this section is to demonstrate that ALPHA is a very flexible method, encoding several classical as well as modern optimization methods for special choices of the parameters of the method. In order to achieve this goal, it is enough to focus on the smooth case, i.e., on Algorithm 2. Similar reasoning can be applied to the proximal case.

Note that in ALPHA we have the liberty to *choose* the sampling  $\hat{S}$  and the sequence  $\{\theta_k\}_{k \geq 0}$ . As we have already seen, by modifying the sequence we can obtain *simple* (i.e., nonaccelerated) and *accelerated* variants of the method. By the choice of the sampling, we can force the method to be *deterministic* or *randomized*. In the latter case, there are many ways of choosing the distribution of the sampling. Here we will constrain ourselves to a basic classification between *uniform samplings* (samplings for which  $p_i = p_{i'}$  for all  $i \in [n]$ ) and non-uniform or *importance samplings*. This is summarized in Table 1.

The deterministic variants of ALPHA (Algorithm 3 and 4) are obtained by choosing the sampling which always selects all blocks:  $\hat{S} = [n]$  with probability 1. The ESO assumption in this special case has the form

$$f(x + h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \|h\|_v^2, \quad \forall x, h \in \mathbb{R}^N, \quad (29)$$

which simply requires the gradient of  $f$  to be 1-Lipschitz with respect to the norm  $\|\cdot\|_v$ . Note that  $\|h\|_v^2 = h^\top \tilde{\mathbf{B}} h$ , where  $\tilde{\mathbf{B}}$  is the block diagonal matrix defined by  $\tilde{\mathbf{B}} \stackrel{\text{def}}{=} \text{Diag}(v_1 \mathbf{B}_1, \dots, v_n \mathbf{B}_n)$ . Likewise, if there is just a single block in our block setup (i.e., if  $n = 1$ ) then it is natural to only consider a sampling which picks this block with probability 1, which again results in a deterministic

Parameters of Algorithm 2	$\hat{S}$		$\theta_k$		
Setting	$ \hat{S}  = n$	$ \hat{S}  < n$		$\theta_{k+1} = \theta_k$	$\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$
		$p_i = p_{i'}, \forall i, i' \in [n]$	$p_i \neq p_{i'}, \exists i, i' \in [n]$		
Characteristic	Deterministic	Randomized		Simple	Accelerated
		Uniform sampling	Importance sampling		
Special cases					
Algorithm 3	✓			✓	
Algorithm 4	✓				✓
Algorithm 5	✓	✓	✓	✓	
Algorithm 6	✓	✓	✓		✓

Table 1: Special cases of Algorithm 2.

method. However, in this case the norm  $\|\cdot\|_v$  can be an arbitrary Euclidean norm (that is, it does not need to be block diagonal).

In the randomized variants of ALPHA (Algorithm 5 and 6) we allow for the sampling to have an *arbitrary distribution*.

#### 4.1 Special case 1: gradient descent

By specializing Algorithm 2 to the choice  $\hat{S} = [n]$  and  $\theta_k = 1$  for all  $k$ , we obtain classical gradient descent (with fixed stepsize). Indeed, note that in this special case we have

$$x_k = y_k = z_k, \quad \forall k \geq 1. \quad (30)$$

Recall that the ESO assumption reduces to (29) when  $\hat{S} = [n]$ .

---

#### Algorithm 3 Gradient Descent (GD) for solving (14)

---

- 1: **Parameters:** vector  $v \in \mathbb{R}_{++}^n$  such that (29) holds
  - 2: **Initialization:** choose  $x_0 \in \mathbb{R}^N$
  - 3: **for**  $k \geq 0$  **do**
  - 4:   **for**  $i \in [n]$  **do**
  - 5:      $x_{k+1}^i = x_k^i - \frac{1}{v_i} \mathbf{B}_i^{-1} \nabla_i f(x_k)$
  - 6:   **end for**
  - 7: **end for**
- 

The complexity of the method is a corollary of Theorem 3.1.

**Corollary 4.1.** *For any optimal solution  $x_*$  of (14), the output of Algorithm 3 for all  $k \geq 1$  satisfies:*

$$f(x_k) - f(x_*) \leq \frac{\|x_0 - x_*\|_v^2}{2k}. \quad (31)$$

In particular, for  $\epsilon > 0$ , if

$$k \geq \frac{\|x_0 - x_*\|_v^2}{2\epsilon}, \quad (32)$$

then  $f(x_k) - f(x_*) \leq \epsilon$ .

*Proof.* By letting  $y = x_k$  in (20) we know that:

$$f(x_{k+1}) \leq f(x_k) + \frac{\theta_k^2}{2} \|z_k - x_k\|^2 \stackrel{(30)}{=} f(x_k), \quad \forall k \geq 1.$$

Note that for this special case  $x_k = z_k$ . Therefore,

$$f(x_k) - f(x_*) = \min_{l=1, \dots, k} f(x_l) - f(x_*) \leq \frac{\|x_0 - x_*\|_v^2}{2k}, \quad \forall k \geq 1,$$

where the second inequality follows from applying Theorem 3.1 to  $\theta_0 = 1$  and  $\hat{S} = [n]$ .  $\square$

Corollary 4.1 is a basic result and can be found in many textbooks on convex optimization; see for example [23].

## 4.2 Special case 2: accelerated gradient descent

Let us still keep  $\hat{S} = [n]$ , but assume now the sequence  $\{\theta_k\}_{k \geq 0}$  is chosen according to (17). In this case, Algorithm 2 reduces to accelerated gradient descent.

---

**Algorithm 4** Accelerated Gradient Descent (AGD) for solving (14)

---

- 1: **Parameters:** positive vector  $v \in \mathbb{R}_{++}^n$  such that (29) holds
  - 2: **Initialization:** choose  $x_0 \in \mathbb{R}^N$ , set  $z_0 = x_0$  and  $\theta_0 = 1$
  - 3: **for**  $k \geq 0$  **do**
  - 4:   **for**  $i \in [n]$  **do**
  - 5:      $z_{k+1}^i = z_k^i - \frac{1}{v_i \theta_k} \mathbf{B}_i^{-1} \nabla_i f((1 - \theta_k)x_k + \theta_k z_k)$
  - 6:   **end for**
  - 7:    $x_{k+1} = (1 - \theta_k)x_k + \theta_k z_{k+1}$
  - 8:    $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k}{2}$
  - 9: **end for**
- 

Note that only two sequences  $\{x_k, z_k\}_{k \geq 0}$  are explicitly used in Algorithm 4. This is achieved by replacing  $y_k$  in Algorithm 2 by  $(1 - \theta_k)x_k + \theta_k z_k$ . The following result follows directly from Theorem 3.2 by letting  $\theta_0 = 1$  and  $p_i = 1$  for all  $i \in [n]$ .

**Corollary 4.2.** *For any optimal solution  $x_*$  of (14), the output of Algorithm 4 for all  $k \geq 1$  satisfies:*

$$f(x_k) - f(x_*) \leq \frac{2\|x_0 - x_*\|_v^2}{(k+1)^2}. \quad (33)$$

In particular, for  $\epsilon > 0$ , if

$$k \geq \sqrt{\frac{2\|x_0 - x_*\|_v^2}{\epsilon}} - 1, \quad (34)$$

then  $f(x_k) - f(x_*) \leq \epsilon$ .

Algorithm 4 is a special case of Algorithm 1 in [41]. The complexity bound (33) was also proved in [41, Corollary 1]. See also [6].

### 4.3 Special case 3: Parallel coordinate descent

We now allow the method (Algorithm 2) to use an arbitrary sampling  $\hat{S}$ , but keep  $\theta_k = \theta_0$  for all  $k \geq 1$ . This leads to Algorithm 5.

---

**Algorithm 5** Parallel Coordinate Descent (PCD) for solving (14)

---

- 1: **Parameters:** proper sampling  $\hat{S}$  with probability vector  $p = (p_1, \dots, p_n)$ , vector  $v \in \mathbb{R}_{++}^n$  for which  $(f, \hat{S}) \sim ESO(v)$
  - 2: **Initialization:** choose  $x_0 \in \mathbb{R}^N$ , set  $z_0 = x_0$  and  $\theta_0 = \min_i p_i$
  - 3: **for**  $k \geq 0$  **do**
  - 4:    $y_k = (1 - \theta_0)x_k + \theta_0 z_k$
  - 5:   Generate a random set of blocks  $S_k \sim \hat{S}$
  - 6:    $z_{k+1} \leftarrow z_k$
  - 7:   **for**  $i \in S_k$  **do**
  - 8:      $z_{k+1}^i = z_k^i - \frac{p_i}{v_i \theta_0} \mathbf{B}_i^{-1} \nabla_i f(y_k)$
  - 9:   **end for**
  - 10:    $x_{k+1} = y_k + \theta_0 p^{-1} \cdot (z_{k+1} - z_k)$
  - 11: **end for**
- 

Note that in classical non-accelerated coordinate descent methods, only a single sequence of iterates is needed. This is indeed the case for our method as well, in the special case when the sampling  $\hat{S}$  is uniform and  $\theta_0 = \mathbb{E}[|\hat{S}|]/n$ , so that the three sequences are equal to each other. We now state a direct corollary of Theorem 3.1.

**Corollary 4.3.** *For any optimal solution  $x_*$  of (14), the output of Algorithm 5 for all  $k \geq 1$  satisfies:*

$$\begin{aligned} & \max \left\{ \mathbb{E} \left[ f \left( \frac{x_k + \min_i p_i \sum_{l=1}^{k-1} x_l}{1 + (k-1) \min_i p_i} \right) \right], \min_{l=1, \dots, k} \mathbb{E} [f(x_l)] \right\} - f(x_*) \\ & \leq \frac{(1 - \min_i p_i) (f(x_0) - f(x_*)) + \frac{1}{2} \|x_0 - x_*\|_v^2}{(k-1) \min_i p_i + 1}. \end{aligned}$$

In particular, for  $\epsilon > 0$ , if

$$k \geq \frac{(1 - \min_i p_i) (f(x_0) - f(x_*)) + \frac{1}{2} \|x_0 - x_*\|_v^2}{\min_i p_i \epsilon} - \frac{1}{\min_i p_i} + 1, \quad (35)$$

then

$$\max \left\{ \mathbb{E} \left[ f \left( \frac{x_k + \min_i p_i \sum_{l=1}^{k-1} x_l}{1 + (k-1) \min_i p_i} \right) \right], \min_{l=1, \dots, k} \mathbb{E} [f(x_l)] \right\} - f(x_*) \leq \epsilon.$$

In the special case when  $\hat{S}$  is the serial uniform sampling, the three sequences  $\{x_k, y_k, z_k\}_{k \geq 0}$  coincide, and one can show that the following bound holds:

$$\mathbb{E}[f(x_k)] - f(x_*) \leq \frac{n}{k-1+n} \left[ \left(1 - \frac{1}{n}\right) (f(x_0) - f(x_*)) + \frac{1}{2} \|x_0 - x_*\|_v^2 \right].$$



Randomized coordinate descent with serial and importance sampling (in a form different from Algorithm 5) was considered by Nesterov [24]. In the special case of serial uniform sampling when Algorithm 5 is the same as in [24], the following convergence rate was proved in [24]:

$$\mathbb{E}[f(x_k)] - f(x_*) \leq \frac{2n}{k+4} \mathcal{R}^2(x_0)$$

where  $\mathcal{R}(x_0)$  is a weighted level-set distance to the set of optimal points  $X_*$ :

$$\mathcal{R}(x_0) \stackrel{\text{def}}{=} \max_x \left\{ \max_{x_* \in X_*} \|x_0 - x_*\|_v^2 : f(x) \leq f(x_0) \right\}.$$

Our result does not require the level sets of  $f$  to be bounded.

#### 4.4 Special case 4: Accelerated parallel coordinate descent

To obtain the accelerated coordinate descent method, as a special case of Algorithm 2, we only need to let the sequence  $\{\theta_k\}_{k \geq 0}$  satisfy (17).

---

**Algorithm 6** Accelerated parallel coordinate descent (APCD) for solving (14)

---

- 1: **Parameters:** proper sampling  $\hat{S}$  with probability vector  $p = (p_1, \dots, p_n)$ , vector  $v \in \mathbb{R}_{++}^n$  for which  $(f, \hat{S}) \sim ESO(v)$ ,  $\theta_0 = 1$
  - 2: **Initialization:** choose  $x_0 \in \mathbb{R}^N$  and set  $z_0 = x_0$
  - 3: **for**  $k \geq 0$  **do**
  - 4:    $y_k = (1 - \theta_k)x_k + \theta_k z_k$
  - 5:   Generate a random set of blocks  $S_k \sim \hat{S}$
  - 6:    $z_{k+1} \leftarrow z_k$
  - 7:   **for**  $i \in S_k$  **do**
  - 8:      $z_{k+1}^i = z_k^i - \frac{p_i}{v_i \theta_k} \mathbf{B}_i^{-1} \nabla_i f(y_k)$
  - 9:   **end for**
  - 10:    $x_{k+1} = y_k + \theta_k p^{-1} \cdot (z_{k+1} - z_k)$
  - 11:    $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k}{2}$
  - 12: **end for**
- 

The convergence result then follows directly as a corollary of Theorem 3.2.

**Corollary 4.4.** *For any optimal solution  $x_*$  of (14), the output of Algorithm 6 for all  $k \geq 1$  satisfies:*

$$\mathbb{E}[f(x_k)] - f(x_*) \leq \frac{2\|x_0 - x_*\|_{v \circ p^{-2}}^2}{(k+1)^2}. \quad (36)$$

In particular, for  $\epsilon > 0$ , if

$$k \geq \sqrt{\frac{2\|x_0 - x_*\|_{v \circ p^{-2}}^2}{\epsilon}} - 1, \quad (37)$$

then  $\mathbb{E}[f(x_k)] - f(x_*) \leq \epsilon$ .

When specialized to the serial uniform sampling (sampling  $\hat{S}$  for which  $\mathbb{P}(\hat{S} = \{i\}) = 1/n$  for  $i \in [n]$ ), the bound 36 simplifies to:

$$\mathbb{E}[f(x_k)] - f(x_*) \leq \frac{2n^2 \|x_0 - x_*\|_v^2}{(k+1)^2}. \quad (38)$$

An accelerated coordinate descent method for unconstrained minimization in the special case of serial uniform sampling was first proposed and analyzed by Nesterov [24], where the following bound was proved:

$$\mathbb{E}[f(x_k)] - f(x_*) \leq \left(\frac{n}{k+1}\right)^2 \left[2\|x_0 - x_*\|_v^2 + \frac{1}{n^2} (f(x_0) - f(x_*))\right]. \quad (39)$$

Comparing (38) and (39), it is clear that we obtain a better bound. An accelerated coordinate descent method (APPROX) utilizing an *arbitrary uniform sampling* was studied by Fercoq and Richtárik [6]. Algorithm 6, when restricted to a uniform sampling, is similar to this method. The main difference is in the value of  $\theta_0$ . Indeed, in [6],  $\theta_0$  is chosen to be  $\mathbb{E}[|\hat{S}|]/n$ , while our analysis allows  $\theta_0$  to be chosen as large as 1. This lead to larger stepsizes and a simplified and improved convergence bound.

Each serial sampling  $\hat{S}$  is uniquely characterized by the vector of probabilities  $p = (p_1, \dots, p_n)$  where  $p_i$  is defined by (8). Suppose that the function  $f$  has block-Lipschitz gradient with constants  $L_1, \dots, L_n$ :

$$f(x + \mathbf{U}_i h^i) \leq f(x) + \langle \nabla_i f(x), h^i \rangle + \frac{L_i}{2} \|h^i\|_i^2, \quad \forall i \in [n], h^i \in \mathbb{R}^{N_i}, x \in \mathbb{R}^N. \quad (40)$$

If  $\hat{S}$  is a serial sampling, then

$$\begin{aligned} \mathbb{E}[f(x + h_{[\hat{S}]})] &= \sum_{i=1}^n p_i f(x + \mathbf{U}_i h^i) \stackrel{(40)}{\leq} f(x) + \sum_{i=1}^n p_i \langle \nabla_i f(x), h^i \rangle + \sum_{i=1}^n \frac{p_i L_i}{2} \|h^i\|_i^2 \\ &= f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2} \|h\|_{L_{op}}^2, \end{aligned}$$

which means that  $(f, \hat{S}) \sim ESO(L)$ , where  $L = (L_1, \dots, L_n) \in \mathbb{R}_{++}^n$ . We can now find a sampling  $\hat{S}$  for which the complexity bound (37) is minimized. This leads to the choice:

$$p_i^* \stackrel{\text{def}}{=} \frac{(L_i \|x_*^i - x_0^i\|_i^2)^{\frac{1}{3}}}{\sum_{j=1}^n (L_j \|x_*^j - x_0^j\|_j^2)^{\frac{1}{3}}}, \quad i = 1, \dots, n. \quad (41)$$

The optimal serial sampling given by (41) is not very useful without (at least some) knowledge of  $x_*$ , which is not known. However, note that the formula (41) confirms the intuition that blocks with larger  $L_i$  and larger distance to the optimal block  $\|x_*^i - x_0^i\|_i$  should be picked (and hence updated) more often.

## 5 Efficient implementation

As mentioned in Section 2.2, Algorithm 1 requires full-dimensional operations at each iteration unless  $\theta_k p_i = 1$  for all  $i \in [n]$  and  $k \in \mathbb{N}$ . In this section we provide an equivalent form of Algorithm 1 which is suitable for efficient implementation under some additional assumptions on the computation of the gradient  $\nabla f$ .

## 5.1 Equivalent form

Focusing on the iterates  $x_k, y_k, z_k$  in Algorithm 1 only, the general algorithm can schematically be written as follows:

$$y_k \leftarrow (1 - \theta_k)x_k + \theta_k z_k \quad (42)$$

$$z_{k+1} \leftarrow \begin{cases} \arg \min_{z \in \mathbb{R}^{N_i}} \{ \langle \nabla_i f(y_k), z \rangle + \frac{\theta_k v_i}{2p_i} \|z - z_k^i\|_i^2 + \psi^i(z) \} & i \in S_k \\ z_k^i & i \notin S_k \end{cases} \quad (43)$$

$$x_{k+1} \leftarrow y_k + \theta_k p^{-1} \cdot (z_{k+1} - z_k) \quad (44)$$

Consider the change of variables from  $\{x_k, y_k, z_k\}$  to  $\{z_k, g_k\}$  where

$$g_k = \alpha_k^{-1}(y_k - z_k) \quad (45)$$

and  $\{\alpha^k\}_{k \geq 0}$  is a sequence defined by:

$$\alpha_0 = 1, \quad \alpha_k = (1 - \theta_k)\alpha_{k-1}, \quad \forall k \geq 1. \quad (46)$$

Note that, in all the special cases presented in Section 4 and 6, either  $\theta_k < 1$  for all  $k \geq 1$  or  $\theta_k = 1$  for all  $k \geq 1$ . The latter case does not require the full-dimensional operation  $y_k = (1 - \theta_k)x_k + \theta_k z_k$  because the three sequences equal to each other. We thus only address the case when  $\theta_k < 1$  for all  $k \geq 1$  which implies that  $\alpha_k \neq 0$  for all  $k \geq 1$ .

Then from  $\{z_k, g_k\}$  and  $\{\alpha_k\}$  we can recover  $\{x_k, y_k\}$  as follows:

$$y_k \stackrel{(45)}{=} z_k + \alpha_k g_k, \quad x_{k+1} \stackrel{(44)+(45)}{=} (z_k + \alpha_k g_k) + \theta_k p^{-1} \cdot (z_{k+1} - z_k). \quad (47)$$

Moreover,  $g_{k+1}$  can be computed recursively as follows:

$$\begin{aligned} g_{k+1} &\stackrel{(45)}{=} \alpha_{k+1}^{-1}(y_{k+1} - z_{k+1}) \stackrel{(42)}{=} \alpha_{k+1}^{-1}(1 - \theta_{k+1})(x_{k+1} - z_{k+1}) \\ &\stackrel{(46)}{=} \alpha_k^{-1}(x_{k+1} - z_{k+1}) \stackrel{(47)}{=} g_k - \alpha_k^{-1}(e - \theta_k p^{-1}) \cdot (z_{k+1} - z_k), \end{aligned}$$

where  $e \in \mathbb{R}^n$  is the vector of all ones. Therefore the updating scheme (42)–(44) can thus be written in the form:

$$z_{k+1} \leftarrow \begin{cases} \arg \min_{z \in \mathbb{R}^{N_i}} \{ \langle \nabla_i f(\alpha_k g_k + z_k), z \rangle + \frac{\theta_k v_i}{2p_i} \|z - z_k^i\|_i^2 + \psi^i(z) \} & i \in S_k \\ z_k^i & i \notin S_k \end{cases} \quad (48)$$

$$g_{k+1} \leftarrow g_k - \alpha_k^{-1}(e - \theta_k p^{-1}) \cdot (z_{k+1} - z_k) \quad (49)$$

$$\alpha_{k+1} \leftarrow (1 - \theta_{k+1})\alpha_k \quad (50)$$

Hence Algorithm 1 can be written in the following equivalent form.

---

**Algorithm 7** Efficient equivalent of Algorithm 1
 

---

- 1: **Parameters:** proper sampling  $\hat{S}$  with probability vector  $p = (p_1, \dots, p_n)$ ,  $v \in \mathbb{R}_{++}^n$ , sequence  $\{\theta_k\}_{k \geq 0}$
  - 2: **Initialization:** choose  $x_0 \in \text{dom } \psi$ , set  $z_0 = x_0$ ,  $g_0 = 0$  and  $\alpha_0 = 1$
  - 3: **for**  $k \geq 0$  **do**
  - 4:   Generate a random set of blocks  $S_k \sim \hat{S}$
  - 5:    $z_{k+1} \leftarrow z_k$ ,  $g_{k+1} \leftarrow g_k$
  - 6:   **for**  $i \in S_k$  **do**
  - 7:      $t_k^i = \arg \min_{t \in \mathbb{R}^{N_i}} \left\{ \langle \nabla_i f(\alpha_k g_k + z_k), t \rangle + \frac{\theta_k v_i}{2 p_i} \|t\|_i^2 + \psi^i(z_k^i + t) \right\}$
  - 8:      $z_{k+1}^i \leftarrow z_k^i + t_k^i$
  - 9:      $g_{k+1}^i \leftarrow g_k^i - \alpha_k^{-1} (1 - \theta_k p_i^{-1}) t_k^i$
  - 10:     $\alpha_{k+1} = (1 - \theta_{k+1}) \alpha_k$
  - 11:   **end for**
  - 12: **end for**
  - 13: **OUTPUT:**  $x_{k+1} = z_k + \alpha_k g_k + \theta_k p^{-1} \cdot (z_{k+1} - z_k)$
- 

## 5.2 Cost of a single iteration

In order to perform Step 7, it is important that we have access to  $\nabla_i f(y_k) = \nabla_i f(\alpha_k g_k + z_k)$  *without* actually computing  $y_k$ . In [6], the authors show that this is possible for problems (1) where  $f$  can be written as:

$$f(x) = \sum_{j=1}^m \phi_j(e_j^\top \mathbf{A}x), \quad \forall x \in \mathbb{R}^N, \quad (51)$$

for some matrix  $\mathbf{A} \in \mathbb{R}^{m \times N}$ . Let us write

$$u_k = \mathbf{A}g_k, \quad w_k = \mathbf{A}z_k, \quad k \geq 1.$$

For  $i \in [n]$  and  $j \in [m]$  denote by  $\mathbf{A}_{ji}$  the  $i$ th block of the  $j$ th row vector of the matrix  $\mathbf{A}$ , i.e.,  $\mathbf{A}_{ji} = U_i^\top \mathbf{A}^\top e_j$ . For each  $i \in [n]$ , denote by  $I_i$  the number of rows containing a non-zero  $i$ th block, i.e.,

$$I_i \stackrel{\text{def}}{=} \{j \in [m] : \mathbf{A}_{ji} \neq 0\}.$$

Then for  $f$  taking the form of (51) we have

$$\nabla_i f(y_k) = \nabla_i f(\alpha_k g_k + z_k) = \sum_{j=1}^m \mathbf{A}_{ji} \phi_j'(\alpha_k u_k^j + w_k^j) = \sum_{j \in I_i} \mathbf{A}_{ji} \phi_j'(\alpha_k u_k^j + w_k^j),$$

where by abuse of notation  $u_k^j$  and  $w_k^j$  denote respectively the  $j$ th element of the vectors  $u_k$  and  $w_k$ . With the knowledge of the vectors  $u_k$  and  $w_k$ , computing  $\nabla_i f(y_k)$  requires  $O(|I_i N_i|)$  operations. Now in order to keep record of the vectors  $u_k$  and  $w_k$ , we use the following equality:

$$w_{k+1} = \mathbf{A}z_{k+1} = \mathbf{A}z_k + \mathbf{A}(z_{k+1} - z_k) = w_k + \sum_{i \in S_k} \mathbf{A}U_i t_k^i, \quad (52)$$

and

$$u_{k+1} = \mathbf{A}g_{k+1} = \mathbf{A}g_k + \mathbf{A}(g_{k+1} - g_k) = u_k + \sum_{i \in S_k} \alpha_k^{-1}(1 - \theta_k p_i^{-1}) \mathbf{A} \mathbf{U}_i t_k^i. \quad (53)$$

Since  $\mathbf{A} \mathbf{U}_i$  is a matrix with  $|I_i N_i|$  nonzero elements, the updating schemes (52) and (53) require then

$$\sum_{i \in S_k} O(I_i N_i)$$

operations. Note that this is also the total complexity of gradient computation  $\nabla_i f(y_k)$  at  $k$ th iteration. Denote by  $\text{nnz}(\mathbf{A})$  the total number of nonzero blocks of the matrix  $\mathbf{A}$ , i.e.,

$$\text{nnz}(\mathbf{A}) \stackrel{\text{def}}{=} \sum_{i=1}^n I_i.$$

Let us consider the special case when  $|\hat{S}| = \tau$  and  $N_i = N/n$  for all  $i \in [n]$ . In this case, the expected one iteration computational complexity is:

$$\mathbb{E} \left[ \sum_{i \in S_k} O \left( I_i \frac{N}{n} \right) \right] = O \left( \sum_{i=1}^n \frac{\tau N}{n^2} I_i \right) = O \left( \frac{\tau N \text{nnz}(\mathbf{A})}{n^2} \right).$$

To make it more direct to understand, let us consider the case when each block contains only one coordinate, i.e.,  $N = n$ . Then the latter expected one iteration complexity becomes

$$\mathbb{E} \left[ \sum_{i \in S_k} O \left( I_i \frac{N}{n} \right) \right] = O \left( \frac{\tau \text{nnz}(\mathbf{A})}{n} \right).$$

Hence, in this case the one iteration complexity in expectation of Algorithm 7 is of order  $O(\tau \bar{\omega})$  where  $\bar{\omega}$  is the average number of nonzero elements of the columns of  $\mathbf{A}$ . Not considering the time spent on synchronization and handling read/write conflicts, the average processing time would be  $O(\bar{\omega})$  if we use a parallel implementation with  $\tau$  processors.

## 6 Proximal minimization

In this section we present and prove complexity results for ALPHA (Algorithm 1) as applied to the general problem (1) involving the proximal term. We leave the discussion concerning special cases to the appendix.

### 6.1 Complexity results

In the presence of the proximal term  $\psi$ , the same complexity bounds as those given in Theorems 3.1 and 3.2 hold for the output of Algorithm 1, with the exception that  $\theta_0$  is only allowed to be chosen between  $(0, \min_i p_i]$ . We now state the formal complexity theorems, first in the non-accelerated and then in the accelerated case.

**Theorem 6.1** (ALPHA – proximal & nonaccelerated). *Let  $\hat{S}$  be arbitrary proper sampling and  $v \in \mathbb{R}_{++}^n$  be such that  $(f, \hat{S}) \sim \text{ESO}(v)$ . Choose  $\theta_k = \theta_0 \in (0, \min_i p_i]$  for all  $k \geq 0$ . Then for any  $y \in \mathbb{R}^N$ , the iterates  $\{x_k\}_{k \geq 1}$  of Algorithm 1 satisfy:*

$$\max \left\{ \mathbb{E}[F(\hat{x}_k)], \min_{l=1, \dots, k} \mathbb{E}[F(x_l)] \right\} - F(y) \leq \frac{C}{(k-1)\theta_0 + 1}, \forall k \geq 1 \quad (54)$$

where

$$\hat{x}_k = \frac{x_k + \theta_0 \sum_{l=1}^{k-1} x_l}{1 + (k-1)\theta_0}$$

and

$$C = (1 - \theta_0) (F(x_0) - F(y)) + \frac{\theta_0^2}{2} \|x_0 - y\|_{v \circ p}^2.$$

**Theorem 6.2** (ALPHA – proximal & accelerated). *Let  $\hat{S}$  be arbitrary proper sampling and  $v \in \mathbb{R}_{++}^n$  be such that  $(f, \hat{S}) \sim \text{ESO}(v)$ . Choose  $\theta_0 \in (0, \min_i p_i]$  and define the sequence  $\{\theta_k\}_{k \geq 0}$  by*

$$\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2 - \theta_k^2}}{2}.$$

Then for any  $y \in \mathbb{R}^N$  such that  $C \geq 0$ , the iterates  $\{x_k\}_{k \geq 1}$  of Algorithm 1 satisfy:

$$\mathbb{E}[F(x_k)] - F(y) \leq \frac{4C}{((k-1)\theta_0 + 2)^2}, \quad (55)$$

where

$$C = (1 - \theta_0) (F(x_0) - F(y)) + \frac{\theta_0^2}{2} \|x_0 - y\|_{v \circ p}^2.$$

In the remainder of the section we will provide the complexity analysis.

Our approach is similar to that presented in [6], but with many modifications required because we allow for an arbitrary sampling. We begin with some technical lemmas.

## 6.2 Technical lemmas

Lemma 6.1 shows that each individual block  $x_k^i$  of the variable  $x_k$  is a convex combination of all the history blocks  $z_0^i, \dots, z_k^i$ . Note that due to the importance sampling, the combination coefficients  $\gamma_{k,0}^i, \dots, \gamma_{k,k}^i$  is now block-dependent, in contrast with the block-independent coefficients proved in [6].

**Lemma 6.1.** *Let  $\{x_k, z_k\}_{k \geq 0}$  be the iterates of Algorithm 1. Then for all  $k \in \mathbb{N}$  and  $i \in [n]$  we have*

$$x_k^i = \sum_{l=0}^k \gamma_{k,l}^i z_l^i, \quad (56)$$

where for each  $i$ , the coefficients  $\{\gamma_{k,l}^i\}_{l=0, \dots, k}$  are defined recursively by setting  $\gamma_{0,0}^i = 1$ ,  $\gamma_{1,0}^i = 1 - \theta_0 p_i^{-1}$ ,  $\gamma_{1,1}^i = \theta_0 p_i^{-1}$  and for  $k \geq 1$ ,

$$\gamma_{k+1,l}^i = \begin{cases} (1 - \theta_k) \gamma_{k,l}^i & l = 0, \dots, k-1 \\ (1 - \theta_k) \gamma_{k,k}^i + \theta_k - \theta_k p_i^{-1} & l = k \\ \theta_k p_i^{-1} & l = k+1 \end{cases} \quad (57)$$

so that the following identity holds,

$$\gamma_{k+1,k}^i + \gamma_{k+1,k+1}^i = (1 - \theta_k)\gamma_{k,k}^i + \theta_k, \quad \forall k \in \mathbb{N}, \quad i \in [n]. \quad (58)$$

Moreover, if  $\theta_0 \in (0, \min_i p_i]$  and  $\{\theta_k\}_{k \geq 0}$  is a decreasing positive sequence, then for all  $k \in \mathbb{N}$  and  $i \in [n]$ , the coefficients  $\{\gamma_{k,l}^i\}_{l=0,\dots,k}$  are all positive and sum to 1.

*Proof.* Fix any  $i \in [n]$ . We proceed by induction on  $k$ . It is clear from  $x_0 = z_0$  that  $\gamma_{0,0}^i = 1$ . Since  $x_1 = y_0 + \theta_0 p^{-1} \cdot (z_1 - z_0)$  and  $y_0 = x_0$ , we get that  $x_1^i = (1 - \theta_0 p_i^{-1})z_0^i + \theta_0 p_i^{-1}z_1^i$  thus  $\gamma_{1,0}^i = 1 - \theta_0 p_i^{-1}$  and  $\gamma_{1,1}^i = \theta_0 p_i^{-1}$ . Assuming that (56) holds for some  $k \geq 1$ , then

$$\begin{aligned} x_{k+1}^i &= y_k^i + \theta_k p_i^{-1}(z_{k+1}^i - z_k^i) = (1 - \theta_k)x_k^i + \theta_k z_k^i - \theta_k p_i^{-1}z_k^i + \theta_k p_i^{-1}z_{k+1}^i \\ &\stackrel{(56)}{=} (1 - \theta_k) \sum_{l=0}^k \gamma_{k,l}^i z_l^i + \theta_k z_k^i - \theta_k p_i^{-1}z_k^i + \theta_k p_i^{-1}z_{k+1}^i \\ &= \sum_{l=0}^{k-1} (1 - \theta_k)\gamma_{k,l}^i z_l^i + ((1 - \theta_k)\gamma_{k,k}^i + \theta_k - \theta_k p_i^{-1})z_k^i + \theta_k p_i^{-1}z_{k+1}^i. \end{aligned}$$

Therefore the recursive equation (57) holds. The identity (58) can then be verified by direct substitution. Next we assume that  $\theta_0 \in (0, \min_i p_i]$  and  $\{\theta_k\}_{k \geq 0}$  is a decreasing positive sequence and show that the linear combination in (56) is a convex combination. Let  $k \geq 1$ . Since  $\theta_k \leq 1$ , we deduce from (57) that  $\{\gamma_{k+1,l}^i\}_{l=0,\dots,k-1}$  are positive if  $\{\gamma_{k,l}^i\}_{l=0,\dots,k-1}$  are positive. Moreover,

$$\begin{aligned} \gamma_{k+1,k}^i &\stackrel{(57)}{=} (1 - \theta_k)\gamma_{k,k}^i + \theta_k - \theta_k p_i^{-1} = \theta_k(1 - \gamma_{k,k}^i) + \gamma_{k,k}^i - \theta_k p_i^{-1} \\ &\stackrel{(57)}{=} \theta_k(1 - \gamma_{k,k}^i) + (\theta_{k-1} - \theta_k)p_i^{-1} \geq \theta_k(1 - \gamma_{k,k}^i). \end{aligned}$$

Then using  $\theta_k \geq 0$ , we conclude that  $\gamma_{k+1,k}^i \geq 0$  if  $\gamma_{k,k}^i \leq 1$ . Besides, we have:

$$\begin{aligned} \sum_{l=0}^{k+1} \gamma_{k+1,l}^i &= \sum_{l=0}^{k-1} \gamma_{k+1,l}^i + \gamma_{k+1,k}^i + \gamma_{k+1,k+1}^i \\ &\stackrel{(57)}{=} (1 - \theta_k) \sum_{l=0}^{k-1} \gamma_{k,l}^i + (1 - \theta_k)\gamma_{k,k}^i + \theta_k - \theta_k p_i^{-1} + \theta_k p_i^{-1} = (1 - \theta_k) \sum_{l=0}^k \gamma_{k,l}^i + \theta_k. \end{aligned}$$

We deduce from the above facts that the coefficients  $\{\gamma_{k+1,l}^i\}_{l=0,\dots,k+1}$  are all positive and sum to 1 if the same holds for  $\{\gamma_{k,l}^i\}_{l=0,\dots,k}$ . Since  $\theta_0 \leq \min_i p_i$ , we know that  $\{\gamma_{1,0}^i, \gamma_{1,1}^i\}$  are positive and sum to 1. It follows that the same property holds for all  $k \in \mathbb{N}$ .  $\square$

**Lemma 6.2.** For  $k \in \mathbb{N}$  and  $i \in [n]$ , define

$$\hat{\psi}_k^i \stackrel{\text{def}}{=} \sum_{l=0}^k \gamma_{k,l}^i \psi^i(z_l^i). \quad (59)$$

Moreover,

$$\mathbb{E}_k[\hat{\psi}_{k+1}^i] = (1 - \theta_k)\hat{\psi}_k^i + \theta_k \psi^i(\tilde{z}_{k+1}^i), \quad \forall k \in \mathbb{N}, \quad i \in [n]. \quad (60)$$

*Proof.*

$$\begin{aligned}
\mathbb{E}_k[\hat{\psi}_{k+1}^i] &= \sum_{l=0}^k [\gamma_{k+1,l}^i \psi^i(z_l^i)] + \theta_k p_i^{-1} \mathbb{E}_k[\psi^i(z_{k+1}^i)] \\
&\stackrel{(10)}{=} \sum_{l=0}^k [\gamma_{k+1,l}^i \psi^i(z_l^i)] + \theta_k p_i^{-1} ((1-p_i)\psi^i(z_k^i) + p_i \psi^i(\tilde{z}_{k+1}^i)) \\
&= \sum_{l=0}^k [\gamma_{k+1,l}^i \psi^i(z_l^i)] + (p_i^{-1} - 1)\theta_k \psi^i(z_k^i) + \theta_k \psi^i(\tilde{z}_{k+1}^i) \\
&\stackrel{(57)}{=} (1-\theta_k) \sum_{l=0}^{k-1} [\gamma_{k,l}^i \psi^i(z_l^i)] + (\gamma_{k+1,k}^i + (p_i^{-1} - 1)\theta_k) \psi^i(z_k^i) + \theta_k \psi^i(\tilde{z}_{k+1}^i) \\
&\stackrel{(57)}{=} (1-\theta_k) \sum_{l=0}^{k-1} [\gamma_{k,l}^i \psi^i(z_l^i)] + (\gamma_{k+1,k}^i + \gamma_{k+1,k+1}^i - \theta_k) \psi^i(z_k^i) + \theta_k \psi^i(\tilde{z}_{k+1}^i) \\
&\stackrel{(58)}{=} (1-\theta_k) \sum_{l=0}^k [\gamma_{k,l}^i \psi^i(z_l^i)] + \theta_k \psi^i(\tilde{z}_{k+1}^i) \\
&\stackrel{(59)}{=} (1-\theta_k) \hat{\psi}_k^i + \theta_k \psi^i(\tilde{z}_{k+1}^i).
\end{aligned}$$

□

The next result was previously stated and used in [6].

**Lemma 6.3** ([2, 41]). *Let*

$$\xi(z) \stackrel{\text{def}}{=} f(y_k) + \langle \nabla f(y_k), z - y_k \rangle + \frac{\theta_k}{2} \|z - z_k\|_{p^{-1} \circ v}^2, \quad z \in \mathbb{R}^N.$$

*Then, for  $\tilde{z}_{k+1}$  defined in (9) we have:*

$$\psi(\tilde{z}_{k+1}) + \xi(\tilde{z}_{k+1}) \leq \psi(y) + \xi(y) - \frac{\theta_k}{2} \|\tilde{z}_{k+1} - y\|_{p^{-1} \circ v}^2, \quad y \in \mathbb{R}^N. \quad (61)$$

For a proof, see for example [2, Lemma 3.2].

### 6.3 Recursion

For all  $k \geq 0$  define:

$$\hat{\psi}_k \stackrel{\text{def}}{=} \sum_{i=1}^n \hat{\psi}_k^i, \quad \hat{F}_k \stackrel{\text{def}}{=} \hat{\psi}_k + f(x_k). \quad (62)$$

We next prove an inequality similar to the one we established in the smooth case (20).

**Lemma 6.4.** *Let  $\hat{S}$  be an arbitrary proper sampling and  $v \in \mathbb{R}_{++}^n$  be such that  $(f, \hat{S}) \sim \text{ESO}(v)$ . Let  $\{\theta_k\}_{k \geq 0}$  be arbitrary sequence of positive numbers in  $(0, 1]$  and fix  $y \in \mathbb{R}^N$ . Then for the sequence of iterates produced by Algorithm 1 and all  $k \geq 0$ , the following recursion holds:*

$$\mathbb{E}_k \left[ \hat{F}_{k+1} + \frac{\theta_k^2}{2} \|z_{k+1} - y\|_{v \circ p^{-2}}^2 \right] \leq \left[ \hat{F}_k + \frac{\theta_k^2}{2} \|z_k - y\|_{v \circ p^{-2}}^2 \right] - \theta_k (\hat{F}_k - F(y)). \quad (63)$$



*Proof.* If  $(f, \hat{S}) \sim ESO(v)$ , then,

$$\begin{aligned}
\mathbb{E}_k[f(x_{k+1})] &\stackrel{(11)}{=} \mathbb{E}_k[f(y_k + \theta_k p^{-1} \cdot (\tilde{z}_{k+1} - z_k)_{[S_k]})] \\
&\stackrel{(13)}{\leq} f(y_k) + \theta_k \langle \nabla f(y_k), \tilde{z}_{k+1} - z_k \rangle + \frac{\theta_k^2}{2} \|\tilde{z}_{k+1} - z_k\|_{p^{-1}ov}^2 \\
&= (1 - \theta_k) f(y_k) - \theta_k \langle \nabla f(y_k), z_k - y_k \rangle \\
&\quad + \theta_k (f(y_k) + \langle \nabla f(y_k), \tilde{z}_{k+1} - y_k \rangle + \frac{\theta_k}{2} \|\tilde{z}_{k+1} - z_k\|_{p^{-1}ov}^2) \\
&\stackrel{(12)}{=} (1 - \theta_k) (f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle) \\
&\quad + \theta_k (f(y_k) + \langle \nabla f(y_k), \tilde{z}_{k+1} - y_k \rangle + \frac{\theta_k}{2} \|\tilde{z}_{k+1} - z_k\|_{p^{-1}ov}^2). \tag{64}
\end{aligned}$$

We first write

$$\begin{aligned}
\mathbb{E}_k[\hat{F}_{k+1}] &= \mathbb{E}_k\left[\sum_{i=1}^n \hat{\psi}_{k+1}^i + f(x_{k+1})\right] \stackrel{(60)}{=} \sum_{i=1}^n \left[(1 - \theta_k) \hat{\psi}_k^i + \theta_k \psi^i(z_{k+1}^i)\right] + \mathbb{E}_k[f(x_{k+1})] \\
&= (1 - \theta_k) \hat{\psi}_k + \theta_k \psi(\tilde{z}_{k+1}) + \mathbb{E}_k[f(x_{k+1})], \tag{65}
\end{aligned}$$

and then bound the expectation of  $\hat{F}_{k+1}$  as follows:

$$\begin{aligned}
&\mathbb{E}_k[\hat{F}_{k+1}] \\
&\stackrel{(65)+(64)}{\leq} (1 - \theta_k) \hat{\psi}_k + (1 - \theta_k) (f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle) \\
&\quad + \theta_k (\psi(\tilde{z}_{k+1}) + f(y_k) + \langle \nabla f(y_k), \tilde{z}_{k+1} - y_k \rangle + \frac{\theta_k}{2} \|\tilde{z}_{k+1} - z_k\|_{p^{-1}ov}^2) \\
&\stackrel{(61)}{\leq} (1 - \theta_k) \hat{\psi}_k + (1 - \theta_k) (f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle) \\
&\quad + \theta_k (\psi(y) + f(y_k) + \langle \nabla f(y_k), y - y_k \rangle + \frac{\theta_k}{2} \|z_k - y\|_{p^{-1}ov}^2 - \frac{\theta_k}{2} \|\tilde{z}_{k+1} - y\|_{p^{-1}ov}^2) \\
&\leq (1 - \theta_k) \hat{\psi}_k + (1 - \theta_k) f(x_k) \\
&\quad + \theta_k (\psi(y) + f(y) + \frac{\theta_k}{2} \|z_k - y\|_{p^{-1}ov}^2 - \frac{\theta_k}{2} \|\tilde{z}_{k+1} - y\|_{p^{-1}ov}^2) \\
&= (1 - \theta_k) \hat{F}_k + \theta_k F(y) + \frac{\theta_k^2}{2} (\|z_k - y\|_{p^{-1}ov}^2 - \|\tilde{z}_{k+1} - y\|_{p^{-1}ov}^2) \\
&\stackrel{(10)}{=} (1 - \theta_k) \hat{F}_k + \theta_k F(y) + \frac{\theta_k^2}{2} \mathbb{E}_k [\|z_k - y\|_{p^{-2}ov}^2 - \|y - z_{k+1}\|_{p^{-2}ov}^2], \quad \forall y \in \mathbb{R}^N.
\end{aligned}$$

Therefore, for all  $y \in \mathbb{R}^N$ ,

$$\mathbb{E}_k \left[ \hat{F}_{k+1} - F(y) + \frac{\theta_k^2}{2} \|y - z_{k+1}\|_{p^{-2}ov}^2 \right] \leq (1 - \theta_k) (\hat{F}_k - F(y)) + \frac{\theta_k^2}{2} \|z_k - y\|_{p^{-2}ov}^2. \quad \square$$

## 6.4 Proof of Theorems 6.1 and 6.2

Using the same reasoning as that in the proof of Theorems 3.1 and 3.2, we analyze recursion of Lemma 6.4 and obtain:

- If  $\theta_k = \theta_0 \in (0, 1]$  for all  $k \geq 0$ , then

$$\frac{\mathbb{E} \left[ \hat{F}_k - F(y) + \theta_0 \sum_{l=1}^{k-1} (\hat{F}_l - F(y)) \right]}{1 + (k-1)\theta_0} \leq \frac{(1 - \theta_0)(\hat{F}_0 - F(y)) + \frac{\theta_0^2}{2} \|x_0 - y\|_{vop^{-2}}}{1 + \theta_0(k-1)}, \quad \forall k \geq 1.$$

- If  $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$  for all  $k \geq 0$ ,  $\theta_0 \in (0, 1]$  and  $\hat{F}_0 \leq F(y)$ , then

$$\mathbb{E} \left[ \hat{F}_{k+1} - F(y) \right] \leq \frac{4 \left( (1 - \theta_0)(\hat{F}_0 - F(y)) + \frac{\theta_0^2}{2} \|x_0 - y\|_{vop^{-2}} \right)}{(\theta_0(k-1) + 2)^2}, \quad \forall k \geq 1. \quad (66)$$

Finally, by Lemma 6.1, for the latter two choices of  $\{\theta_k\}$ , if in addition  $\theta_0 \in (0, \min_i p_i]$ , then for  $k \geq 1$ , each block of the vector  $x_k$  is a convex combination of the corresponding blocks of the vectors  $z_0, \dots, z_k$ . By the convexity of each function  $\psi^i$ , we get:

$$\psi(x_k) = \sum_{i=1}^n \psi^i(x_k^i) = \sum_{i=1}^n \psi^i \left( \sum_{l=0}^k \gamma_{k,l}^i z_l^i \right) \leq \sum_{i=1}^n \sum_{l=0}^k \gamma_{k,l}^i \psi^i(z_l^i) = \hat{\psi}_k. \quad (67)$$

Hence, Theorem 6.1 and 6.2 hold by the fact that  $F(x_k) \leq \hat{F}_k$  for all  $k \in \mathbb{N}$  and  $\hat{F}_0 = F(x_0)$ . Note that the condition  $\theta_0 \in (0, \min_i p_i]$  is only needed to prove (67). Thus it can be relaxed to  $\theta_0 \in (0, 1]$  if  $\psi \equiv 0$ , in which case  $\hat{F}_k = F(x_k)$  and Theorem 3.1 and 3.2 follows.

## 7 Conclusion

In this paper we propose a general randomized coordinate descent method which can be specialized to serial or parallel and accelerated or non-accelerated variants, with or without importance sampling. Based on the technical assumption which captures in a compact way certain smoothness properties of the function in a random subspace spanned by the sampled coordinates, we provide a unified complexity analysis which allows to derive as direct corollary the convergence results for the multiple variants of the general algorithm. We focused on the minimization of non-strongly convex function. Further study on a unified algorithm and complexity analysis for both strongly and non-strongly convex objective functions can be investigated.

## References

- [1] Alfred Auslender and Marc Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.
- [2] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.
- [3] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014.

- [4] Michael Elad, Boaz Matalon, and Michael Zibulevsky. Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization. *Applied and Computational Harmonic Analysis*, pages 346–367, 2007.
- [5] Olivier Fercoq, Zheng Qu, Peter Richtárik, and Martin Takáč. Fast distributed coordinate descent for minimizing non-strongly convex losses. *IEEE International Workshop on Machine Learning for Signal Processing*, 2014.
- [6] Olivier Fercoq and Peter Richtárik. Accelerated, parallel and proximal coordinate descent. *SIAM Journal on Optimization (after minor revision)*, *arXiv:1312.5799*, 2013.
- [7] Olivier Fercoq and Peter Richtárik. Smooth minimization of nonsmooth functions by parallel coordinate descent. *arXiv:1309.5885*, 2013.
- [8] C-J. Hsieh, K-W. Chang, C-J. Lin, S.S. Keerthi, and S. Sundarajan. A dual coordinate descent method for large-scale linear SVM. In *ICML*, 2008.
- [9] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.
- [10] Jakub Konečný, Jie Lu, Peter Richtárik, and Martin Takáč. mS2GD: Mini-batch semi-stochastic gradient descent in the proximal setting. *arXiv:1410.4744*, 2014.
- [11] Jakub Konečný, Zheng Qu, and Peter Richtárik. Semi-stochastic coordinate descent. *arXiv:1412.6293*, 2014.
- [12] Jakub Konečný and Peter Richtárik. S2GD: Semi-stochastic gradient descent methods. *arXiv:1312.1666*, 2014.
- [13] Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. *arXiv:1305.1922*, 2013.
- [14] Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated proximal coordinate gradient method and its application to regularized empirical risk minimization. Technical Report MSR-TR-2014-94, July 2014.
- [15] Ji Liu and Stephen J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.
- [16] Ji Liu, Stephen J. Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *Journal of Machine Learning Research*, 16:285–322, 2015.
- [17] Zhaosong Lu and Lin Xiao. On the complexity analysis of randomized block-coordinate descent methods. *arXiv:1305.4723*, 2013.
- [18] Mehrdad Mahdavi and Rong Jin. Mixedgrad: An  $o(1/t)$  convergence rate algorithm for stochastic smooth optimization. *arXiv:1307.7192v1*, 2013.
- [19] Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

- [20] Ion Necoara and Dragos Clipici. Efficient parallel coordinate descent algorithm for convex optimization problems with separable constraints: Application to distributed {MPC}. *Journal of Process Control*, 23(3):243 – 253, 2013.
- [21] Ion Necoara and Andrei Patrascu. A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints. *Computational Optimization and Applications*, 57:307–337, 2014.
- [22] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- [23] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Springer Netherlands, 1 edition.
- [24] Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.
- [25] Zheng Qu and Peter Richtárik. Coordinate descent methods with arbitrary sampling II: Expected separable overapproximation. *Technical report*, 2014.
- [26] Zheng Qu, Peter Richtárik, and Tong Zhang. Randomized dual coordinate ascent with arbitrary sampling. *arXiv:1411.5873*, 2014.
- [27] Peter Richtárik. Randomized coordinate descent for big data optimization (theory). Technical report, School of Mathematics, University of Edinburgh, June 2014.
- [28] Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. *arXiv:1310.2059*, 2013.
- [29] Peter Richtárik and Martin Takáč. On optimal probabilities in stochastic coordinate descent methods. *arXiv:1310.3438*, 2013.
- [30] Peter Richtárik and Martin Takáč. Efficient serial and parallel coordinate descent method for huge-scale truss topology design. In *Operations Research Proceedings*, pages 27–32. Springer, 2012.
- [31] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(2):1–38, 2014.
- [32] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, pages 1–52, 2015.
- [33] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *arXiv:1309.2388*, 2013.
- [34] Shai Shalev-Shwartz, Yoram Singer, Nati Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, pages 3–30, 2011.

- [35] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for  $l_1$ -regularized loss minimization. *J. Mach. Learn. Res.*, 12:1865–1892, July 2011.
- [36] Shai Shalev-Shwartz and Tong Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems 26*, pages 378–385. 2013.
- [37] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, pages 1–41, 2014.
- [38] Martin Takáč, Avleen Bijral, Peter Richtárik, and Nathan Srebro. Mini-batch primal and dual methods for SVMs. In *ICML*, 2013.
- [39] Rachael Tappenden, Peter Richtárik, and Burak Büke. Separable approximations and decomposition methods for the augmented lagrangian. *Optimization Methods and Software*, 2014.
- [40] Rachael Tappenden, Peter Richtárik, and Jacek Gondzio. Inexact block coordinate descent method: complexity and preconditioning. *arXiv:1304.5530*, 2013.
- [41] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Submitted to SIAM Journal on Optimization*, 2008.
- [42] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.
- [43] Lijun Zhang, Mehrdad Mahdavi, and Rong Jin. Linear convergence with condition number independent access of full gradients. *NIPS*, 2013.
- [44] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML*, 2004.
- [45] Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *ICML*, 2015.
- [46] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling. *ICML*, 2015.

## A Special cases of ALPHA in the proximal setup

Extending the discussion presented in Section 4 which focused on the smooth case, we now present four special cases of Algorithm 1 for solving problem (1).

### A.1 Special case 1: proximal gradient descent

Specializing Algorithm 1 to  $\hat{S} = [n]$  and  $\theta_k = 1$  for all  $k \geq 1$ , we obtain the classical proximal gradient descent algorithm. Note that in this case the three sequences  $\{x_k, y_k, z_k\}_{k \geq 0}$  reduce to one sequence.

---

#### Algorithm 8 Proximal Gradient Descent for solving 1

---

- 1: **Parameters:** vector  $v \in \mathbb{R}_{++}^n$  such that (29) holds
  - 2: **Initialization:** choose  $x_0 \in \text{dom } \psi$
  - 3: **for**  $k \geq 0$  **do**
  - 4:   **for**  $i \in [n]$  **do**
  - 5:      $x_{k+1}^i = \arg \min_{x \in \mathbb{R}^{N_i}} \{ \langle \nabla_i f(x_k), x \rangle + \frac{v_i}{2} \|x - x_k^i\|_i^2 + \psi^i(x) \}$
  - 6:   **end for**
  - 7: **end for**
- 

**Corollary A.1.** *For any optimal solution  $x_*$  of (1), the output of Algorithm 8 for all  $k \geq 1$  satisfies:*

$$F(x_k) - F(x_*) \leq \frac{\|x_0 - x_*\|_v^2}{2k}. \quad (68)$$

In particular, for  $\epsilon > 0$ , if

$$k \geq \frac{\|x_0 - x_*\|_v^2}{2\epsilon}, \quad (69)$$

then  $F(x_k) - F(x_*) \leq \epsilon$ .

The proof follows as a corollary of Theorem 6.1, with additional remark (30) which holds in this special case. The reader can refer to the proof of Corollary 4.1.

Corollary A.1 can be found in classical textbooks on convex optimization, see for example [24].

### A.2 Special case 2: accelerated proximal gradient descent

By choosing  $\hat{S} = [n]$  (with probability 1),  $\theta_0 = 1$  and the sequence  $\{\theta_k\}_{k \geq 0}$  according to (17), ALPHA (Algorithm 1) reduces to accelerated proximal gradient descent.

---

**Algorithm 9** Accelerated proximal gradient descent [1]

---

- 1: **Parameters:** vector  $v \in \mathbb{R}_{++}^n$  such that (29) holds
  - 2: **Initialization:** choose  $x_0 \in \text{dom}(\psi)$ , set  $z_0 = x_0$  and  $\theta_0 = 1$
  - 3: **for**  $k \geq 0$  **do**
  - 4:   **for**  $i \in [n]$  **do**
  - 5:      $z_{k+1}^i = \arg \min_{z \in \mathbb{R}^{N_i}} \{ \langle \nabla_i f((1 - \theta_k)x_k + \theta_k z_k), z \rangle + \frac{\theta_k v_i}{2p_i} \|z - z_k^i\|^2 + \psi^i(z) \}$
  - 6:   **end for**
  - 7:    $x_{k+1} = (1 - \theta_k)x_k + \theta_k z_{k+1}$
  - 8:    $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$
  - 9: **end for**
- 

**Corollary A.2.** *For any optimal solution  $x_*$  of (1), the output of Algorithm 9 for all  $k \geq 1$  satisfies:*

$$F(x_k) - F(x_*) \leq \frac{2\|x_0 - x_*\|_v^2}{(k+1)^2}.$$

In particular, for  $0 < \epsilon < \frac{1}{2}\|x_0 - x_*\|_v^2$ , if

$$k \geq \sqrt{\frac{2\|x_0 - x_*\|_v^2}{\epsilon}} - 1, \tag{70}$$

then  $F(x_k) - F(x_*) \leq \epsilon$ .

As discussed for the unconstrained case in Section 4.2, Algorithm 9 and Corollary A.2 can be attributed to Tseng [41].

### A.3 Special case 3: Parallel Coordinate Descent Method (PCDM)

Let  $\hat{S}$  be a proper uniform sampling (i.e., sampling for which  $p_i = \mathbb{P}(i \in \hat{S}) > 0$  is the same for all  $i \in [n]$ , necessarily equal to  $\mathbb{E}[|\hat{S}|]/n$ ). Furthermore, letting  $\tau = \mathbb{E}[|\hat{S}|]$ , choose  $\theta_k = \tau/n$  for all  $k$ . For this choice of the parameters (30) holds, and hence for Algorithm 1 reduces to the PCDM method of Richtárik and Takáč [32].

---

**Algorithm 10** Parallel Coordinate Descent Method (PCDM) [32]

---

- 1: **Parameters:** uniform proper random sampling  $\hat{S}$ , positive vector  $v \in \mathbb{R}_{++}^n$  such that  $(f, \hat{S}) \sim \text{ESO}(v)$
  - 2: **Initialization:** choose  $x_0 \in \text{dom} \psi$  and set  $\tau = \mathbb{E}[|\hat{S}|]$
  - 3: **for**  $k \geq 0$  **do**
  - 4:   Generate  $S_k \sim \hat{S}$
  - 5:    $x_{k+1} \leftarrow x_k$
  - 6:   **for**  $i \in S_k$  **do**
  - 7:      $x_{k+1}^i = \arg \min_{x \in \mathbb{R}^{N_i}} \{ \langle \nabla_i f(x_k), x \rangle + \frac{v_i \tau}{2\tau} \|x - x_k^i\|^2 + \psi^i(x) \}$
  - 8:   **end for**
  - 9: **end for**
- 

By applying Theorem 6.1 and using the same reasoning as in the proof of Corollary 4.1 we obtain the following *new* convergence result for PCDM.

**Corollary A.3.** For any optimal solution  $x_*$  of (1), the output of Algorithm 10 for all  $k \geq 1$  satisfies:

$$\mathbb{E}[F(x_k)] - F(x_*) \leq \frac{n}{(k-1)\tau + n} \left[ \left(1 - \frac{\tau}{n}\right) (F(x_0) - F(x_*)) + \frac{1}{2} \|x_0 - x_*\|_v^2 \right]. \quad (71)$$

In particular, for  $0 < \epsilon < (1 - \tau/n) (F(x_0) - F(x_*)) + \frac{1}{2} \|x_0 - x_*\|_v^2$ , if

$$k \geq \frac{(n - \tau) (F(x_0) - F(x_*)) + \frac{n}{2} \|x_0 - x_*\|_v^2}{\tau \epsilon} - \frac{n}{\tau} + 1,$$

then  $\mathbb{E}[F(x_k) - F(x_*)] \leq \epsilon$ .

A high-probability result involving the level-set distance

$$\mathcal{R}_v(x_0, x_*) \stackrel{\text{def}}{=} \max_x \{ \|x - x_*\|_v^2 : F(x) \leq F(x_0) \} < +\infty \quad (72)$$

was provided in [32] for PCDM (Algorithm 10). Although not explicitly stated in the paper, it is apparent from the proof that their approach yields the following rate:

$$\mathbb{E}[F(x_k)] - F(x_*) \leq \frac{2n \max\{\mathcal{R}_v(x_0, x_*), F(x_0) - F(x_*)\}}{2n \max\{\mathcal{R}_v(x_0, x_*) / (F(x_k) - F(x_*)), 1\} + \tau k} \quad (73)$$

Since  $\mathcal{R}_v(x_0, x_*) \geq \|x_0 - x_*\|_v^2$ , it is clear that for sufficiently large  $k$ , our rate (71) is better than (73).

#### A.4 Special case 4: APPROX with importance sampling

Finally, let  $\{\theta_k\}_{k \geq 0}$  be chosen in accordance with (17). In this case, ALPHA (Algorithm 1) reduces to an accelerated coordinate descent method with arbitrary sampling  $\hat{S}$  for solving the proximal minimization problem 1.

---

**Algorithm 11** APPROXis (APPROX [6] with importance sampling)

---

- 1: **Parameters:** proper random sampling  $\hat{S}$  with probability vector  $p = (p_1, \dots, p_n)$ ,  $v \in \mathbb{R}_{++}^n$  such that  $(f, \hat{S}) \sim ESO(v)$
  - 2: **Initialization:** choose  $x_0 \in \text{dom}(\psi)$ , set  $z_0 = x_0$  and  $\theta_0 = \min_i p_i$
  - 3: **for**  $k \geq 0$  **do**
  - 4:    $y_k = (1 - \theta_k)x_k + \theta_k z_k$
  - 5:   Generate  $S_k \sim \hat{S}$
  - 6:    $z_{k+1} \leftarrow z_k$
  - 7:   **for**  $i \in S_k$  **do**
  - 8:      $z_{k+1}^i = \arg \min_{z \in \mathbb{R}^{N_i}} \{ \langle \nabla_i f(y_k), z \rangle + \frac{\theta_k v_i}{2p_i} \|z - z_k^i\|_i^2 + \psi^i(z) \}$
  - 9:   **end for**
  - 10:    $x_{k+1} = y_k + \theta_k p^{-1} \cdot (z_{k+1} - z_k)$
  - 11:    $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k}{2}$
  - 12: **end for**
- 

APPROXis is a generalization of APPROX [6] from a uniform sampling to an arbitrary sampling: we recover APPROX if  $\hat{S}$  is uniform with  $\tau = \mathbb{E}[|\hat{S}|]$  and  $\theta_0 = \tau/n$ .



**Corollary A.4.** For any optimal solution  $x_*$  of (1), the output of Algorithm 11 for all  $k \geq 1$  satisfies:

$$\mathbb{E}[F(x_k) - F(x_*)] \leq \frac{4 \left[ \left(1 - \min_i p_i\right) (F(x_0) - F(x_*)) + \min_i \frac{p_i^2}{2} \|x_0 - x_*\|_{\text{vop}-2}^2 \right]}{((k-1) \min_i p_i + 2)^2}.$$

In particular, for  $\epsilon > 0$ , if

$$k \geq \frac{2 \sqrt{\left(1 - \min_i p_i\right) (F(x_0) - F(x_*)) + \min_i \frac{p_i^2}{2} \|x_0 - x_*\|_{\text{vop}-2}^2}}{\min_i p_i \sqrt{\epsilon}} - \frac{2}{\min_i p_i} + 1, \quad (74)$$

then  $\mathbb{E}[F(x_k) - F(x_*)] \leq \epsilon$ .

*Proof.* This is a direct corollary of Theorem 6.2 by taking  $\theta_0 = \min_i p_i$ .  $\square$

If  $\hat{S}$  is a uniform sampling with  $\tau = \mathbb{E}[|\hat{S}|]$  and  $\theta_0 = \tau/n$ , then we recover the convergence result established for APPROX [6, Theorem 3], as well as all the special cases that APPROX can recover, including the fast distributed coordinate descent method Hydra<sup>2</sup> [5].