2023 European Control Conference (ECC)
June 13-16, 2023. Bucharest, Romania

Can random proximal coordinate descent be accelerated on nonseparable convex composite minimization problems?

Flavia Chorobura[1], François Glineur[2] and Ion Necoara[1,3]

***Abstract*— In this paper we consider convex composite optimization problems, where first term is smooth, while the second term is proximal easy but nonseparable (possibly nonsmooth). For this problem we adapt the accelerated proximal coordinate descent algorithm from [7], initially developed for convex composite problems having the second term separable. We study convergence to a coordinate-wise minimizer point, and derive convergence rate in expected function values of order $\mathcal{O}((\mathrm{C} + (\mathbb{E}[S_k^{\#}])_+)/k^2)$. The first term, C, coincides with the usual constant appearing in the rate of accelerated gradient type methods, while the second one, $S_k^{\#}$, measures the nonseparability of the second term in the objective along the iterates. We conjecture that the second term, $S_k^{\#}$, is bounded as this is what we observe in all our numerical simulations and that coordinate descent can be accelerated.**

I. INTRODUCTION

In this paper we study the convergence behavior of an accelerated random proximal (block) coordinate descent method for solving composite optimization problems of the form:

$$F^* = \min_{x \in \mathbb{R}^n} F(x) := f(x) + \psi(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function with a (block) coordinate-wise Lipschitz gradient and $\psi : \mathbb{R}^n \to \bar{\mathbb{R}}$ is convex, proximal easy along cooordinates, nonseparable function (possibly nonsmooth), where $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$. Optimization problems having this composite structure permit to handle coupling functions ψ (e.g., $\|Ax\|^p$, with A linear operator and $p \geq 1$, group sparsity, TV regularizations and indicator functions), and arise in many applications [4], [9], [10].

State of the art (Accelerated) proximal coordinate descent methods have fast convergence and a small cost per iteration, hence they are widely used to solve large-scale problems [11]. The criteria for choosing at each iteration the coordinate over which we minimize a local approximation differs among these methods (e.g., greedy, cyclic or random coordinate search), see also the survey [17]. However, when dealing with composite problems involving a nonsmooth term, the studies [2], [3], [5], [7], [10], [11], [14], [15] require the latter term to be separable, i.e., $\psi(x) = \sum_{i=1}^n \psi_i(x_i)$, where x_i is the ith (block) component of x. In particular, accelerated coordinate descent algorithms for solving the convex optimization problem (1) with $\psi \equiv 0$ were proposed in [2], [11], [14] and sublinear rates of order $\mathcal{O}\left(1/k^2\right)$ were obtained in function values (k denotes the iteration counter). Further, for the composite optimization problem (1) with ψ separable, the standard random proximal coordinate descent achieves a convergence rate of order $\mathcal{O}\left(1/k\right)$ [10], [15], while accelerated variants have rate of order $\mathcal{O}\left(1/k^2\right)$ [7]. For problem (1) with nonseparable function ψ full proximal coordinate descent schemes were proposed recently in [1], [8]. It was obtained in [8] linear rate when the objective function is strongly convex, while, in [1] sublinear rate of order $\mathcal{O}(1/k^{\frac{3}{2}})$ was derived for an accelerated coordinate descent algorithm.

Contributions In this paper we extend the accelerated proximal coordinate descent algorithm proposed in [7], initially developed for convex composite problems having the second term separable, to the case of nonseparable ψ. We study convergence to a coordinate-wise minimizer, and derive convergence rate in expected function values of order $\mathcal{O}((\mathrm{C}+(\mathbb{E}[S_k^{\#}])_+)/k^2)$. Note that a coordinate-wise minimizer point it is not necessarily a solution of the problem (1) (see Definition 1). However, any solution of (1) is a coordinate-wise minimizer. More specifically, our contributions are:
(i) We extend the accelerated proximal coordinate descent algorithm proposed in [7] to convex composite problems having the second term nonseparable. This method does not require the computation of (the block of) the full proximal operator at each iteration. Instead, one needs to compute the proximal operator along a single block of coordinates. Hence, in general, our method has low cost per iteration and can be applied to larger classes of problems than [1], [8].
(ii) We show that the expected difference between function values and the function evaluated at some coordinate-wise minimizer $x^{\#}$, i.e., $\mathbb{E}[F(x_k) - F(x^{\#})]$, is bounded by $\mathcal{O}((\mathrm{C} + (\mathbb{E}[S_k^{\#}])_+)/k^2)$. First term, C, coincides with the usual constant appearing in the rate of accelerated gradient type methods, while the second term, $S_k^{\#}$, measures the nonseparability of the second function ψ along the iterates.
(iii) Although, we cannot prove a bound on $S_k^{\#}$, we conjecture that this quantity is bounded and that the algorithm converges with a rate $\mathcal{O}\left(1/k^2\right)$ to a coordinate-wise minimizer of the nonseparable problem. Indeed, in our preliminary numerical simulations, we observed empirically that $S_k^{\#}$ is bounded and the algorithm converges with a rate $\mathcal{O}\left(1/k^2\right)$.

II. PRELIMINARIES

In this section we present the basic assumptions for composite problem (1), some definitions and preliminary results.

[1]Automatic Control and System Engineering Department, University Politehnica Bucharest, Spl. Independentei, 060042 Bucharest, Romania, flavia.chorobura@stud.acs.upb.ro.
[2]ICTEAM Institute and CORE, Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium, Francois.Glineur@uclouvain.be
[3]Gheorghe Mihoc-Caius Iacob Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, 050711 Bucharest, Romania, ion.necoara@upb.ro.

We consider the following problem settings. Let $U \in \mathbb{R}^{n\times n}$ be a column permutation of the $n \times n$ identity matrix and further decompose it into N submatrices $U = [U_1, ..., U_N]$, with $U_i \in \mathbb{R}^{n\times n_i}$ and $\sum_{i=1}^N n_i = n$. Then, any vector $x \in \mathbb{R}^n$ can be written uniquely as $x = \sum_{i=1}^N U_i x^{(i)}$, where $x^{(i)} = U_i^T x \in \mathbb{R}^{n_i}$. Throughout the paper the following assumptions will be valid.

Assumptions 1: A.1. The gradient of f is (block) coordinate-wise Lipschitz continuous with constants L_i:

$$\|U_i^T(\nabla f(x + U_i h) - \nabla f(x))\| \le L_i \|h\| \tag{2}$$

for all $x \in \mathbb{R}^n, h \in \mathbb{R}^{n_i}$ and $i = 1 : N$.
A.2. A solution exists for (1) (i.e., optimal value $F^* > -\infty$).
A.3. Functions f and ψ are convex.
A.4. Function ψ is simple in the sense that ψ restricted to any block of coordinates (i.e., any subspace generated by $U_i \in \mathbb{R}^{n\times n_i}$) is proximal easy.

If Assumption A1 holds, then we have the relation:

$$f(x + U_i h) - f(x) - \langle U_i^T \nabla f(x), h\rangle \le \frac{L_i}{2}\|h\|^2 \tag{3}$$

for all $x \in \mathbb{R}^n, h \in \mathbb{R}^{n_i}$ and $i = 1 : N$. We consider the following norms: $\|x\|_L^2 = \sum_{i=1}^N L_i \|x^{(i)}\|^2$ and $\|x\|^2 = \sum_{i=1}^N \|x^{(i)}\|^2$. For a given function ψ and a point x, the partial functions along the subspaces generated by U_i are denoted as: $\psi_i^x(d) = \psi(x + U_i d) \quad \forall i = 1 : N$. Let us define a coordinate-wise minimizer for problem (1).

Definition 1: A point $x^\# \in \mathbb{R}^n$ is a coordinate-wise minimizer of (1) if the following holds:

$$F(x^\#) \le F(x^\# + U_i d) \quad \forall d \in \mathbb{R}^{n_i}, \quad \forall i = 1 : N, \tag{4}$$

or equivalently, for all $i = 1 : N$,

$$\exists \quad g_i^\# \in \partial\psi_i^{x^\#}(0) \quad \text{such that} \quad \nabla_i f(x^\#) + g_i^\# = 0. \tag{5}$$

Note that, since $F_i^{x^\#}$ is convex function, (4) holds iff $F_i^{x^\#}(d) \ge F_i^{x^\#}(0) + \langle 0, d\rangle$ or equivalently, $0 \in \partial F_i^{x^\#}(0) = \nabla_i f(x^\#) + \partial\psi_i^{x^\#}(0)$. It is clear that a minimizer of convex problem (1) is always a coordinate-wise minimizer. However, the converse is not always true, unless extra conditions on ψ are satisfied, such as ψ being differentiable or separable, see also [6]. Let us recall a basic result related to the optimality conditions for $\min_{x\in \mathrm{dom}\,\phi} \theta(x) + \phi(x)$, where θ is differentiable function and ϕ is convex function on the convex domain $\mathrm{dom}\,\phi$, i.e., if y_* is a (local) minimizer, we have [12]:

$$\langle \nabla\theta(y_*), y - y_*\rangle + \phi(y) \ge \phi(y_*) \quad \forall y \in \mathrm{dom}\,\phi. \tag{6}$$

III. Algorithm APPROX

In this section, we extend the accelerated proximal coordinate descent algorithm proposed in [7], called APPROX, to the case of nonseparable ψ. Note that in step 4 we do not require the computation of (block of) the full proximal operator of ψ. Instead, we need to compute only the proximal operator along a single block of coordinates. When ψ is separable function, this algorithm is the same as in [7].

Algorithm 1 Algorithm APPROX

Given a starting point $x_0 \in \mathbb{R}^n$. Set $z_0 = x_0$ and $\theta_0 = \frac{1}{N}$.
for $k \ge 0$ **do**
1. Set $y_k = (1 - \theta_k)x_k + \theta_k z_k$
2. Choose index $i_k \in \{1, ..., N\}$ uniformly at random
3. Set $z_{k+1} = z_k$
4. Compute $z_{k+1}^{(i_k)} = \arg\min_{z\in\mathbb{R}^{n_{i_k}}} \langle U_{i_k}^T \nabla f(y_k), z - y_k^{(i_k)}\rangle + \frac{N\theta_k L_{i_k}}{2}\|z - z_k^{(i_k)}\|^2 + \psi(z_k + U_{i_k}(z - z_k^{(i_k)}))$
5. Update: $x_{k+1} = y_k + N\theta_k(z_{k+1} - z_k)$

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

end for

The sequence θ_k satisfies [7]:

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} = \frac{1}{\theta_k^2} \text{ and } \theta_k \le \frac{2}{k + 2N}. \tag{7}$$

Our convergence follows similar lines as in [7]. Let us define:

$$\tilde{z}_{k+1} = \arg\min_{z=(z^{(1)},\cdots,z^{(N)})\in\mathbb{R}^n} \langle \nabla f(y_k), z - y_k\rangle + \frac{N\theta_k}{2}\|z - z_k\|_L^2 + \sum_{i=1}^N \psi(z_k + U_i(z^{(i)} - z_k^{(i)})). \tag{8}$$

Note that, we have: $z_{k+1}^{(i)} = \begin{cases} \tilde{z}_{k+1}^{(i)}, & i = i_k \\ z_k^{(i)}, & i \ne i_k. \end{cases}$

Lemma 1: Given $x^\#$ a coordinate-wise minimizer and

$$\xi(u) := f(y_k) + \langle \nabla f(y_k), u - y_k\rangle + \frac{N\theta_k}{2}\|u - z_k\|_L^2,$$

then the following relation holds:

$$\sum_{i=1}^N \psi(z_k + U_i(\tilde{z}^{(i)} - z_k^{(i)})) + \xi(\tilde{z}_{k+1}) \le \sum_{i=1}^N \psi(z_k + U_i((x^\#)^{(i)} - z_k^{(i)})) + \xi(\hat{x}) - \frac{N\theta_k}{2}\|\hat{x} - \tilde{z}_{k+1}\|_L^2.$$

Proof: From (6), we have for all $z \in \mathbb{R}^n$:

$$\sum_{i=1}^N \psi(z_k + U_i(\tilde{z}^{(i)} - z_k^{(i)})) \le \sum_{i=1}^N \psi(z_k + U_i(z^{(i)} - z_k^{(i)})) + \sum_{i=1}^N \langle U_{i_k}^T \nabla f(y_k) + N\theta_k L_i(\tilde{z}_{k+1}^{(i)} - z_k^{(i)}), z^{(i)} - \tilde{z}_{k+1}^{(i)}\rangle.$$

Choosing $z = x^\#$, we further get:

$$\sum_{i=1}^N \psi(z_k + U_i(\tilde{z}^{(i)} - z_k^{(i)})) + f(y_k) + \langle \nabla f(y_k), \tilde{z}_{k+1} - y_k\rangle + \frac{N\theta_k}{2}\|\tilde{z}_{k+1} - z_k\|_L^2 \le \sum_{i=1}^N \psi(z_k + U_i((x^\#)^{(i)} - z_k^{(i)})) + \frac{N\theta_k}{2}\|\tilde{z}_{k+1} - z_k\|_L^2 + f(y_k) + \langle \nabla f(y_k), x^\# - \tilde{z}_{k+1}\rangle + \langle \nabla f(y_k), \tilde{z}_{k+1} - y_k\rangle + \sum_{i=1}^N N\theta_k L_{i_k}\langle \tilde{z}_{k+1}^{(i)} - z_k^{(i)}, (x^\#)^{(i_k)} - \tilde{z}_{k+1}^{(i_k)}\rangle$$

Hence

$$\sum_{i=1}^{N}\psi(z_k+U_i(\tilde{z}^{(i)}-z_k^{(i)}))+f(y_k)$$
$$+\langle\nabla f(y_k),\tilde{z}_{k+1}-y_k\rangle+\frac{N\theta_k}{2}\|\tilde{z}_{k+1}-z_k\|_L^2$$
$$\leq\sum_{i=1}^{N}\psi(z_k+U_i((x^{\#})^{(i)}-z_k^{(i)}))+\langle\nabla f(y_k),x^{\#}-y_k\rangle$$
$$+f(y_k)+\frac{N\theta_k}{2}\|x^{\#}-z_k\|_L^2-\frac{N\theta_k}{2}\|x^{\#}-\tilde{z}_{k+1}\|_L^2,$$

which proves our statement. ∎

Lemma 2: [7] Iterates of Algorithm 1, (x_k,z_k), satisfy:

$$x_k=\sum_{l=0}^{k}\gamma_k^l z_l, \tag{9}$$

where the coefficients $\gamma_k^0,\gamma_k^1,\cdots,\gamma_k^k$ are nonnegative and sum to one. That is, x_k is a convex combination of the vectors z_0, z_1,$\cdots$,z_k. In particular, the coeficients are defined recursively, setting $\gamma_0^0=1$, $\gamma_1^0=0$, $\gamma_1^1=1$, and for $k\geq 1$:

$$\gamma_{k+1}^l=\begin{cases}(1-\theta_k)\gamma_k^l, & l=0,\cdots,k-1,\\ \theta_k\left(1-N\theta_{k-1}\right)+N\left(\theta_{k-1}-\theta_k\right), & l=k\\ N\theta_k, & l=k+1.\end{cases}$$

Moreover, for $k\geq 0$, the following identity holds:

$$\gamma_{k+1}^k+(N-1)\theta_k=(1-\theta_k)\gamma_k^k. \tag{10}$$

Proof: See Lemma 2 in [7]. ∎

Define $(x)_+=\max\{x,0\}$ and the following quantities:

$$D_k^{\#}=\sum_{i=1}^{N}\psi(z_k+U_i((x^{\#})^{(i)}-z_k^{(i)})) \tag{11}$$
$$-(N-1)\psi(z_k)-\psi(x^{\#}),\quad S_k^{\#}=\sum_{j=0}^{k-1}\frac{1}{\theta_j}D_j^{\#},$$
$$C=\left(1-\frac{1}{N}\right)\left(F(x_0)-F^{\#}\right)_+\frac{1}{2}\|x^{\#}-x_0\|_L^2.$$

In the next theorem we consider $F^{\#}=F(x^{\#})$, with $x^{\#}$ a coordinate-wise minimizer of problem (1).

Theorem 2: Let Assumption 1 hold. Then, the iterates $(x_k)_{k\geq 1}$ of Algorithm 1 (APPROX) satisfy:

$$\mathbb{E}[F(x_k)-F^{\#}]\leq\frac{4N^2}{(k-1+2N)^2}\left[C+\frac{1}{N^2}\left(\mathbb{E}[S_k^{\#}]\right)_+\right].$$

Proof: Let us define:

$$\hat{F}_k=\sum_{l=0}^{k}\gamma_k^l\psi(z_l)+f(x_k) \tag{12}$$

From Lemma 2 and convexity of ψ, we have:

$$\hat{F}_k\geq F(x_k) \tag{13}$$

From Lemma 2

$$\hat{F}_{k+1}=\sum_{l=0}^{k+1}\gamma_{k+1}^l\psi(z_l)+f(x_{k+1})$$
$$=\sum_{l=0}^{k}\gamma_{k+1}^l\psi(z_l)+N\theta_k\psi(z_{k+1})+f(x_{k+1})$$

Using Assumption 1.A1, we further have:

$$\hat{F}_{k+1}\leq\sum_{l=0}^{k}\gamma_{k+1}^l\psi(z_l)+\frac{N^2\theta_k^2L_{i_k}}{2}\|z_{k+1}^{(i_k)}-z_k^{(i_k)}\|^2$$
$$f(y_k)+N\theta_k\psi(z_{k+1})+N\theta_k\langle U_{i_k}^T\nabla f(y_k),z_{k+1}^{(i_k)}-z_k^{(i_k)}\rangle$$
$$\leq\sum_{l=0}^{k}\gamma_{k+1}^l\psi(z_l)+\frac{N^2\theta_k^2L_{i_k}}{2}\|z_{k+1}^{(i_k)}-z_k^{(i_k)}\|^2$$
$$\theta_k\left(N\psi(z_{k+1})+f(y_k)+N\langle U_{i_k}^T\nabla f(y_k),z_{k+1}^{(i_k)}-y_k^{(i_k)}\rangle\right)$$
$$+(1-\theta_k)f(y_k)+N\theta_k\langle U_{i_k}^T\nabla f(y_k),y_k^{(i_k)}-z_k^{(i_k)}\rangle.$$

Taking expectation w.r.t. the block index i_k, we have:

$$\mathbb{E}_{i_k}[\hat{F}_{k+1}]\leq\sum_{l=0}^{k}\gamma_{k+1}^l\psi(z_l)+\theta_kN\mathbb{E}_{i_k}[\psi(z_{k+1})]$$
$$+\theta_k\left(f(y_k)+\langle\nabla f(y_k),\tilde{z}_{k+1}-y_k\rangle+\frac{N\theta_k}{2}\|\tilde{z}_{k+1}-z_k\|_L^2\right)$$
$$+(1-\theta_k)f(y_k)+\theta_k\langle\nabla f(y_k),y_k-z_k\rangle.$$

Note that from Step 1 of algorithm APPROX we have $\theta_k(y_k-z_k)=(1-\theta_k)(x_k-y_k)$. Moreover, using Lemma 1 and definition of $D_k^{\#}$, we get:

$$\mathbb{E}_{i_k}[\hat{F}_{k+1}]\leq\sum_{l=0}^{k}\gamma_{k+1}^l\psi(z_l)+\theta_k\langle\nabla f(y_k),x^{\#}-y_k\rangle$$
$$(1-\theta_k)f(y_k)+\theta_k\sum_{i=1}^{N}\psi(z_k+U_i((x^{\#})^{(i)}-z_k^{(i)}))$$
$$+\theta_kf(y_k)+(1-\theta_k)\langle\nabla f(y_k),x_k-y_k\rangle$$
$$+\frac{N\theta_k^2}{2}\|x^{\#}-z_k\|_L^2-\frac{N\theta_k^2}{2}\|x^{\#}-\tilde{z}_{k+1}\|_L^2$$
$$=\sum_{l=0}^{k}\gamma_{k+1}^l\psi(z_l)+\theta_k(N-1)\psi(z_k)+\theta_kD_k^{\#}$$
$$+(1-\theta_k)f(y_k)+(1-\theta_k)\langle\nabla f(y_k),x_k-y_k\rangle$$
$$+\theta_k\left(\psi(x^{\#})+f(y_k)+\langle\nabla f(y_k),x^{\#}-y_k\rangle\right)$$
$$+\frac{N\theta_k^2}{2}\|x^{\#}-z_k\|_L^2-\frac{N\theta_k^2}{2}\|x^{\#}-\tilde{z}_{k+1}\|_L^2.$$

From convexity of f, we obtain:

$$\mathbb{E}_{i_k}[\hat{F}_{k+1}]\leq(1-\theta_k)\sum_{l=0}^{k}\gamma_k^l\psi(z_l)+(1-\theta_k)f(x_k)$$
$$+\theta_k\psi(x^{\#})+\theta_kf(x^{\#})+\theta_kD_k^{\#}$$
$$+\frac{N^2\theta_k^2}{2}\|x^{\#}-z_k\|_L^2-\frac{N^2\theta_k^2}{2}\mathbb{E}_{i_k}[\|x^{\#}-z_{k+1}\|_L^2].$$

From Lemma 2 and relation (12), we have:

$$\mathbb{E}_{i_k}[\hat{F}_{k+1}]\leq\theta_kF^{\#}+(1-\theta_k)\hat{F}_k+\theta_kD_k^{\#}$$
$$+\frac{N^2\theta_k^2}{2}\|x^{\#}-z_k\|_L^2-\frac{N^2\theta_k^2}{2}\mathbb{E}_{i_k}[\|x^{\#}-z_{k+1}\|_L^2].$$

Dividing both sides in the last inequality by θ_k^2, using (7) and rearranging the terms, we obtain:

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2}\mathbb{E}_{i_k}[\hat{F}_{k+1}-F^{\#}]+\frac{N^2}{2}\mathbb{E}_{i_k}[\|x^{\#}-z_{k+1}\|_L^2]$$
$$\leq\frac{1-\theta_k}{\theta_k^2}\left(\hat{F}_k-F^{\#}\right)+\frac{N^2}{2}\|x^{\#}-z_k\|_L^2+\frac{1}{\theta_k}D_k^{\#}.$$

Taking expectation with respect to $(i_0,\cdots,i_k)$ in the inequality above and unrolling the recurrence, we get:

$$\frac{1-\theta_k}{\theta_k^2}\mathbb{E}[\hat{F}_k-F^{\#}]+\frac{N^2}{2}\mathbb{E}[\|x^{\#}-z_k\|_L^2]$$
$$\leq\frac{1-\theta_0}{\theta_0^2}\left(\hat{F}_0-F^{\#}\right)+\frac{N^2}{2}\|x^{\#}-z_0\|_L^2+\sum_{j=0}^{k-1}\frac{1}{\theta_j}\mathbb{E}[D_j^{\#}].$$

From (13) and the inequality above, we obtain for $k\geq 1$:

$$\mathbb{E}[F(x_k)-F^{\#}]\leq\mathbb{E}[\hat{F}_k-F^{\#}]$$
$$\leq\frac{1-\theta_0}{\theta_0^2}\theta_{k-1}^2\left(\hat{F}_0-F^{\#}\right)$$
$$+\frac{N^2}{2}\theta_{k-1}^2\|x^{\#}-z_0\|_L^2+\theta_{k-1}^2\sum_{j=0}^{k-1}\frac{1}{\theta_j}\mathbb{E}[D_j^{\#}].$$

Using (7) and the last inequality: we obtain for $k\geq 1$:

$$\mathbb{E}[F(x_k)-F^{\#}]\leq\theta_{k-1}^2N^2\left(1-\frac{1}{N}\right)\left(F(x_0)-F^{\#}\right)_+$$
$$+\theta_{k-1}^2N^2\left(\frac{1}{2}\|x^{\#}-x_0\|_L^2+\frac{1}{N^2}\left(\sum_{j=0}^{k-1}\frac{1}{\theta_j}\mathbb{E}[D_j^{\#}]\right)_+\right)$$
$$\leq\frac{4N^2}{(k-1+2N)^2}\left(C+\frac{1}{N^2}\left(\sum_{j=0}^{k-1}\frac{1}{\theta_j}\mathbb{E}[D_j^{\#}]\right)_+\right),$$

where we used that $\hat{F}_0=F(x_0)$, $z_0=x_0$, $\theta_0=\frac{1}{N}$ and $\theta_{k-1}\leq\frac{2}{k-1+2N}$. This concludes our statement. ■

Note that if the second term that measure nonseparability of ψ along the iterates:

$$\mathbb{E}[S_k^{\#}]=\sum_{j=0}^{k-1}\frac{1}{\theta_j}\mathbb{E}[D_j^{\#}]$$

is bounded, then from Theorem 2, we have that the convergence rate of algorithm APPROX is of the order $\mathcal{O}\left(1/k^2\right)$. However, it is still an open question wether this term is bounded, although in our numerical simulations (see also next section), we observed empirically that $S_k^{\#}$ is bounded. Since an optimal point x^* is also a coordinate-wise minimizer, we also get from previous theorem:

$$\mathbb{E}[F(x_k)-F^*]$$
$$\leq\frac{4N^2}{(k-1+2N)^2}\left[C+\frac{1}{N^2}\left(\sum_{j=0}^{k-1}\frac{1}{\theta_j}\mathbb{E}[D_j^*]\right)_+\right],$$

where D_j^* has the same expression as in (11), but evaluated in x^*. S_k^* is defined similarly. Note that if ψ is separable we have $D_j^*=0$ for all $j\geq 0$. Hence, we recover the convergence rate in [7] for this particular case.

IV. Simulations

In this section we use Algorithm 1 for solving several nonseparable problems. For a given coordinate-wise minimizer $x^{\#}$, we plot the behaviour along the iterates of the following quantities: (1) Sequence $D_k^{\#}$ as defined in (11); (2) Sequence $S_k^{\#}$ as defined in (11); (3) $F(x_k)-F(x^{\#})$ and the estimates: $\mu_k=\frac{4N^2C}{(k-1+2N)^2}$ and $\nu_k=\mu_k+\frac{4(S_k^{\#})_+}{(k-1+2N)^2}$. If $x^{\#}$ is a minimizer, μ_k is the bound obtained in [7].

In the plots, we denoted a minimizer point by x^*, in the case the algorithm APPROX converges to such a point. For the nonseparable function ψ we consider three choices: $M/6\|x\|^3$, $M\|x\|$ and the TV regularization. For the smooth function f we consider the loss in the quadratic or logistic regression. In all simulations the starting point x_0 was generated from a standard normal distribution and only one coordinate was updated per iteration (i.e., $n_i=1$ for all $i=1:N$). In all the examples we run the algorithm 10 times and we plot the average of the results.

A. Quadratic with cubic regularization

The first composite function we consider corresponds to the subproblem in the cubic Newton method [13]:

$$F(x)=\frac{1}{2}x^TAx+b^Tx+\frac{M}{6}\|x\|^3, \tag{14}$$

with $A\in\mathbb{R}^{n\times n}$ a positive semidefinite matrix and $b\in\mathbb{R}^n$. We generate $b\in\mathbb{R}^n$ from a standard normal distribution $\mathcal{N}(0,1)$. The matrix A was considered as $A=B^TB$, with $B\in\mathbb{R}^{m\times n}$ a matrix generated from a standard normal distribution $\mathcal{N}(0,1)$. In the simulations we consider $n=100$, $m=10$ and $M=1$. Note that in this case the objective function is differentiable and uniformly convex, hence there is a unique minimizer and coincides with the coordinate-wise minimizer. The results are given in Figure 1. We observe that the algorithm APPROX converges to the unique minimizer x^*. Note that for this problem we have S_k^* negative. Hence, the expected difference of the function values is bounded by μ_k (the theoretical bound obtained in [7] for separable ψ). In other words, the algorithm converges with a rate $\mathcal{O}\left(1/k^2\right)$.

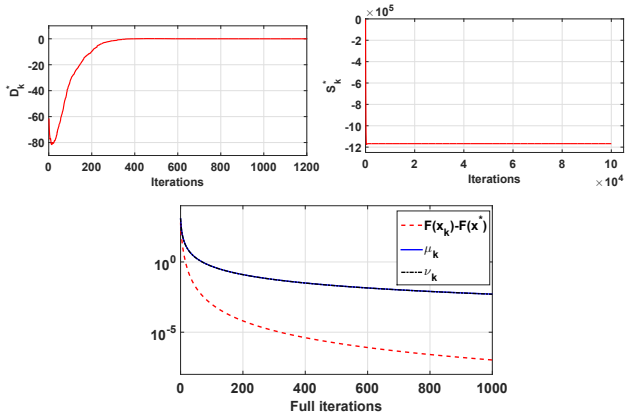

Fig. 1. Behaviour of $D_k^{\#}=D_k^*$, $S_k^{\#}=S_k^*$, $F(x_k)-F^*$, μ_k and ν_k along iterations: quadratic with cubic regularization.

B. Logistic regression with cubic regularization

Then, we consider the logistic loss with cubic regularization:

$$F(x) = \frac{1}{T}\sum_{j=1}^{T} \log\left(1+\exp(a_j^T x)\right) + \frac{M}{6}\|x\|^3, \qquad (15)$$

with $a_j \in \mathbb{R}^n$, for $j = 1:T$. The vectors $a_j \in \mathbb{R}^n$ were generated from a standard normal distribution $\mathcal{N}(0,1)$. In the simulations, we choose $n = 100$, $T = 1000$ and $M = 1$. Also in this case the objective function is differentiable and uniformly convex, hence there is a unique minimizer and coincides with the coordinate-wise minimizer. The results are given in Figure 2. The algorithm APPROX also converges to the minimizer point x^*. Note that also for this problem we have S_k^* negative. Hence, the expected difference of the function values is bounded by μ_k. In other words, the algorithm converges with a rate $\mathcal{O}\left(1/k^2\right)$ also in this case.

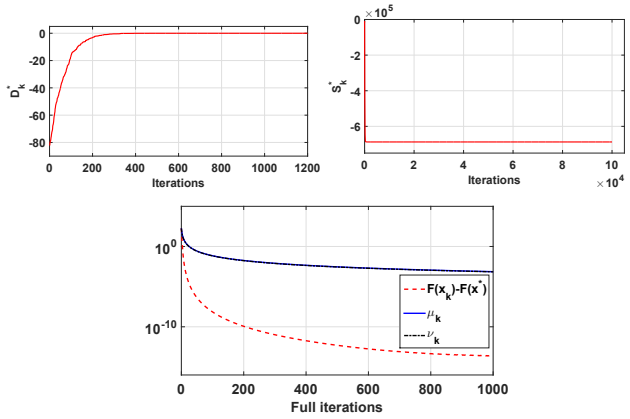

Fig. 2. Behaviour of $D_k^{\#} = D_k^*$, $S_k^{\#} = S_k^*$, $F(x_k) - F^*$, μ_k and ν_k along iterations: logistic regression with cubic regularization.

C. Quadratic with ℓ_2 regularization

Consider the composite objective function:

$$F(x) = \frac{1}{2}x^T A x + b^T x + M\|x\| \qquad (16)$$

with A a positive semidefinite matrix. The matrix $A = B^T B$, with $B \in \mathbb{R}^{m\times n}$ a matrix generated from an uniform distribution with values between 0 and 1. Define $e \in \mathbb{R}^n$ as the vector of all ones. The vector b is chosen $b = (1/2)e$. In the simulations we consider $n = 100$, $m = 10$ and $M = 1$. The results are reported in Figure 3. Note that, the function in (16) is differentiable for all $x \neq 0$. Hence, the only point that can be coordinate-wise minimizer, without being a minimizer, is $x = 0$. For example, we know that $x^{\#} = 0$ is a coordinate-wise minimizer, provided that $|b^{(i)}| \leq M$ for all $i = 1:n$. Indeed, consider a_{ii}, the ith entry on the diagonal of A and e_i the ith vector of the canonical basis. Then, we have for all $d \in \mathbb{R}$:

$$F(x^{\#} + e_i d) = \frac{1}{2}a_{ii}d^2 + b^{(i)}d + M|d| \geq 0 = F(x^{\#}),$$

where we used that $a_{ii} \geq 0$ since the matrix A is positive semidefinite. Hence, for $b = 1/2e$ and $M = 1$, we have that $x^{\#} = 0$ is coordinate-wise minimizer, but not a (global) minimizer. However, we observed that the algorithm APPROX converges always to the (global) minimizer x^*. Note that for this problem we have $D_k^* \geq 0$, but S_k^* is bounded. Moreover we observe that the expected difference of the function values is bounded by μ_k (the theoretical bound obtained in [7]). Hence, the algorithm still converges with rate $\mathcal{O}\left(1/k^2\right)$.

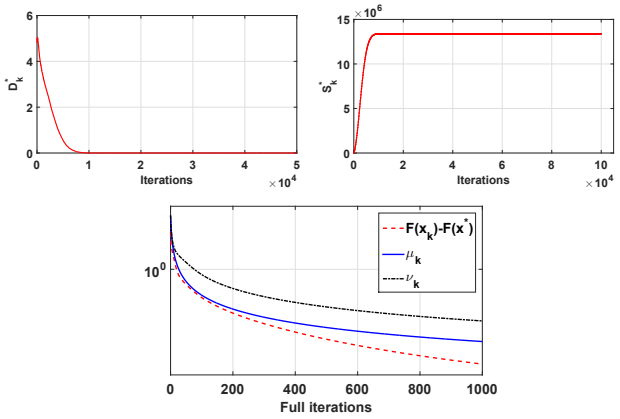

Fig. 3. Behaviour of D_k^*, S_k^*, $F(x_k) - F^*$, μ_k and ν_k along iterations: quadratic with ℓ_2 regularization.

D. Logistic regression with ℓ_2 regularization

Consider the composite objective function:

$$F(x) = \frac{1}{T}\sum_{j=1}^{T} \log\left(1+\exp(a_j^T x)\right) + M\|x\|, \qquad (17)$$

with a_j generated from a standard normal distribution $\mathcal{N}(0,1)$. In the simulations, we choose $n = 100$, $T = 1000$ and $M = 1$. The simulations are in Figure 4. We observed that the algorithm APPROX converges to a minimizer point x^*. Note that for this problem we have $D_k^* \geq 0$, but S_k^* is bounded. Moreover we observe that the expected difference of the function values is bounded by μ_k. Hence, the algorithm still converges with a rate $\mathcal{O}\left(1/k^2\right)$.

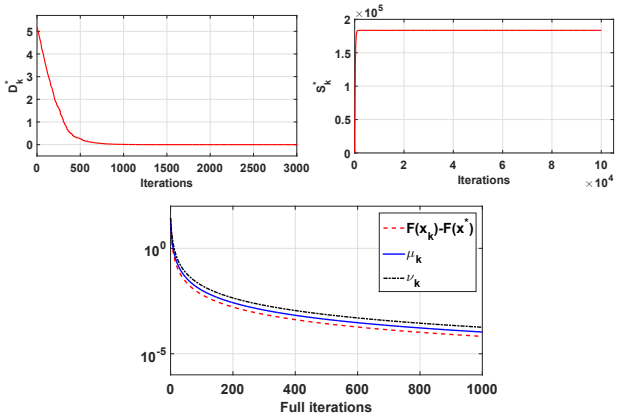

Fig. 4. Behaviour of D_k^*, S_k^*, $F(x_k) - F^*$, μ_k and ν_k along iterations: logistic regression with ℓ_2 regularization.

E. Quadratic with TV Regularization

Finally, we consider the following two dimensional quadratic with TV regularization composite function:

$$F(x_1, x_2) = x_1^2 + x_2^2 - x_1 x_2 + x_1 + x_2 + |x_1 - x_2|. \quad (18)$$

Note that, the unique optimal point of this problem is $(x_1^*, x_2^*) = (-1, -1)$ and the optimal value is $F^* = -1$. Moreover, any point (α, α), with $\alpha \in [-2, 0]$, is a coordinate-wise minimizer for (18). Indeed, we have that $F(\alpha, \alpha) = \alpha^2 + 2\alpha$. Moreover, for any $d \in \mathbb{R}$, we have:

$$\begin{aligned} F(\alpha + d, \alpha) &= \alpha^2 + 2\alpha + d^2 + (\alpha + 1)d + |d| \\ &\geq F(\alpha, \alpha) + (\alpha + 1)d + |d| \geq F(\alpha, \alpha). \end{aligned}$$

Similarly, we can show that $F(\alpha, \alpha + d) \geq F(\alpha, \alpha)$. We first run the algorithm APPROX five times, the starting point and the solution found by the algorithm is presented in Table I. Note that APPROX converges to differents coordinate-wise minimizers even when we initialize the algorithm with the same starting point. Finally, we plot in Figure 5, $D_k^\#, D_k^*$, $S_k^\#, S_k^*$, $F(x_k) - F^\#, F(x_k) - F^*$, μ_k and ν_k for both the coordinate-wise minimizer to which the algorithm converges (left) and for the optimal point (right). Note that $D_k^\#$ converges to zero and $S_k^\#$ is bounded. Hence, we observe that the sequence $F(x_k)$ converges to $F^\#$ with rate $\mathcal{O}\left(1/k^2\right)$. On other hand, we observe that the sequence D_k^* converges to a value different from zero and S_k^* diverges. Moreover, the sequence $F(x_k)$ does not converge to F^*.

Starting point	Solution
(0.1747, 0.0150)	(-0.0061, -0.0061)
(-0.6718, 0.5756)	(-0.6718, -0.6718)
(0.5377, 1.8339)	(-0.1466, -0.1466)
(0.5377, 1.8339)	(-0.2802, -0.2802)
(0.5377, 1.8339)	(-0.5289, -0.5289)

TABLE I
STARTING POINTS AND SOLUTIONS FOUND WITH APPROX.

V. CONCLUSIONS

In this paper we have extended the accelerated coordinated proximal gradient method from [7] to composite convex optimization problems where both terms are nonseparable. We have shown that the diference of function values, $F(x_k) - F(x^\#)$ (where $x^\#$ is a coordinate-wise minimizer), converges with a rate of order $\mathcal{O}((\mathrm{C} + (\mathbb{E}[S_k^\#])_+)/k^2)$. We conjecture that the second term $S_k^\#$ is bounded and consequently algorithm APPROX converges with rate $\mathcal{O}(1/k^2)$ to a coordinate-wise minimizer, as this is observed in all our numerical simulations. For future work, we plan to derive a theoretical bound for $S_k^\#$.

ACKNOWLEDGMENT

The research leading to these results has received funding from: ITN-ETN project TraDE-OPT funded by the EU Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 861137; UEFISCDI PN-III-P4-PCE-2021-0720, under project L2O-MOC, nr. 70/2022; NO Grants 2014-2021 RO-NO-2019-0184, under project ELO-Hyp, contract no. 24/2020.

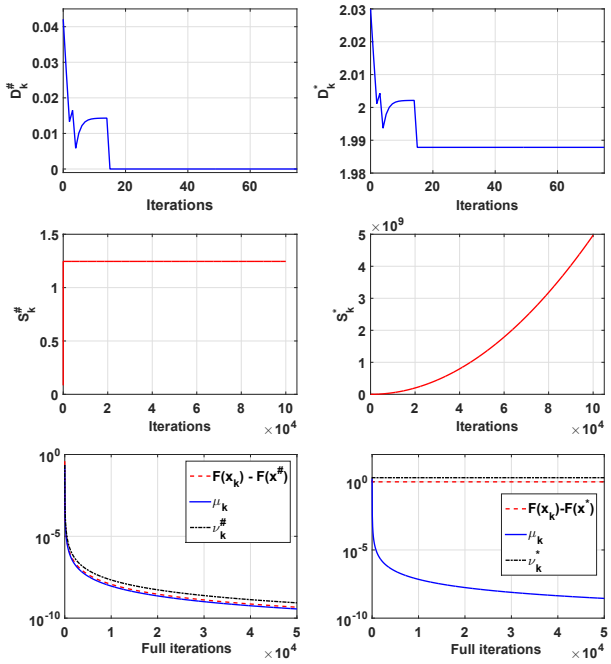

Fig. 5. Behaviour of $D_k^\#, S_k^\#, F(x_k) - F^\#$ (left) , and $D_k^*, S_k^*, F(x_k) - F^*$ (right): quadratic with TV regularization.

REFERENCES

[1] A. Aberdam and A. Beck *An Accelerated Coordinate Gradient Descent Algorithm for Non-separable Composite Optimization*, J. Opt. Theory Applications, doi: 10.1007/s10957-021-01957-1, 2021.

[2] Z. Allen-Zhu, Z. Qu, P. Richtarik and Y. Yuan. *Even Faster Accelerated Coordinate Descent Using Non-Uniform Sampling*. Proceedings of Int. Conference on Machine Learning, PMLR 48:1110–1119, 2016.

[3] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optimization, 23(4): 2037–2060, 2013.

[4] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.

[5] J. Bolte, S. Sabach and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146: 459–494, 2014.

[6] A. Daneshmand, F. Facchinei, V. Kungurtsev and G. Scutari, *Hybrid Random/Deterministic Parallel Algorithms for Convex and Nonconvex Big Data Optimization*, IEEE Tran. Sig. Proc., 63: 3914–3929, 2015.

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

[8] P. Latafat, A. Themelis and P. Patrinos *Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems*, Mathematical Programming, 2021.

[9] T. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[10] I. Necoara and D. Clipici, *Parallel random coordinate descent methods for composite minimization: convergence analysis and error bounds*, SIAM Journal on Optimization, 26(1): 197-226, 2016.

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

[12] Yu. Nesterov, *Inexact basic tensor methods*, Core Paper 23, 2019.

[13] Yu. Nesterov and B.T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Progr., 108: 177–205, 2006.

[14] Yu. Nesterov and S.U. Stich, *Efficiency of the accelerated coordinate descent method on structured optimization problems*, SIAM Journal on Optimization, 27(1): 110–123, 2017.

[15] P. Richtarik and M. Takac, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Mathematical Programming, 144: 1–38, 2014.

[16] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, SIAM J. on Optimization, 2008, submitted.

[17] S.J. Wright, *Coordinate descent algorithms*, Mathematical Programming, 151(1): 3–34, 2015.