2016/03

Efficiency of accelerated coordinate descent method on structured optimization problems

Yu. Nesterov and S. Stich

February 1, 2016

DISCUSSION PAPER

Center for Operations Research
and Econometrics

Voie du Roman Pays, 34
B-1348 Louvain-la-Neuve
Belgium
http://www.uclouvain.be/core

CORE DISCUSSION PAPER
2016/03

Efficiency of accelerated coordinate descent method on structured optimization problems

Yu. Nesterov[1] and S. Stich[2]

February 1, 2016

Abstract

In this paper we prove a new complexity bound for a variant of Accelerated Coordinate Descent Method [7]. We show that this method often outperforms the standard Fast Gradient Methods (FGM, [3, 6]) on optimization problems with dense data. In many important situations, the computational expenses of oracle and method itself at each iteration of our scheme are perfectly balanced (both depend linearly on dimensions of the problem). As application examples, we consider unconstrained convex quadratic minimization, and the problems arising in Smoothing Technique [6]. On some special problem instances, the provable acceleration factor with respect to FGM can reach the square root of the number of variables. Our theoretical conclusions are confirmed by numerical experiments.

Keywords: Convex optimization, structural optimization, fast gradient methods, coordinate descent methods, complexity bounds

[1] Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium.
[2] CORE/ICTEAM (UCL). This research was supported by Swiss Science Foundation (SNF).

The research results presented in this paper have been supported by a grant “Action de recherche concertée ARC 04/09-315” from the “Direction de la recherche scientifique - Communauté française de Belgique”. Scientific responsibility rests with the authors.

1 Introduction

Motivation. In the last years, coordinate descent methods attract more and more attention of the Optimization Community. Its popularity is based mainly on the fact that they can be applied to problems of a very big size. Starting from the paper [7], it became possible to provide the randomized variants of these schemes with very attractive worst-case efficiency guarantees, which take into account a very high sparsity of the data. Consequently, the further developments of these methods were naturally related to the needs of Big-Data machinery: parallelization, distributed computing, etc (see, for example, [4, 5]). However, in this paper we show that the coordinate descent strategies can be useful even for the problems of moderate-size when the data is dense.

In [7], there was proposed a variant of Fast Gradient Method [3], where the gradient step was replaced by a step along coordinate direction (we call this method Accelerated Coordinate Descent Method, ACDM for short). It was suggested to choose the corresponding active coordinate randomly, in accordance to uniform distribution. The expected complexity of this scheme for finding an ϵ-solution for unconstrained minimization problem is of the order

$$O\left(\frac{n}{\epsilon^{1/2}} \max_{1 \le i \le n} L_i\right) \tag{1.1}$$

iterations, where L_i is the uniform upper bound on the ith diagonal element of the Hessian of the objective function, and n is the number of variables. At the same time, in [7] it was also mentioned that this scheme is not appropriate for Huge-Scale optimization problems since it needs at least one full-dimensional vector operation at each iteration.

Complexity bound (1.1) was improved in [2] up to the level

$$O\left(\left[\frac{n}{\epsilon} \sum_{i=1}^{n} L_i\right]^{1/2}\right) \tag{1.2}$$

iterations. For choosing the active coordinate, the authors suggest to use probabilities $L_i \left[\sum_{k=1}^{n} L_k\right]^{-1}$, $i = 1, \dots, n$. Finally, in our paper we get the further improvement in the complexity of ACDM, up to the level

$$O\left(\frac{1}{\epsilon^{1/2}} \sum_{i=1}^{n} L_i^{1/2}\right) \tag{1.3}$$

iterations. The probabilities we use now are defined as $L_i^{1/2} \left[\sum_{k=1}^{n} L_k^{1/2}\right]^{-1}$. This is the first time when we get the complexity estimate of ACDM, which does not depend explicitly in the dimension of the space of variables.

Another important result of our paper consists in finding interesting applications, where the new scheme becomes dominant. We show that in *all* unconstrained convex optimization problems obtained by Smoothing Technique [6], our method provably outperforms the standard Fast Gradient Methods. For some classes of problems, the gain in the computational time reaches the square root of the dimension. This improvement is mainly achieved due to the fact, that in many situations the computational expenses at

each iteration of our method are perfectly balanced with the computational time spent for updating the results of matrix-vector products (both depend linearly in the dimension of the problem). For the standard first-order methods, this is not true even if we apply them for unconstrained minimization of convex quadratic function with dense matrix. For the latter problem, the worst-case estimates of computational time of our method are provably better than the estimates of unbeatable Conjugate Gradients.[1)] Note also that for problems with explicit minimax structure, it is always possible to compute good bounds for the constants L_i, $i = 1, \ldots, n$ (see Section 3.3).

Contents. In Section 2, we present a new version of ACD-method for solving the problem of unconstrained minimization of strongly convex function with Lipschitz continuous partial derivatives. The probability of choosing component i to be active is define as $L_i^{1/2} \left[\sum_{k=1}^{n} L_k^{1/2} \right]^{-1}$, where L_i is the corresponding Lipschitz constant. Our scheme, complexity analysis, and efficiency estimates are nonstandard since they all are *continuous* in the convexity parameter of the objective function. In order to obtain the efficiency estimates and the rules of the method just for differentiable convex function, we need to pass to the limit in the corresponding expressions, tending the convexity parameter to zero.[2)]

In Section 3, we present some applications, where the new method has the best known worst-case bounds for the total computational time. In Section 3.1 we develop a general model of the objective function, which allows to update and compute efficiently the directional derivatives. Our key observation is that in many cases a single directional derivative can be easily computed, often in linear time. After that, we analyze the behavior of the new ACDM on the problems of quadratic minimization (Section 3.2) and in the framework of Smoothing Technique (Section 3.3)). In both cases, we show that our method has better worst-case guarantees in computational time, as compared with the total computational time of the standard FGM.

We conclude the paper by presenting the results of preliminary computational experiments (Section 4). At our class of test problems, new ACDM always outperforms the standard Fast Gradient Method with automatic adjustment of the Lipschitz constant for the gradient.

Notation. In what follows, we assume that the finite-dimensional linear vector space of variables $\mathbb{E}$, $\dim \mathbb{E} = N$, is represented as a direct product of n-dimensional spaces $\mathbb{E}^{(i)}$, $\dim \mathbb{E}^{(i)} = n_i$:

$$\mathbb{E} \;=\; \bigotimes_{i=1}^{n} \mathbb{E}^{(i)}, \quad N \;=\; \sum_{i=1}^{n} n_i.$$

We denote by $\mathbb{E}_*^{(i)}$, $i \in \{1 : n\}$, the corresponding dual spaces. Thus, $\mathbb{E}_* = \bigotimes_{i=1}^{n} \mathbb{E}_*^{(i)}$. Value

[1)] Of course, this result does not contradict to the well known fact on optimality of conjugate gradient methods. Note that coordinate descent methods belong to another family of optimization schemes, which *do not* generate minimization sequences belonging to Krylov spaces.

[2)] When this paper was already finished, we found a very recent paper [1], where there was analyzed a version of ACDM with the same distribution of probabilities. This version can be also used for minimizing strongly convex functions. However, it becomes inefficient as the convexity parameter goes to zero.

of linear function $s^{(i)} \in \mathbb{E}_*^{(i)}$ at point $x^{(i)} \in \mathbb{E}^{(i)}$ is denoted by $\langle s^{(i)}, x^{(i)} \rangle$. We define

$$\langle s, x \rangle \stackrel{\text{def}}{=} \sum_{i=1}^{n} \langle s^{(i)}, x^{(i)} \rangle, \quad x \in \mathbb{E},\ s \in \mathbb{E}_*.$$

We define also the *partition operators* $U_i : \mathbb{E}_i \to \mathbb{E}$, $i = 1, \ldots, n$, by identity

$$x = \left(x^{(1)}, \ldots, x^{(n)}\right) = \sum_{i=1}^{n} U_i x^{(i)}, \quad x^{(i)} \in \mathbb{E}^{(i)},\ i \in \{1:n\}.$$

If $\mathbb{E} = \mathbb{R}^N$, Then the matrices U_i are composed by columns of the unit $N \times N$-matrix:

$$I_N = (U_1, \ldots, U_n).$$

For a linear operator A, acting from one linear vector space $\mathbb{E}'$ to another linear vector space $\mathbb{E}''_*$, we define its adjoint operator by identity

$$\langle Au, v \rangle = \langle A^* v, u \rangle, \quad u \in \mathbb{E}',\ v \in \mathbb{E}''.$$

Clearly, $A^* : \mathbb{E}'' \to \mathbb{E}'_*$.

For all spaces $\mathbb{E}^{(i)}$, we fix self-adjoint positive-definite operators $B_i : \mathbb{E}^{(i)} \to \mathbb{E}_*^{(i)}$ (notation: $B_i = B_i^* \succ 0$), $i = 1, \ldots, n$. Using these operators, we can introduce in these spaces the *scalar products* and *Euclidean norms*:

$$\langle x^{(i)}, y^{(i)} \rangle_i \stackrel{\text{def}}{=} \langle B_i x^{(i)}, y^{(i)} \rangle, \quad \|x^{(i)}\|_i^2 \stackrel{\text{def}}{=} \langle B_i x^{(i)}, x^{(i)} \rangle, \quad x^{(i)}, y^{(i)} \in \mathbb{E}^{(i)},\ i \in \{1:n\}.$$

Similarly, for the dual spaces, we have the following definitions:

$$\langle s^{(i)}, v^{(i)} \rangle_i^* \stackrel{\text{def}}{=} \langle s^{(i)}, B_i^{-1} v^{(i)} \rangle,\ \|s^{(i)}\|_i^* \stackrel{\text{def}}{=} \langle s^{(i)}, B_i^{-1} s^{(i)} \rangle,\ s^{(i)}, v^{(i)} \in \mathbb{E}_*^{(i)},\ i \in \{1:n\}.$$

Thus, we get valid Cauchy-Schwartz inequalities:

$$\langle s^{(i)}, x^{(i)} \rangle \leq \|s^{(i)}\|_i^* \cdot \|x^{(i)}\|_i, \quad x^{(i)} \in \mathbb{E}^{(i)},\ s^{(i)} \in \mathbb{E}_*^{(i)},\ i \in \{1:n\}. \tag{1.4}$$

In order to define the norms for the whole space $\mathbb{E}$, we use the scaling coefficients $L = (L_1, \ldots, L_n)$ (to be defined later in (2.3)), and the tolerance parameter $\alpha \in [0, 1]$. For $x = (x^{(1)}, \ldots, x^{(n)}) \in \mathbb{E}$ and $s = (s^{(1)}, \ldots, s^{(n)}) \in \mathbb{E}_*$ denote

$$\begin{aligned} \langle s, x \rangle &= \sum_{i=1}^{n} \langle s^{(i)}, x^{(i)} \rangle, \\ \|x\|_{[\alpha]}^2 &= \sum_{i=1}^{n} L_i^{\alpha} \|x^{(i)}\|_i^2, \qquad (1.5) \\ \|s\|_{[\alpha]*}^2 &= \sum_{i=1}^{n} L_i^{-\alpha} \left(\|s^{(i)}\|_i^*\right)^2. \end{aligned}$$

Clearly, for all $x \in \mathbb{E}$ and $s \in \mathbb{E}^*$ we have

$$\langle s, x \rangle \leq \|x\|_{[\alpha]} \cdot \|s\|_{[\alpha]*}. \tag{1.6}$$

In the case $\mathbb{E} = \mathbb{R}^N$, we have $\|x\|_{[\alpha]}^2 = \langle B_\alpha x, x\rangle$, $\|s\|_{[\alpha]*}^2 = \langle s, B_\alpha^{-1} s\rangle$, with

$$B_\alpha \;=\; \sum_{i=1}^n L_i^\alpha U_i B_i U_i^T, \quad B_\alpha^{-1} \;=\; \sum_{i=1}^n L_i^{-\alpha} U_i B_i^{-1} U_i^T.$$

For a differentiable function $f(x)$, $x \in \operatorname{dom} f \subseteq \mathbb{E}$, denote by $\nabla f(x) \in \mathbb{E}_*$ its *gradient*. Then, its *partial derivatives* are defined as follows:

$$\nabla_i f(x) \;\stackrel{\text{def}}{=}\; U_i^T - \nabla f(x) \;\in\; E_*^{(i)}, \quad i \in \{1:n\}.$$

If function f is convex, then for any $x \in \operatorname{dom} f$ and any partial displacement $h^{(i)} \in E^{(i)}$ satisfying condition $x + U_i h^{(i)} \in \operatorname{dom} f$ (we call it *feasible*), we have

$$\begin{aligned} f(x + U_i h^{(i)}) \;&\geq\; f(x) + \langle \nabla f(x), U_i h^{(i)}\rangle \;=\; f(x) + \langle U_i^T \nabla f(x), h^{(i)}\rangle \\ &=\; f(x) + \langle \nabla_i f(x), h^{(i)}\rangle, \quad i \in \{1:n\}. \end{aligned} \tag{1.7}$$

2 Accelerated Coordinate Descent Method

Consider the following optimization problem:

$$\min_{x \in \mathbb{E}} f(x), \tag{2.1}$$

where function f is convex and continuously differentiable on $\mathbb{E}$. We assume that this problem is solvable and $x_* \in E$ is its optimal solution.

Global behavior of function $f(\cdot)$ is described by the following characteristics.

- *Parameter of strong convexity* $\sigma_\alpha \geq 0$, such that

$$f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2}\sigma_{1-\alpha}\|y - x\|_{[1-\alpha]}^2, \quad \forall x, y \in \mathbb{E}. \tag{2.2}$$

- *Lipschitz constants* L_i *for partial derivatives*:

$$\begin{aligned} \|\nabla_i f(x + U_i h^{(i)}) - \nabla_i f(x)\|_i^* \;&\leq\; L_i \|h^{(i)}\|_i, \\ \forall\, x \in \mathbb{E},\; h^{(i)} &\in \mathbb{E}^{(i)}, \quad i \in \{1:n\}. \end{aligned} \tag{2.3}$$

These inequalities are equivalent to the following conditions:

$$\begin{aligned} f(x + U_i h^{(i)}) \;&\leq\; f(x) + \langle \nabla_i f(x), h^{(i)}\rangle + \tfrac{1}{2} L_i \|h^{(i)}\|_i^2, \\ \forall\, x \in \mathbb{E},\; h^{(i)} &\in \mathbb{E}^{(i)},\; i \in \{1:n\}. \end{aligned} \tag{2.4}$$

For the sake of simplicity, we assume that parameters σ_α and $L \stackrel{\text{def}}{=} (L_1, \ldots, L_n)$ are known.

Let us define now the *partial gradient step* at point $x \in \mathbb{E}$ along the active coordinate $i \in \{1:n\}$:

$$h^{(i)}(x) \;\stackrel{\text{def}}{=}\; -B_i^{-1} \nabla_i f(x). \tag{2.5}$$

In view of inequality (2.4), for any stepsize $\tau \in \mathbb{R}$, we have

$$\begin{aligned} f(x + \tau U_i h^{(i)}(x)) - f(x) &\leq \tau \langle \nabla_i f(x), h^{(i)}(x) \rangle + \tfrac{\tau^2}{2} L_i \|h^{(i)}(x)\|_i^2 \\ &= -\tau(1 - \tfrac{1}{2}\tau L_i)(\|\nabla_i f(x)\|_i^*)^2. \end{aligned} \tag{2.6}$$

Finally, we need to define a random generator $j = \mathcal{R}_\beta(L)$, $\beta \in [0, 1]$, which generates random numbers $j \in \{1 : n\}$ with the following probabilities:

$$\pi_\beta[i] \equiv \text{Prob}\,(j = i) \stackrel{\text{def}}{=} \tfrac{1}{S_\beta} L_i^\beta, \quad i \in \{1 : n\}, \tag{2.7}$$

where $S_\beta = \sum_{i=1}^n L_i^\beta$.

For solving the problem (2.1), consider the following method.

Method $ACDM_\alpha(x_0)$

1. Define $v_0 = x_0 \in \mathbb{E}$, $A_0 = 0$, $B_0 = 1$, and $\beta = \frac{\alpha}{2}$.

2. For $t \geq 0$, iterate:

1) Choose active coordinate $i_t = \mathcal{R}_\beta(L)$.

2) Find parameter $a_{t+1} > 0$ from equation $a_{t+1}^2 S_\beta^2 = A_{t+1} B_{t+1}$,

where $A_{t+1} = A_t + a_{t+1}$ and $B_{t+1} = B_t + \sigma_{1-\alpha} a_{t+1}$.

3) Define $\alpha_t = \frac{a_{t+1}}{A_{t+1}}$, $\beta_t = \frac{\sigma_{1-\alpha} a_{t+1}}{B_{t+1}}$, and $y_t = \frac{(1-\alpha_t)x_t + \alpha_t(1-\beta_t)v_t}{1-\alpha_t\beta_t}$.

4) Compute $\nabla_{i_t} f(y_t)$. Update $x_{t+1} = y_t + \frac{1}{L_{i_t}} U_{i_t} h^{(i_t)}(y_t)$,

and $v_{t+1} = (1 - \beta_t)v_t + \beta_t y_t + \frac{a_{t+1}}{L_{i_t}^{1-\alpha} B_{t+1} \pi_\beta[i_t]} U_{i_t} h^{(i_t)}(y_t)$.

(2.8)

Denote $w_t = (1 - \beta_t)v_t + \beta_t y_t$. Then

$$y_t = \tfrac{(1-\alpha_t)x_t}{1-\alpha_t\beta_t} + \tfrac{\alpha_t(1-\beta_t)}{1-\alpha_t\beta_t} \cdot \tfrac{w_t - \beta_t y_t}{1-\beta_t} = \tfrac{(1-\alpha_t)x_t + \alpha_t w_t}{1-\alpha_t\beta_t} - \tfrac{\alpha_t\beta_t y_t}{1-\alpha_t\beta_t}.$$

Thus, in method (2.8) we have the following representation:

$$y_t = (1 - \alpha_t)x_t + \alpha_t w_t. \tag{2.9}$$

Method (2.8) generates random output, which depends on particular implementation of the collection of i.i.d.-variables $\mathcal{I}_t = \{i_0, \ldots, i_t\}$ (define $\mathcal{I}_{-1} = \emptyset$). In what follows, notation $E_{\mathcal{I}_t}(\cdot)$ denotes the expectation of corresponding random variables.

Theorem 1 *Let sequences $\{x_t\}_{t\geq 0}$ and $\{v_t\}_{t\geq 0}$ be generated by method (2.8). Then, for any $t \geq 0$ we have*

$$2A_t E_{\mathcal{I}_{t-1}}(f(x_t) - f(x_*)) + B_t E_{\mathcal{I}_{t-1}}(\|v_t - x_*\|^2_{[1-\alpha]}) \;\leq\; \|x_0 - x_*\|^2_{[1-\alpha]}, \tag{2.10}$$

where

$$\begin{aligned} A_t &\geq \tfrac{1}{4\sigma_{1-\alpha}}\left[(1+\gamma)^t - (1-\gamma)^t\right]^2 \geq \tfrac{1}{4S^2_\beta}t^2, \\ B_t &\geq \tfrac{1}{4}\left[(1+\gamma)^t + (1-\gamma)^t\right]^2, \end{aligned} \tag{2.11}$$

and $\gamma = \frac{\sigma^{1/2}_{1-\alpha}}{2S_{\alpha/2}}$.

Proof:
Denote $r_t^2 = \|v_t - x_*\|^2_{[1-\alpha]}$. Then

$$\|v_{t+1} - x_*\|^2_{[1-\alpha]} \;=\; \sum_{i\neq i_t} L_i^{1-\alpha}\|w_t^{(i)} - x_*^{(i)}\|_i^2 + L_{i_t}^{1-\alpha}\left\|w_t^{(i_t)} - x_*^{(i_t)} + \tfrac{a_{t+1}h^{(i_t)}(y_t)}{L_{i_t}^{1-\alpha}B_{t+1}\pi_\beta[i_t]}\right\|^2_{i_t}$$

$$= \|w_t - x_*\|^2_{1-\alpha} - \tfrac{2a_{t+1}}{B_{t+1}\pi_\beta[i_t]}\langle \nabla_{i_t} f(y_t), w_t^{(i_t)} - x_*^{(i_t)}\rangle + \tfrac{a_{i+1}^2}{L_{i_t}^{1-\alpha}B_{t+1}^2\pi_\beta^2[i_t]}\left(\|\nabla_{i_t} f(y_t)\|^*_{i_t}\right)^2.$$

Since $\|w_t - x_*\|^2_{1-\alpha} \leq (1-\beta_t)r_t^2 + \beta_t\|y_t - x_*\|^2_{1-\alpha}$, we can continue as folows:

$$\begin{aligned} B_{t+1}r_{t+1}^2 &\overset{(2.6)}{\leq} B_t r_t^2 + \beta_t B_{t+1}\|y_t - x_*\|^2_{1-\alpha} - \tfrac{2a_{t+1}}{\pi_\beta[i_t]}\langle \nabla_{i_t} f(y_t), w_t^{(i_t)} - x_*^{(i_t)}\rangle \\ &\quad + \tfrac{2a_{i+1}^2 L_{i_t}^\alpha}{B_{t+1}\pi_\beta^2[i_t]}(f(y_t) - f(y_t + \tfrac{1}{L_{i_t}}U_{i_t}h^{(i_t)}(y_t)) \\ &\overset{(2.7)}{=} B_t r_t^2 + \beta_t B_{t+1}\|y_t - x_*\|^2_{1-\alpha} - \tfrac{2a_{t+1}}{\pi_\beta[i_t]}\langle \nabla_{i_t} f(y_t), w_t^{(i_t)} - x_*^{(i_t)}\rangle \\ &\quad + 2\tfrac{a_{t+1}^2}{B_{t+1}}S_\beta^2(f(y_t) - f(y_t + \tfrac{1}{L_{i_t}}U_{i_t}h^{(i_t)}(y_t)). \end{aligned}$$

Note that $E_{i_t} f(x_{t+1}) = \sum\limits_{i=1}^n \pi_\beta[i] f(y_t + \frac{1}{L_i}U_i h^{(i)}(y_t))$. Therefore, taking expectation of the above inequality in random variable i_t, we obtain

$$\begin{aligned} E_{i_t}(B_{t+1}r_{t+1}^2) \;\leq\; & B_t r_t^2 + a_{t+1}\sigma_{1-\alpha}\|y_t - x_*\|^2_{1-\alpha} + 2a_{t+1}\langle \nabla f(y_t), x_* - w_t\rangle \\ & + 2\tfrac{a_{t+1}^2}{B_{t+1}}S_\beta^2(f(y_t) - E_{i_t}(f(x_{t+1}))). \end{aligned} \tag{2.12}$$

Since $w_t \overset{(2.9)}{=} y_t + \frac{1-\alpha_t}{\alpha_t}(y_t - x_t)$, we obtain

$$a_{t+1}\langle \nabla f(y_t), x_* - w_t\rangle \;=\; a_{t+1}\langle \nabla f(y_t), x_* - y_t + \tfrac{1-\alpha_t}{\alpha_t}(x_t - y_t)\rangle$$

$$\overset{(2.2)}{\leq} a_{t+1}(f(x_*) - f(y_t)) - \tfrac{1}{2}a_{t+1}\sigma_{1-\alpha}\|y_t - x_*\|^2_{1-\alpha} + a_{t+1}\tfrac{1-\alpha_t}{\alpha_t}(f(x_t) - f(y_t))$$

$$\overset{(2.8)_2}{=} a_{t+1}f(x_*) - A_{t+1}f(y_t) + A_t f(x_t) - \tfrac{1}{2}a_{t+1}\sigma_{1-\alpha}\|y_t - x_*\|^2_{1-\alpha}.$$

Substituting this inequality in (2.12), we obtain

$$E_{i_t}(B_{t+1}r_{t+1}^2) \;\leq\; B_t r_t^2 + 2A_t(f(x_t) - f(x_*)) - 2A_{t+1}(E_{i_t}(f(x_{t+1}) - f(x_*)).$$

It remains to take the expectation in $\mathcal{I}_{t-1}$ and sum up all previous inequalities. We obtain

$$2A_t E_{\mathcal{I}_{t-1}}(f(x_t) - f(x_*)) + B_t E_{\mathcal{I}_{t-1}}(r_t^2) \;\leq\; r_0^2 \;=\; \|x_0 - x_*\|_{[1-\alpha]}^2.$$

Let us estimate now the growth of coefficients A_t and B_t. Note that $B_t = 1 + \sigma_{1-\alpha}A_t$. Therefore, equation for finding parameter a_{t+1} in method (2.8) looks as follows:

$$(A_{t+1} - A_t)^2 S_\beta^2 = A_{t+1}(1 + \sigma_{1-\alpha}A_{t+1}).$$

Denote $C_t = \sigma_{1-\alpha}^{1/2} A_t^{1/2}$, $t \geq 0$. Then

$$\sigma_{1-\alpha}^{-1} C_{t+1}^2(1 + C_{t+1}^2) \;=\; \sigma_{1-\alpha}^{-2} S_\beta^2 (C_{t+1}^2 - C_t^2)^2 \;\leq\; 4\sigma_{1-\alpha}^{-2} S_\beta^2 (C_{t+1} - C_t)^2 C_{t+1}^2.$$

Thus, $C_{t+1} - C_t \geq \gamma(1 + C_{t+1}^2)^{1/2} \geq C_t + \gamma(1 + C_t^2)^{1/2}$ with $\gamma = \frac{\sigma_{1-\alpha}^{1/2}}{2S_\beta}$. Now, by induction we can easily check that $C_t \geq \frac{1}{2}\left[(1+\gamma)^t - (1-\gamma)^t\right] \geq \gamma t$ for $t \geq 0$. Indeed, in this case,

$$1 + C_t^2 \;\geq\; 1 + \tfrac{1}{4}(1+\gamma)^{2t} + \tfrac{1}{4}(1-\gamma)^{2t} - \tfrac{1}{2}(1-\gamma^2)^t \;\geq\; \tfrac{1}{4}\left[(1+\gamma)^t + (1-\gamma)^t\right]^2.$$

Hence,

$$\begin{aligned} C_{t+1} \;&\geq\; \tfrac{1}{2}\left[(1+\gamma)^t - (1-\gamma)^t\right] + \tfrac{\gamma}{2}\left[(1+\gamma)^t + (1-\gamma)^t\right] \\ &=\; \tfrac{1}{2}\left[(1+\gamma)^{t+1} + (1-\gamma)^{t+1}\right]. \end{aligned}$$

Thus, $A_t \geq \frac{1}{4\sigma_{1-\alpha}}\left[(1+\gamma)^t - (1-\gamma)^t\right]^2 \geq \frac{1}{4S_\beta^2}t^2$, and

$$B_t \;=\; 1 + \sigma_{1-\alpha}A_t \;\geq\; 1 + \tfrac{1}{4}\left[(1+\gamma)^t - (1-\gamma)^t\right]^2 \;\geq\; \tfrac{1}{4}\left[(1+\gamma)^t + (1-\gamma)^t\right]^2.$$

□

Note that method (2.8) and its efficiency bounds (2.10), (2.11) are continuous in the convexity parameter $\sigma_{1-\alpha}$. As $\sigma_{1-\alpha} \to 0$, we get a monotone decrease of values B_t to one, and values A_t go to their lower bounds $\frac{t^2}{4S_{\alpha/2}^2}$.

Remark 1 *The first coordinate descent version of method (2.8) with $\alpha = 0$ (uniform distribution) was suggested in [7]. In [2], this method was extended onto arbitrary values of $\alpha \in [0, 1]$. However, in [2] the authors used another random strategies ($\pi_i = L_i^\alpha / S_\alpha$). As a result, they get weaker complexity bounds. Indeed, in order to solve problem (2.1) with accuracy ϵ, they need $O\left(\frac{\sqrt{nS_\alpha}}{\sigma_{1-\alpha}^{1/2}} \ln\frac{1}{\epsilon}\right)$ iterations (see Theorem 4 in [2]). Our method requires $O\left(\frac{S_{\alpha/2}}{\sigma_{1-\alpha}^{1/2}} \ln\frac{1}{\epsilon}\right)$ iterations. It is easy to see that we always have*

$$\sqrt{nS_\alpha} \;\geq\; S_{\alpha/2},$$

and sometimes the gain can reach a factor of order $\sqrt{n}$. We give the corresponding examples in Section 3.

3 Examples of applications

3.1 Favorable structure of objective function

Let us compare now the complexity bounds of the Accelerated Coordinate Descent Method (2.8) with complexity bounds of the standard Fast Gradient Methods (e.g. [6]). For the sake of simplicity, we assume that in problem (2.1) we have $\dim \mathbb{E}^{(i)} = 1$, $i \in \{1:n\}$. Thus, $\dim E = N \equiv n$. Moreover, let us assume that the objective function in (2.1) is twice continuously differentiable. Therefore,

$$L_i(f) \;=\; \sup_{x \in E} \langle \nabla^2 f(x) e_i, e_i \rangle, \quad x \in E,\ i \in \{1:n\}, \tag{3.1}$$

where e_i is the ith coordinate vector in $\mathbb{E}$.

Let us define also the Lipschitz constant for the gradient of objective function in (2.1):

$$L(f) \;=\; \sup_{x \in \mathbb{E}} \max_{\|h\| \le 1} \langle \nabla^2 f(x) h, h \rangle. \tag{3.2}$$

Assuming that $\|e_i\| \le 1$ for all $i = 1, \dots, n$, we clearly have $L_i(f) \le L(f)$, $i \in \{1:n\}$.

For our comparison, let us choose $\alpha = 1$. Then all distances in $\mathbb{E} \equiv \mathbb{R}^n$ are measured in the standard Euclidean norm, which does not depend on the Lipschitz constants for the derivatives. For the sake of notation, denote $\|\cdot\| \equiv \|\cdot\|_{[0]}$. Denote $R = \|x_0 - x_*\|$ and let us assume that $\sigma_0 = 0$ (no strong convexity).

In this situation, fast gradient methods solve problem (2.1) up to accuracy ϵ in $O\left(I_{FGM} \stackrel{\text{def}}{=} \frac{L^{1/2}(f)}{\epsilon^{1/2}} R\right)$ iterations (e.g. [8]). At each iteration, they need to update n-dimensional vectors and to call oracle (a constant number of times). Denoting the corresponding computational expenses by T_{FGM}, we get the following bound for total computational cost:

$$C_{FGM} \;=\; I_{FGM} \cdot T_{FGM} \;=\; \frac{L^{1/2}(f)}{\epsilon^{1/2}} R \cdot T_{FGM}.$$

Similarly, in view of Theorem 1, for solving problem (2.1) up to accuracy ϵ, method (2.8) needs $O\left(I_{ACDM} \stackrel{\text{def}}{=} \frac{S_{1/2}}{\epsilon^{1/2}} R\right)$ iterations. Thus, its total computational cost is

$$C_{ACDM} \;=\; I_{ACDM} \cdot T_{ACDM} \;=\; \frac{S_{1/2}}{\epsilon^{1/2}} R \cdot T_{ACDM}.$$

Note that $S_{1/2} \le n L^{1/2}(f)$. Therefore, in order to ensure $C_{ACDM} \le C_{FGM}$, we need to find problems, for which $T_{ACDM} \le \frac{1}{n} T_{FGM}$.

Let the objective function f in problem (2.1) has the following structure:

$$f(x) \;=\; F(Ax, x), \tag{3.3}$$

where $F(s, x) : \mathbb{R}^{m+n} \to \mathbb{R}$ is a convex differentiable function, and A is an $m \times n$-matrix. Our main structural assumption on function F is that the complexity T_F of its first-order oracle is linear:

$$T_F = O(m + n). \tag{3.4}$$

This time is required for computing the function value $F(s,x)$ and the gradient

$$\nabla F(s,x) = (\nabla_s F(s,x), \nabla_x F(s,x)) \in \mathbb{R}^m \times \mathbb{R}^n.$$

Note that $\nabla f(x) = \nabla_x F(Ax,x) + A^T \nabla_s F(Ax,x)$. Let us estimate now the complexity of one iteration of our methods, assuming that matrix A is dense and

$$m \geq O(n). \tag{3.5}$$

For Fast Gradient Method, the most expensive computation at each iteration is the call of oracle. In accordance of our assumptions, computation of the function value and the gradient needs $O(mn)$ arithmetic operations. All other costs (update of n-dimensional vectors, computation of scalar products, etc.) need $O(m+n)$ operations. Thus, we conclude that

$$T_{FGM} = O(mn). \tag{3.6}$$

For ACD-method (2.8), at each iteration we need to know only the value of directional derivative $\nabla_{i_t} f(y_t)$. If the vector Ay_t is already computed, this needs $O(m+n)$ operations. Therefore, during the process (2.8) we need to *update recursively* these vectors. For this, we need to update also the products Ax_t, Av_t, and Aw_t. These operations need just computation of convex combinations of some already computed vectors with the cost $O(n)$. Only two operations for computing Ax_{t+1} and Av_{t+1} need addition of i_tth column of matrix A with some factors, and their cost is $O(m)$. Thus, we conclude that in our case

$$T_{ACDM} \;=\; O(m+n) \;\overset{(3.5)}{\leq}\; \tfrac{1}{n} T_{FGM}. \tag{3.7}$$

Hence, for all optimization problem (2.1) with above structure we have $C_{ACDM} \leq C_{FGM}$.

In the next two parts of this section we give examples of objective functions, for which ACD-method (2.8) can outperform the standard schemes by a dimensionally dependent factor. For these examples, we can guarantee that $L_i(f) << L(f)$, $i \in \{1:n\}$.

3.2 Unconstrained minimization of quadratic function

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive-definite matrix, and $F(s,x) = \frac{1}{2}\langle s,x\rangle - \langle b,x\rangle$. Then, all structural assumptions of Section 3.1 are satisfied, and we conclude that for problem

$$\min_{x \in \mathbb{R}^n} [f(x) = \tfrac{1}{2}\langle Ax,x\rangle - \langle b,x\rangle] \tag{3.8}$$

we have $C_{ACDM} \leq C_{FGM}$.

Let us assume now that matrix A has positive elements, which have same order of magnitude:

$$0 \;<\; \kappa_1 \;\leq\; A^{(i,j)} \;\leq\; \kappa_2, \quad i,j \in \{1:n\}, \tag{3.9}$$

and $\kappa_2 \leq O(\kappa_1)$. Then,

$$S_{1/2} \;\leq n\kappa_2^{1/2}. \tag{3.10}$$

On the other hand,

$$L(f) \;=\; \lambda_{\max}(A) \;\geq\; \kappa_1 \lambda_{\max}(1_n 1_n^T) \;=\; n\kappa_1, \tag{3.11}$$

where $1_n \in \mathbb{R}^n$ is the vector of all ones. This implies that

$$S_{1/2} \quad \leq \quad \sqrt{\tfrac{n\kappa_2}{\kappa_1}} \cdot L^{1/2}(f).$$

In other words, assumption (3.9) implies $C_{ACDM} \leq O\left(\frac{1}{n^{1/2}}\right) C_{FGM}$.

3.3 Smoothing Technique

Smoothing technique [6] can be applied to objective functions with sufficiently simple dual representation:

$$f(x) \quad = \quad \max_{u \in Q} \{\langle Ax, u\rangle - \phi(u)\}, \tag{3.12}$$

where $Q \subset \mathbb{R}^m$ is a closed convex bounded set, and function ϕ is convex on Q. Let us measure distances in $\mathbb{R}^m$ by some norm $\|\cdot\|_X$. We assume that

$$\|e_i\|_X \quad \leq \quad 1, \quad i = 1, \ldots, n, \tag{3.13}$$

where e_i is ith coordinate vector in $\mathbb{R}^n$.

Function f defined by (3.12) is typically nonsmooth. However, optimization problem in (3.12) must be simple enough since we assume it solvable in a closed form (otherwise, the value $f(x)$ is not computable). In this situation, it is often possible to approximate f by a convex function with Lipschitz continuous gradient.

Indeed, let prox-function $d(u)$ be differentiable and strongly convex on Q in some norm $\|\cdot\|_U$ with convexity parameter one:

$$\langle \nabla d(u_1) - \nabla d(u_2), u_1 - u_2\rangle \quad \geq \quad \|u_1 - u_2\|_U^2, \quad u_1, u_2 \in U. \tag{3.14}$$

Assume that $d(u) \geq 0$ for all $u \in Q$ and $d(u_0) = 0$ at some prox-center $u_0 \in Q$.

Denote

$$f_\mu(x) \quad = \quad \max_{u \in Q} \{\langle Ax, u\rangle - \phi(u) - \mu d(u)\}, \tag{3.15}$$

where $\mu > 0$ is the smoothness parameter. Then f_μ approximates f with accuracy $O(\mu)$, and its gradient is Lipschitz continuous with constant $L(f_\mu) = \frac{1}{\mu}\|A\|^2$, where

$$\|A\| \quad = \quad \max_{x,u} \{\langle Ax, u\rangle : \ \|x\|_X \leq 1, \ \|u\|_U \leq 1\}.$$

Note that $\|A\| \overset{(3.13)}{\geq} \max_u \{\langle Ae_i, u\rangle : \ \|u\|_U\} = \|Ae_i\|_U^*$ for all $i = 1, \ldots, n$. Therefore,

$$\sum_{i=1}^n \|Ae_i\|_U^* \quad \leq \quad n\|A\|. \tag{3.16}$$

Recall that the gradient of function f_μ is defined as

$$\nabla f_\mu(x) \quad = \quad A^T u_\mu(x), \tag{3.17}$$

where $u_\mu(x)$ is the unique solution of the optimization problem in definition (3.15).

Let us justify now the bounds for $L_i(f_\mu)$, $i \in \{1:n\}$. Consider two points x_1 and $x_2 = x_1 + h$, where h is an arbitrary direction in $\mathbb{R}^n$. Denote $u_i = u_\mu(x_i)$, $i = 1, 2$. From the optimality conditions for optimization problem in (3.15), we have

$$\begin{aligned} \langle Ax_1 - \nabla\phi(u_1) - \mu\nabla d(u_1), u_2 - u_1\rangle &\le 0, \\ \langle Ax_2 - \nabla\phi(u_2) - \mu\nabla d(u_2), u_1 - u_2\rangle &\le 0. \end{aligned}$$

Adding these two inequalities, we get

$$\begin{aligned} \mu\|u_1 - u_2\|_U^2 &\overset{(3.14)}{\le} \mu\langle\nabla d(u_1) - \nabla d(u_2), u_1 - u_2\rangle \\ &\le \langle Ax_1 - Ax_2 - (\nabla\phi(u_1) - \nabla\phi(u_2)), u_1 - u_2\rangle \\ &\le \langle Ah, u_2 - u_1\rangle. \end{aligned}$$

Taking now $h = \tau e_i$, where e_i is the ith coordinate vector in $\mathbb{R}^n$, we obtain:

$$\tau\left(\nabla_i f_\mu(x_2) - \nabla_i f_\mu(x_1)\right) \overset{(3.17)}{=} \tau\langle e_i, A^T(u_2 - u_1)\rangle \ge \mu\|u_1 - u_2\|_U^2$$

$$\ge \tfrac{\mu}{(\|Ae_i\|_U^*)^2}\langle Ae_i, u_1 - u_2\rangle^2 = \tfrac{\mu}{(\|Ae_i\|_U^*)^2}\left(\nabla_i f_\mu(x_1) - \nabla_i f_\mu(x_2)\right)^2.$$

Thus, we can take $L_i(f_\mu) = \frac{1}{\mu}(\|Ae_i\|_U^*)^2$, $i \in \{1:n\}$. Consequently,

$$\sum_{i=1}^n L_i^{1/2}(f_\mu) \overset{(3.16)}{\le} nL^{1/2}(f_\mu). \tag{3.18}$$

If the set Q and function ϕ in (3.15) are simple, then f_μ satisfies all conditions of Section 3.1 (in particular, with known product Ax, vector $u_\mu(x)$ is computable in $O(m)$ operations). Therefore, for its unconstrained minimization, efficiency estimates of ACD-method (2.8) are always not worse than the bounds of any Fast Gradient Method.

Let us present an example, where ACD-method (2.8) is much better than FGM (since $L_i(f_\mu) << L(f_\mu)$, $i \in \{1:n\}$). Assume that all elements of matrix A are positive and have the same order of magnitude:

$$0 < \kappa_1 \le A^{(i,j)} \le \kappa_2, \quad i \in \{1:m\},\ j \in \{1:n\}, \tag{3.19}$$

and $\kappa_2 \le O(\kappa_1)$. Then, clearly $L_i(f_\mu) \le \frac{m}{\mu}\kappa_2^2$. Therefore, $S_{1/2} \le n\kappa_2\left(\frac{m}{\mu}\right)^{1/2}$.

On the other hand,

$$L(f_\mu) = \tfrac{1}{\mu}\lambda_{\max}(A^TA) \ge \tfrac{1}{\mu}\kappa_1^2 m\lambda_{\max}(1_n 1_n^T) = \tfrac{1}{\mu}\kappa_1^2 mn.$$

Thus, comparing the bounds

$$C_{ACDM} = O\left(m \cdot \tfrac{S_{1/2}R}{\epsilon^{1/2}}\right) \le O\left(nm^{3/2} \cdot \tfrac{R}{\mu^{1/2}\epsilon^{1/2}}\right),$$

and

$$C_{FGM} = O\left(mn \cdot \tfrac{L^{1/2}(f_\mu)R}{\epsilon^{1/2}}\right) \ge O\left(n^{3/2}m^{3/2} \cdot \tfrac{R}{\mu^{1/2}\epsilon^{1/2}}\right),$$

we can see that the bound for ACD-method (2.8) is at least in $O\left(n^{1/2}\right)$ times better.

4 Preliminary computational experiments

In our computational experiments, we solved the following problem with randomly generated data:

$$\min_{x \in \mathbb{R}^M} \left\{ f_\mu(x) = \sum_{i=1}^{N} \phi_\mu \left(\langle a_i, x \rangle - c^{(i)} \right) \right\}, \tag{4.1}$$

where

$$\phi_\mu(\tau) \;=\; \begin{cases} \frac{\tau^2}{2\mu}, & \text{if } |\tau| \le \mu, \\ |\tau| - \frac{1}{2}\mu, & \text{if } \tau > \mu. \end{cases}$$

Coefficients of dense vectors a_i, $i = 1, \dots, N$, are uniformly distributed in the interval $[1, 2]$. Coefficients of vector $c = (c^{(1)}, \dots, c^{(N)})^T \in \mathbb{R}^N$ are chosen as $c^{(i)} = \langle a_i, \bar{y} \rangle$, where the entries of vector $\bar{y} \in \mathbb{R}^M$ are uniformly distributed in the interval $[-1, 1]$.

Thus, the optimal value of function f_μ is zero. Therefore, for all methods we use the termination criterion $f_\mu(x) \le \epsilon$ with $\epsilon = 10^{-2}$. We choose also $\mu = \epsilon$.

Among numerous variants of Fast Gradient Methods, we choose the method with the maximal adaptivity to the unknown Lipschitz constant for the gradient of objective function. Its scheme is as follows.

FGM: Choose $x_0 \in \mathbb{E}$ and $L_0 > 0$. Set $v_0 = x_0$.

For $t \ge 0$ iterate:

1) Find the smallest $i_t \ge 0$ such that for

$$a_{t,i_t} = \frac{1}{2^{i_t+1} L_t} \left(1 + \sqrt{1 + 2^{i_t+2} L_t A_t} \right), \quad \tau_{t,i_t} = \frac{a_{t,i_t}}{a_{t,i_t} + A_t},$$

$$y_{t,i_t} = (1 - \tau_{t,i_t}) x_t + \tau_{t,i_t} v_t, \text{ and } x_{t+1,i_t} = y_{t,i_t} - \frac{1}{2^{i_t} L_t} \nabla f(y_{t,i_t})$$

we have $f(y_{t,i_t}) - f(x_{t+1,i_t}) \ge \frac{1}{2^{i_t+1} L_t} \|\nabla f(y_{t,i_t})\|^2$.

2) Set $x_{t+1} = x_{t+1,i_t}$, $v_{t+1} = v_t - a_{t,i_t} \nabla f(y_{t,i_t})$,

$$A_{t+1} = A_t + a_{t,i_t}, \text{ and } L_{t+1} = 2^{i_t - 1} L_t.$$

(4.2)

On the contrary, for Accelerated Coordinate Descent Method (2.8) with parameters $\alpha = 1$ and $\sigma_0 = 0$, we choose the fixed worst-case estimates for the coordinate Lipschitz constants

$$L_i(f_\mu) \;=\; \frac{1}{\mu} \|A^T e_i\|^2, \quad i = 1, \dots, M, \tag{4.3}$$

where $A = (a_1, \dots, a_N)$ and the norm is standard Euclidean. Since we take $\beta \equiv \alpha/2 = \frac{1}{2}$, we get the following distribution of probabilities:

$$\pi_{1/2}[i] \;=\; \frac{\|A^T e_i\|}{\sum\limits_{k=1}^{N} \|A^T e_k\|}, \quad i = 1, \dots, M. \tag{4.4}$$

At the same time, $S_{1/2}^2 = \frac{1}{\mu}\left(\sum\limits_{i=1}^{M} \|A^T e_i\|\right)^2$.

In all our experiments we use the staring point $x_0 = 0 \in \mathbb{R}^M$. In the method below, notation Ax (or, Ay, Av) is used for the value of the linear operator in (4.1), computed at point $x \in \mathbb{R}^M$:

$$Ax \quad \equiv \quad A^T x - c \in \mathbb{R}^N.$$

The scheme of ACD-method for problem (4.1) looks as follows.

ACDM for (4.1): Define $v_0 = x_0 = 0 \in \mathbb{R}^M$, $Av_0 = Ax_0 = -c$, and $A_0 = 0$.

For $t \geq 0$, iterate:

1) Find parameter $a_{t+1} > 0$ from equation $a_{t+1}^2 S_\beta^2 = A_{t+1} + a_{t+1}$.

   Set $A_{t+1} = A_t + a_{t+1}$, and $\tau_t = \frac{a_{t+1}}{A_{t+1}}$.

2) Define $y_t = (1-\tau_t)x_t + \tau_t v_t$. Update $Ay_t = (1-\tau_t)Ax_t + \tau_t Av_t$.

3) Choose i_t in accordance to distribution (4.4) and compute $\nabla_{i_t} f(y_t)$.

4) Update $x_{t+1} = y_t - \frac{1}{L_{i_t}} \nabla_{i_t} f(y_t) e_{i_t}$, $Ax_{t+1} = Ay_t - \frac{1}{L_{i_t}} \nabla_{i_t} f(y_t) A^T e_{i_t}$,

   $v_{t+1} = v_t - a_{t+1} \nabla_{i_t} f(y_t) e_{i_t}$, and $Av_{t+1} = Av_t - \frac{a_{t+1}}{\pi_{1/2}[i_t]} \nabla_{i_t} f(y_t) A^T e_{i_t}$.

(4.5)

Note that the computational cost of all operations in the above method, including the computation of directional derivative $\nabla_i f(y_t) = \langle A^T e_i, \nabla f(y_t) \rangle$, is *linear* in the dimensions of problem (4.1).

In view of its adaptivity, in our experiments, FGM is a priori in a much better position

than ACDM. Nevertheless, the computational results are as follows.

		FGM			ACDM	
N	M	IT	NF	Time (sec)	IT/M	Time (sec)
100	50	4727	18916	0.547	2024	0.578
50	100	4889	19566	0.578	2305	0.672
200	100	11244	44986	4.750	3700	4.000
100	200	12859	51450	5.250	3750	4.203
400	200	25473	101902	40.234	5495	23.125
200	400	26184	104750	40.719	6345	30.157
800	400	55511	222056	358.234	8789	302.203
400	800	61994	247992	397.656	11461	245.657
1600	800	122542	490184	3185.953	13899	1652.733
800	1600	126748	507008	3213.156	19139	2360.719

Table 1. Performance of FGM and ACDM on problem (4.1).

In this table, first two columns display the dimensions of problem (4.1). In all our tests, the matrix A is dense. Therefore, for the largest problem we have more than one million nonzero coefficients. Columns IT and NF show the number of iterations and number of function evaluation of FGM. Column IT/M shows the number of blocks of M iterations in method ACDM. Finally, the column Time displays the total computational time in seconds.

For us, the main characteristics of complexity of the problem for numerical scheme is the total computational time. As we can see, ACDM always outperforms FGM. Its domination is less impressive with respect to the theoretical prediction. However, this can be explained by the ability of method (4.2) to use much smaller estimate of the constant $L(f_\mu)$ than the worst-case theoretical value.

To conclude, we can see that potentially, ACDM is a promising computational scheme, which has good chances to outperform FGM on many important real-life problems. At this moment, as compared with FGM, ACDM has four main drawbacks:

- absence of version with separable constraints;
- impossibility to adjust the worst-case estimates for $L_i(f)$ during the minimization process;
- absence of a reliable stopping criterion;
- impossibility to generate good primal-dual solutions.

In our opinion, any advancement in these directions will be very interesting.

References

[1] Z. Allen-Zhu, Z. Qu, P. Richtárik, and Y. Yuan. Even faster accelerated coordinate descent using non-uniform sampling. arXiv:1512.09103v2 [math.OC] 13 Jan 2016.

[2] Y.T. Lee and A. Sidford. Efficient accelerated coordinate descent methods and fastest algorithms for solving linear systems. arXiv: 1305.1922v1, 8 May, 2013.

[3] Yu. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(\frac{1}{k^2})$. *Doklady AN SSSR* (translated as Soviet Math. Docl.), 1983, v.269, No. 3, 543-547

[4] P. Richtarik and M. Takač. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 1-51 (2012)

[5] P. Richtarik and M. Takač. Distributed coordinate descent method for learning with big data. arXiv: 1310.2059 (2012)

[6] Yu. Nesterov. Smooth minimization of non-smooth functions, *Mathematical Programming* (A), **103** (1), 127-152 (2005).

[7] Yu.Nesterov. Efficiency of coordinate-descent methods on huge-scale optimization problems. *SIOPT*, **22**(2), 341-362 (2012).

[8] Yu. Nesterov. Introductory lectures on Convex Optimization. The basic course. *Kluwer*, Boston (2004).

Recent titles

CORE Discussion Papers

2015/24 Wing Man Wynne LAM. Switching costs in two-sided markets.

2015/25 Philippe DE DONDER, Marie-Louise LEROUX. The political choice of social long term care transfers when family gives time and money.

2015/26 Pierre PESTIEAU and Gregory PONTHIERE. Long-term care and births timing.

2015/27 Pierre PESTIEAU and Gregory PONTHIERE. Longevitiy variations and the welfare State.

2015/28 Mattéo GODIN and Jean HINDRIKS. A review of critical issues on tax design and tax administration in a global economy and developing countries

2015/29 Michel MOUCHART, Guillaume WUNSCH and Federica RUSSO. The issue of control in multivariate systems, A contribution of structural modelling.

2015/30 Jean J. GABSZEWICZ, Marco A. MARINI and Ornella TAROLA. Alliance formation in a vertically differentiated market.

2015/31 Jens Leth HOUGAARD, Juan D. MORENO-TERNERO, Mich TVEDE and Lars Peter ØSTERDAL. Sharing the proceeds from a hierarchical venture.

2015/32 Arnaud DUFAYS and Jeroen V.K. ROMBOUTS. Spare change-point time series models.

2015/33 Wing Man Wynne LAM. Status in organizations.

2015/34 Wing Man Wynne LAM. Competiton in the market for flexible resources : an application to cloud computing.

2015/35 Yurii NESTEROV and Vladimir SHIKHMAN. Computation of Fisher-Gale equilibrium by auction.

2015/36 Maurice QUEYRANNE and Laurence A. WOLSEY. Thight MIP formulations for bounded up/down tiems and interval-dependent start-ups.

2015/37 Paul BELLEFLAMME and Dimitri PAOLINI. Strategic promotion and release decisions for cultural goods.

2015/38 Nguyen Thang DAO and Julio DAVILA. Gender inequality, technologial progress, and the demographic transition.

2015/39 Thomas DEMUYNCK, Bram DE ROCK and Victor GINSBURGH. The transfer paradox in welfare space.

2015/40 Pierre DEHEZ. On Harsanyi dividends and asymmetric values.

2015/41 Laurence A. WOLSEY. Uncapacitated lot-sizing with stock upper bounds, stock fixed costs, stock overloads and backlogging: A tight formulation.

2015/42 Paul BELLEFLAMME. Monopoly price discrimination and privacy: the hidden cost of hiding.

2015/43 Pierre PESTIEAU and Gregory PONTHIERE. Optimal fertility under age-dependent labor productivity.

2015/44 Jacques DREZE. Subjective expected utility with state-dependent but action/observation-independent preferences

2015/45 Joniada MILLA, Ernesto SAN MARTÍN and Sébastien VAN BELLEGEM. Higher education value added using multiple outcomes.

2015/46 Helmuth CREMER, Pierre PESTIEAU and Kerstin ROEDER. Social long-term care insurance with two-sided altruism.

2015/47 Per J. AGRELL and Humberto BREA-SOLÍS. Stationarity of heterogeneity in production technology using latent class modelling.

2015/48 Mattéo GODIN et Jean HINDRIKS. Disparités et convergence économiques : Ensemble mais différents.

2015/49 Maurice QUEYRANNE and Laurence A. WOLSEY. Modeling poset convex subsets.

Recent titles
CORE Discussion Papers – continued

2015/50 Benoît DECERF. A new index combining the absolute and relative aspects of income poverty: Theory and application.

2015/51 Pierre COPÉE, Axel GAUTIER and Mélanie LEFÈVRE. Promoting competition at the digital age with an application to Belgium.

2015/52 Mathieu LEFEBVRE, Sergio PERELMAN and Pierre PESTIEAU. Productivity and performance in the public sector.

2015/53 Oswaldo GRESSANI. Endogeneous quantal response equilibrium for normal form games.

2015/54 Wouter VERGOTE. One-to-one matching problems with location restrictions.

2015/55 Marie-Louise LEROUX, Dario MALDONADO and Pierre PESTIEAU. Compliance, informality and contributive pensions.

2015/56 Michel MOUCHART and Renzo ORSI. Building a structural model : Parameterization and structurality.

2016/01 Luc BAUWENS, Manuela BRAIONE, Giuseppe STORTI. A dynamic component model for forecasting high-dimensional realized covariance matrices.

2016/02 Manuela BRAIONE. A time-varying long run HEAVY model.

2016/03 Yurii NESTEROV and Sebastian STICH. Efficiency of accelerated coordinate descent method on structured optimization problems.

Books

W. GAERTNER and E. SCHOKKAERT (2012), *Empirical Social Choice*. Cambridge University Press.

L. BAUWENS, Ch. HAFNER and S. LAURENT (2012), *Handbook of Volatility Models and their Applications*. Wiley.

J-C. PRAGER and J. THISSE (2012), *Economic Geography and the Unequal Development of Regions*. Routledge.

M. FLEURBAEY and F. MANIQUET (2012), *Equality of Opportunity: The Economics of Responsibility*. World Scientific.

J. HINDRIKS (2012), *Gestion publique*. De Boeck.

M. FUJITA and J.F. THISSE (2013), *Economics of Agglomeration: Cities, Industrial Location, and Globalization*. (2nd edition). Cambridge University Press.

J. HINDRIKS and G.D. MYLES (2013). *Intermediate Public Economics*. (2nd edition). MIT Press.

J. HINDRIKS, G.D. MYLES and N. HASHIMZADE (2013). *Solutions Manual to Accompany Intermediate Public Economics*. (2nd edition). MIT Press.

J. HINDRIKS (2015). *Quel avenir pour nos pensions ? Les grands défis de la réforme des pensions*. De Boeck.

P. BELLEFLAMME and M. PEITZ (2015). *Industrial Organization: Markets and Strategies* (2nd edition). Cambridge University Press.

CORE Lecture Series

R. AMIR (2002), Supermodularity and Complementarity in Economics.

R. WEISMANTEL (2006), Lectures on Mixed Nonlinear Programming.

A. SHAPIRO (2010), Stochastic Programming: Modeling and Theory.