Randomized Similar Triangles Method: A Unifying Framework for Accelerated Randomized Optimization Methods (Coordinate Descent, Directional Search, Derivative-Free Method)

Pavel Dvurechensky * Alexander Gasnikov † Alexander Tiurin ‡

July 26, 2017§

Abstract

In this paper, we consider smooth convex optimization problems with simple constraints and inexactness in the oracle information such as value, partial or directional derivatives of the objective function. We introduce a unifying framework, which allows to construct different types of accelerated randomized methods for such problems and to prove convergence rate theorems for them. We focus on accelerated random block-coordinate descent, accelerated random directional search, accelerated random derivative-free method and, using our framework, provide their versions for problems with inexact oracle information. Our contribution also includes accelerated random block-coordinate descent with inexact oracle and entropy proximal setup as well as derivative-free version of this method.

Keywords: convex optimization, accelerated random block-coordinate descent, accelerated random directional search, accelerated random derivative-free method, inexact oracle, complexity, accelerated gradient descent methods, first-order methods, zero-order methods.
AMS Classification: 90C25, 90C30, 90C06, 90C56, 68Q25, 65K05, 49M27, 68W20, 65Y20, 68W40

*Weierstrass Institute for Applied Analysis and Stochastics, Berlin; Institute for Information Transmission Problems RAS, Moscow, pavel.dvurechensky@wias-berlin.de

†Moscow Institute of Physics and Technology, Moscow; Institute for Information Transmission Problems RAS, Moscow, gasnikov@yandex.ru

‡National Research University Higher School of Economics, Moscow, alexandertiurin@gmail.com

§The results obtained in this paper were presented in December, 2016 (http://www.mathnet.ru/php/seminars.phtml?option_lang=rus&presentid=16180) and in June, 2017 (http://www.lccc.lth.se/index.php?mact=ReglerSeminars,cntnt01,abstractbio,0&cntnt01abstractID=889&cntnt01returnid=116)

Introduction

In this paper, we consider smooth convex optimization problems with simple constraints and inexactness in the oracle information such as value, partial or directional derivatives of the objective function. Different types of randomized optimization algorithms, such as random coordinate descent or stochastic gradient descent for empirical risk minimization problem, have been extensively studied in the past decade with main application being convex optimization problems. Our main focus in this paper is on accelerated randomized methods: random block-coordinate descent, random directional search, random derivative-free method. As opposed to non-accelerated methods, these methods have complexity $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ iterations to achieve objective function residual ε. Accelerated random block-coordinate descent method was first proposed in Nesterov [2012], which was the starting point for active research in this direction. The idea of the method is, on each iteration, to randomly choose a block of coordinates in the decision variable and make a step using the derivative of the objective function with respect to the chosen coordinates. Accelerated random directional search and accelerated random derivative-free method were first proposed in 2011 and published recently in Nesterov and Spokoiny [2017], but there was no extensive research in this direction. The idea of random directional search is to use a projection of the objective's gradient onto a randomly chosen direction to make a step on each iteration. Random derivative-free method uses the same idea, but random projection of the gradient is approximated by finite-difference, i.e. the difference of values of the objective function in two close points. This also means that it is a zero-order method which uses only function values to make a step.

Existing accelerated randomized methods have different convergence analysis. This motivated us to pose the main question, we address in this paper, as follows. *Is it possible to find a crucial part of the convergence rate analysis and use it to systematically construct new accelerated randomized methods?* To some extent, our answer is "'yes"'. We determine three main assumptions and use them to prove convergence rate theorem for our generic accelerated randomized method. Our framework allows both to reproduce known and to construct new accelerated randomized methods. The latter include new accelerated random block-coordinate descent with inexact block derivatives and entropy proximal setup.

Related Work

In the seminal paper Nesterov [2012], Nesterov proposed random block-coordinate descent for convex optimization problems with simple convex separable constraints and accelerated random block-coordinate descent for unconstrained convex optimization problems. In Lee and Sidford [2013], Lee and Sidford proposed accelerated random block-coordinate descent with non-uniform probability of choosing a particular block of coordinates. They also developed an efficient implementation without full-dimensional operations on each iteration. Fercoq and Richtárik in Fercoq and Richtárik [2015] introduced accelerated block-coordinate descent for composite optimization problems, which include problems with separable constraints. Later, Lin, Lu and Xiao in Lin et al. [2014] extended this method for strongly convex

problems. In May 2015, Nesterov and Stich presented an accelerated block-coordinate descent with complexity, which does not explicitly depend on the problem dimension. This result was recently published in Nesterov and Stich [2017]. Similar complexity was obtained also by Allen-Zhu, Qu, Richtárik and Yuan in Allen-Zhu et al. [2016] and by Gasnikov, Dvurechensky and Usmanova in Gasnikov et al. [2016c]. We also mention special type of accelerated block-coordinate descent of Shalev-Shwartz and Zhang developed in Shalev-Shwartz and Zhang [2014] for empirical risk minimization problems. All these accelerated block-coordinate descent methods work in Euclidean setup, when the norm in each block is Euclidean and defined using some positive semidefinite matrix. Non-accelerated block-coordinate methods, but with non-euclidean setup, were considered by Dang and Lan in Dang and Lan [2015]. All the mentioned methods rely on exact block derivatives and exact projection on each step. Inexact projection in the context of non-accelerated random coordinate descent was considered by Tappenden, Richtárik and Gondzio in Tappenden et al. [2016].

Research on accelerated random directional search and accelerated random derivative-free methods started in Nesterov and Spokoiny [2017]. Mostly non-accelerated derivative-free methods were further developed in the context of inexact function values in Gasnikov et al. [2016a,b], Bogolubsky et al. [2016], Gasnikov et al. [2017].

We should also mention that there are other accelerated randomized methods in Frostig et al. [2015], Lin et al. [2015], Zhang and Lin [2015], Allen-Zhu [2017], Lan and Zhou [2017]. Most of these methods were developed deliberately for empirical risk minimization problems and do not fall in the scope of this paper.

Our Approach and Contributions

Our framework has two main components, namely, Randomized Inexact Oracle and Randomized Similar Triangles Method. The starting point for the definition of our oracle is a unified view on random directional search and random block-coordinate descent. In both these methods, on each iteration, a randomized approximation for the objective function's gradient is calculated and used, instead of the true gradient, to make a step. This approximation for the gradient is constructed by a projection on a randomly chosen subspace. For random directional search, this subspace is the line along a randomly generated direction. As a result a directional derivative in this direction is calculated. For random block-coordinate descent, this subspace is given by randomly chosen block of coordinates and block derivative is calculated. One of the key features of these approximations is that they are unbiased, i.e. their expectation is equal to the true gradient. We generalize two mentioned approaches by allowing other types of random transformations of the gradient for constructing its randomized approximation.

The inexactness of our oracle is inspired by the relation between derivative-free method and directional search. In the framework of derivative-free methods, only the value of the objective function is available for use in an algorithm. At the same time, if the objective function is smooth, the directional derivative can be well approximated by the difference of function values at two points which are close to each other. Thus, in the context of zero-

order optimization, one can calculate only an inexact directional derivative. Hence, one can construct only a biased randomized approximation for the gradient when a random direction is used. We combine previously mentioned random transformations of the gradient with possible inexactness of this transformations to construct our *Randomized Inexact Oracle*, which we use in our generic algorithm to make a step on each iteration.

The basis of our generic algorithm is Similar Triangles Method of Tyurin [2017] (see also Dvurechensky et al. [2017]), which is an accelerated gradient method with only one proximal mapping on each iteration, this proximal mapping being essentially the Mirror Descent step. The notable point is that, we only need to substitute the true gradient with our Randomized Inexact Oracle and slightly change one step in the Similar Triangles Method, to obtain our generic accelerated randomized algorithm, which we call Randomized Similar Triangles Method (RSTM), see Algorithm 1. We prove convergence rate theorem for RSTM in two cases: the inexactness of Randomized Inexact Oracle can be controlled and adjusted on each iteration of the algorithm, the inexactness can not be controlled.

We apply our framework to several particular settings: random directional search, random coordinate descent, random block-coordinate descent and their combinations with derivative-free approach. As a corollary of our main theorem, we obtain both known and new results on the convergence of different accelerated randomized methods with inexact oracle.

To sum up, our contributions in this paper are as follows.

- We introduce a general framework for constructing and analyzing different types of accelerated randomized methods, such as accelerated random directional search, accelerated block-coordinate descent, accelerated derivative-free methods. Our framework allows to obtain both known and new methods and their convergence rate guarantees as a corollary of our main Theorem 1.

- Using our framework, we introduce new accelerated methods with inexact oracle, namely, accelerated random directional search, accelerated random block-coordinate descent, accelerated derivative-free method. To the best of our knowledge, such methods with inexact oracle were not known before. See Section 3.

- Based on our framework, we introduce new accelerated random block-coordinate descent with inexact oracle and non-euclidean setup, which was not done before in the literature. The main application of this method is minimization of functions on a direct product of large number of low-dimensional simplexes. See Subsection 3.3.

- We introduce new accelerated random derivative-free block-coordinate descent with inexact oracle and non-euclidean setup. Such method was not known before in the literature. Our method is similar to the method in the previous item, but uses only finite-difference approximations for block derivatives. See Subsection 3.6.

The rest of the paper is organized as follows. In Section 1, we provide the problem statement, motivate and make three our main assumptions, illustrate them by random directional search and random block-coordinate descent. In Section 2, we introduce our main

algorithm, called Randomized Similar Triangles Method, and, based on stated general assumptions, prove convergence rate Theorem 1. Section 3 is devoted to applications of our general framework for different particular settings, namely

- Accelerated Random Directional Search (Subsection 3.1),
- Accelerated Random Coordinate Descent (Subsection 3.2),
- Accelerated Random Block-Coordinate Descent (Subsection 3.3),
- Accelerated Random Derivative-Free Directional Search (Subsection 3.4),
- Accelerated Random Derivative-Free Coordinate Descent (Subsection 3.5),
- Accelerated Random Derivative-Free Block-Coordinate Descent (Subsection 3.6).
- Accelerated Random Derivative-Free Block-Coordinate Descent with Random Approximations for Block Derivatives (Subsection 3.7).

1 Preliminaries

1.1 Notation

Let finite-dimensional real vector space E be a direct product of n finite-dimensional real vector spaces E_i, $i = 1, ..., n$, i.e. $E = \otimes_{i=1}^n E_i$ and $\dim E_i = p_i$, $i = 1, ..., n$. Denote also $p = \sum_{i=1}^n p_i$. Let, for $i = 1, ..., n$, E_i^* denote the dual space for E_i. Then, the space dual to E is $E^* = \otimes_{i=1}^n E_i^*$. Given a vector $x^{(i)} \in E_i$ for some $i \in 1, ..., n$, we denote as $[x^{(i)}]_j$ its j-th coordinate, where $j \in 1, ..., p_i$. To formalize the relationship between vectors in E_i, $i = 1, ..., n$ and vectors in E, we define primal partition operators $U_i : E_i \to E$, $i = 1, ..., n$, by identity

$$x = (x^{(1)}, ..., x^{(n)}) = \sum_{i=1}^n U_i x^{(i)}, \quad x^{(i)} \in E_i, \quad i = 1, ..., n, \quad x \in E. \tag{1}$$

For any fixed $i \in 1, ..., n$, U_i maps a vector $x^{(i)} \in E_i$, to the vector $(0, ..., x^{(i)}, ..., 0) \in E$. The adjoint operator $U_i^T : E^* \to E_i^*$, then, is an operator, which, maps a vector $g = (g^{(1)}, ..., g^{(i)}, ..., g^{(n)}) \in E^*$, to the vector $g^{(i)} \in E_i^*$. Similarly, we define dual partition operators $\widetilde{U}_i : E_i^* \to E^*$, $i = 1, ..., n$, by identity

$$g = (g^{(1)}, ..., g^{(n)}) = \sum_{i=1}^n \widetilde{U}_i g^{(i)}, \quad g^{(i)} \in E_i^*, \quad i = 1, ..., n, \quad g \in E^*. \tag{2}$$

For all $i = 1, ..., n$, we denote the value of a linear function $g^{(i)} \in E_i^*$ at a point $x^{(i)} \in E_i$ by $\langle g^{(i)}, x^{(i)} \rangle_i$. We define

$$\langle g, x \rangle = \sum_{i=1}^n \langle g^{(i)}, x^{(i)} \rangle_i, \quad x \in E, \quad g \in E^*.$$

For all $i = 1, ..., n$, let $\|\cdot\|_i$ be some norm on E_i and $\|\cdot\|_{i,*}$ be the norm on E_i^* which is dual to $\|\cdot\|_i$

$$\|g^{(i)}\|_{i,*} = \max_{\|x^{(i)}\|_i \leq 1} \langle g^{(i)}, x^{(i)} \rangle_i.$$

Given parameters $\beta_i \in \mathbb{R}^n_{++}$, $i = 1, ..., n$, we define the norm of a vector $x = (x^{(1)}, ..., x^{(n)}) \in E$ as

$$\|x\|_E^2 = \sum_{i=1}^n \beta_i \|x^{(i)}\|_i^2.$$

Then, clearly, the dual norm of a vector $g = (g^{(1)}, ..., g^{(n)}) \in E^*$ is

$$\|g\|_{E,*}^2 = \sum_{i=1}^n \beta_i^{-1} \|g^{(i)}\|_i^2.$$

Throughout the paper, we consider optimization problem with feasible set Q, which is assumed to be given as $Q = \otimes_{i=1}^n Q_i \subseteq E$, where $Q_i \subseteq E_i$, $i = 1, ..., n$ are closed convex sets. To have more flexibility and be able to adapt algorithm to the structure of sets Q_i, $i = 1, ..., n$, we introduce *proximal setup*, see e.g. Ben-Tal and Nemirovski [2015]. For all $i = 1, ..., n$, we choose a *prox-function* $d_i(x^{(i)})$ which is continuous, convex on Q_i and

1. admits a continuous in $x^{(i)} \in Q_i^0$ selection of subgradients $\nabla d_i(x^{(i)})$, where $x^{(i)} \in Q_i^0 \subseteq Q_i$, and Q_i^0 is the set of all $x^{(i)}$, where $\nabla d_i(x^{(i)})$ exists;

2. is 1-strongly convex on Q_i with respect to $\|\cdot\|_i$, i.e., for any $x^{(i)} \in Q_i^0, y^{(i)} \in Q_i$, it holds that $d_i(y^{(i)}) - d_i(x^{(i)}) - \langle \nabla d_i(x^{(i)}), y^{(i)} - x^{(i)} \rangle_i \geq \frac{1}{2}\|y^{(i)} - x^{(i)}\|_i^2$.

We define also the corresponding *Bregman divergence* $V_i[z^{(i)}](x^{(i)}) := d_i(x^{(i)}) - d_i(z^{(i)}) - \langle \nabla d_i(z^{(i)}), x^{(i)} - z^{(i)} \rangle_i$, $x^{(i)} \in Q_i, z^{(i)} \in Q_i^0$, $i = 1, ..., n$. It is easy to see that $V_i[z^{(i)}](x^{(i)}) \geq \frac{1}{2}\|x^{(i)} - z^{(i)}\|_i^2, \quad x^{(i)} \in Q_i, z^{(i)} \in Q_i^0$, $i = 1, ..., n$. Standard proximal setups, e.g. Euclidean, entropy, ℓ_1/ℓ_2, simplex can be found in Ben-Tal and Nemirovski [2015]. It is easy to check that, for given parameters $\beta_i \in \mathbb{R}^n_{++}$, $i = 1, ..., n$, the functions $d(x) = \sum_{i=1}^n \beta_i d_i(x^{(i)})$ and $V[z](x) = \sum_{i=1}^n \beta_i V_i[z^{(i)}](x^{(i)})$ are respectively a prox-function and a Bregman divergence corresponding to Q. Also, clearly,

$$V[z](x) \geq \frac{1}{2}\|x - z\|_E^2, \quad x \in Q, \quad z \in Q^0 := \otimes_{i=1}^n Q_i^0. \tag{3}$$

For a differentiable function $f(x)$, we denote by $\nabla f(x) \in E^*$ its gradient.

1.2 Problem Statement and Assumptions

The main problem, we consider, is as follows

$$\min_{x \in Q \subseteq E} f(x), \tag{4}$$

where $f(x)$ is a smooth convex function, $Q = \otimes_{i=1}^n Q_i \subseteq E$, with $Q_i \subseteq E_i$, $i = 1, ..., n$ being closed convex sets.

We now list our main assumptions and illustrate them by two simple examples. More detailed examples are given in Section 3. As the first example here, we consider random directional search, in which the gradient of the function f is approximated by a vector $\langle \nabla f(x), e \rangle e$, where $\langle \nabla f(x), e \rangle$ is the directional derivative in direction e and random vector e is uniformly distributed over the Euclidean sphere of radius 1. Our second example is random block-coordinate descent, in which the gradient of the function f is approximated by a vector $\widetilde{U}_i U_i^T \nabla f(x)$, where $U_i^T \nabla f(x)$ is i-th block derivative and the block number i is uniformly randomly sampled from $1, ..., n$. The common part in both these randomized gradient approximations is that, first, one randomly chooses a subspace which is either the line, parallel to e, or i-th block of coordinates. Then, one projects the gradient on this subspace by calculating either $\langle \nabla f(x), e \rangle$ or $U_i^T \nabla f(x)$. Finally, one lifts the obtained random projection back to the whole space E either by multiplying directional derivative by vector e, or applying dual partition operator $\widetilde{U}_i$. At the same time, in both cases, if one scales the obtained randomized approximation for the gradient by multiplying it by n, one obtains an *unbiased randomized approximation* of the gradient

$$\mathbb{E}_e n \langle \nabla f(x), e \rangle e = \nabla f(x), \quad \mathbb{E}_i n \widetilde{U}_i U_i^T \nabla f(x) = \nabla f(x), \quad x \in Q.$$

We also want our approach to allow construction of derivative-free methods. For a function f with L-Lipschitz-continuous gradient, the directional derivative can be well approximated by the difference of function values in two close points. Namely, it holds that

$$\langle \nabla f(x), e \rangle = \frac{f(x + \tau e) - f(x)}{\tau} + o(\tau),$$

where $\tau > 0$ is a small parameter. Thus, if only the value of the function is available, one can calculate only inexact directional derivative, which leads to biased randomized approximation for the gradient if the direction is chosen randomly. These three features, namely, random projection and lifting up, unbiased part of the randomized approximation for the gradient, bias in the randomized approximation for the gradient, lead us to the following assumption about the structure of our general *Randomized Inexact Oracle*.

Assumption 1 (Randomized Inexact Oracle)**.** We access the function f only through Randomized Inexact Oracle $\widetilde{\nabla} f(x)$, $x \in Q$, which is given by

$$\widetilde{\nabla} f(x) = \rho \mathcal{R}_r (\mathcal{R}_p^T \nabla f(x) + \xi(x)) \in E^*, \tag{5}$$

where $\rho > 0$ is a known constant; $\mathcal{R}_p$ is a random "'projection"' operator from some auxiliary space H to E, and, hence, $\mathcal{R}_p^T$, acting from E^* to H^*, is the adjoint to $\mathcal{R}_p$; $\mathcal{R}_r : H^* \to E^*$ is also some random "'reconstruction"' operator; $\xi(x) \in H^*$ is a, possibly random, vector characterizing the error of the oracle. The oracle is also assumed to satisfy the following properties

$$\mathbb{E} \rho \mathcal{R}_r \mathcal{R}_p^T \nabla f(x) = \nabla f(x), \quad \forall x \in Q, \tag{6}$$

$$\|\mathcal{R}_r \xi(x)\|_{E,*} \leq \delta, \quad \forall x \in Q, \tag{7}$$

where $\delta \geq 0$ is oracle *error level*.

Let us make some comments on this assumption. The nature of the operator $\mathcal{R}_p^T$ is generalization of random projection. For the case of random directional search, $H = \mathbb{R}$, $\mathcal{R}_p^T : E^* \to \mathbb{R}$ is given by $\mathcal{R}_p^T g = \langle g, e \rangle$, $g \in E^*$. For the case of random block-coordinate descent, $H = E_i$, $\mathcal{R}_p^T : E^* \to E_i^*$ is given by $\mathcal{R}_p^T g = U_i^T g$, $g \in E^*$. We assume that there is some additive error $\xi(x)$ in the generalized random projection $\mathcal{R}_p^T \nabla f(x)$. This error can be introduced, for example, when finite-difference approximation of the directional derivative is used. Finally, we lift the inexact random projection $\mathcal{R}_p^T \nabla f(x) + \xi(x)$ back to E by applying operator $\mathcal{R}_r$. For the case of random directional search, $\mathcal{R}_r : \mathbb{R} \to E^*$ is given by $\mathcal{R}_r t = te$, $t \in \mathbb{R}$. For the case of random block-coordinate descent, $\mathcal{R}_r : E_i^* \to E^*$ is given by $\mathcal{R}_r g^{(i)} = \widetilde{U}_i g^{(i)}$, $g^{(i)} \in E_i^*$. The number ρ is the normalizing coefficient, which allows the part $\rho \mathcal{R}_r \mathcal{R}_p^T \nabla f(x)$ to be unbiased randomized approximation for the gradient. This is expressed by equality (6). Finally, we assume that the error in our oracle is bounded, which is expressed by property (7). In our analysis, we consider two cases: the error ξ can be controlled and δ can be appropriately chosen on each iteration of the algorithm; the error ξ can not be controlled and we only know oracle error level δ.

Let us move to the next assumption. As said, our generic algorithm is based on Similar Triangles Method of Tyurin [2017] (see also Dvurechensky et al. [2017]), which is an accelerated gradient method with only one proximal mapping on each iteration. This proximal mapping is essentially the Mirror Descent step. For simplicity, let us consider here an unconstrained minimization problem in the Euclidean setting. This means that $Q_i = E_i = \mathbb{R}^{p_i}$, $\|x^{(i)}\|_i = \|x^{(i)}\|_2$, $i = 1, ..., n$. Then, given a point $u \in E$, a number α, and the gradient $\nabla f(y)$ at some point $y \in E$, the Mirror Descent step is

$$u_+ = \arg\min_{x \in E} \left\{ \frac{1}{2}\|x - u\|_2^2 + \alpha \langle \nabla f(y), x \rangle \right\} = u - \alpha \nabla f(y).$$

Now we want to substitute the gradient $\nabla f(y)$ with our Randomized Inexact Oracle $\widetilde{\nabla} f(y)$. Then, we see that the step $u_+ = u - \alpha \widetilde{\nabla} f(y)$ makes progress only in the subspace onto which the gradient is projected, while constructing the Randomized Inexact Oracle. In other words, $u - u_+$ lies in the same subspace as $\widetilde{\nabla} f(y)$. In our analysis, this is a desirable property and we formalize it as follows.

Assumption 2 (Regularity of Prox-Mapping)**.** The set Q, norm $\|\cdot\|_E$, prox-function $d(x)$, and Randomized Inexact Oracle $\widetilde{\nabla} f(x)$ are chosen in such a way that, for any $u, y \in Q$, $\alpha > 0$, the point

$$u_+ = \arg\min_{x \in Q} \left\{ V[u](x) + \alpha \langle \widetilde{\nabla} f(y), x \rangle \right\} \tag{8}$$

satisfies

$$\langle \mathcal{R}_r \mathcal{R}_p^T \nabla f(y), u - u_+ \rangle = \langle \nabla f(y), u - u_+ \rangle. \tag{9}$$

The interpretation is that, in terms of linear pairing with $u - u_+$, the unbiased part $\mathcal{R}_r \mathcal{R}_p^T \nabla f(y)$ of the Randomized Inexact Oracle makes the same progress as the true gradient $\nabla f(y)$.

Finally, we want to formalize the smoothness assumption for the function f. In our analysis, we use only the smoothness of f in the direction of $u_+ - u$, where $u \in Q$ and u_+ is defined in (8). Thus, we consider two points $x, y \in Q$, which satisfy equality $x = y + a(u_+ - u)$, where $a \in \mathbb{R}$. For the random directional search, it is natural to assume that f has L-Lipschitz-continuous gradient with respect to the Euclidean norm, i.e.

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|_2^2, \quad x, y \in Q. \tag{10}$$

Then, if we define $\|x\|_E^2 = L\|x\|_2^2$, we obtain that, for our choice $x = y + a(u_+ - u)$,

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}\|x - y\|_E^2.$$

Usual assumption for random block-coordinate descent is that the gradient of f is block-wise Lipschitz continuous. This means that, for all $i = 1, ..., n$, block derivative $f_i'(x) = U_i^T \nabla f(x)$ is L_i-Lipschitz continuous with respect to chosen norm $\|\cdot\|_i$, i.e.

$$\|f_i'(x + U_i h^{(i)}) - f_i'(x)\|_{i,*} \leq L_i \|h^{(i)}\|_i, \quad h^{(i)} \in E_i, \quad i = 1, ..., n, \quad x \in Q. \tag{11}$$

By the standard reasoning, using (11), one can prove that, for all $i = 1, ..., n$,

$$f(x + U_i h^{(i)}) \leq f(x) + \langle U_i^T \nabla f(x), h^{(i)} \rangle + \frac{L_i}{2}\|h^{(i)}\|_i^2, \quad h^{(i)} \in E_i, \quad x \in Q. \tag{12}$$

In block-coordinate setting, $\widetilde{\nabla} f(x)$ has non-zero elements only in one, say i-th, block and it follows from (8) that $u_+ - u$ also has non-zero components only in the i-th block. Hence, there exists $h^{(i)} \in E_i$, such that $u_+ - u = U_i h_i$ and $x = y + aU_i h^{(i)}$. Then, if we define $\|x\|_E^2 = \sum_{i=1}^n L_i \|x^{(i)}\|_i^2$, we obtain

$$\begin{aligned} f(x) = f(y + aU_i h^{(i)}) &\overset{(12)}{\leq} f(y) + \langle U_i^T \nabla f(y), a h^{(i)} \rangle + \frac{L_i}{2}\|a h^{(i)}\|_i^2 \\ &= f(y) + \langle \nabla f(y), a U_i h^{(i)} \rangle + \frac{1}{2}\|a U_i h^{(i)}\|_E^2 \\ &= f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}\|x - y\|_E^2. \end{aligned}$$

We generalize these two examples and assume smoothness of f in the following sense.

Assumption 3 (Smoothness)**.** The norm $\|\cdot\|_E$ is chosen in such a way that, for any $u, y \in Q$, $a \in \mathbb{R}$, if $x = y + a(u_+ - u) \in Q$, then

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}\|x - y\|_E^2. \tag{13}$$

Algorithm 1 Randomized Similar Triangles Method (RSTM)

Input: starting point $u_0 \in Q^0 = \otimes_{i=1}^n Q_i^0$, prox-setup: $d(x)$, $V[u](x)$, see Subsection 1.1.

1: Set $k = 0$, $A_0 = \alpha_0 = 1 - \frac{1}{\rho}$, $x_0 = y_0 = u_0$.

2: **repeat**

3: Find α_{k+1} as the largest root of the equation

$$A_{k+1} := A_k + \alpha_{k+1} = \rho^2 \alpha_{k+1}^2. \tag{14}$$

4: Calculate

$$y_{k+1} = \frac{\alpha_{k+1} u_k + A_k x_k}{A_{k+1}}. \tag{15}$$

5: Calculate

$$u_{k+1} = \arg\min_{x \in Q} \{V[u_k](x) + \alpha_{k+1} \langle \widetilde{\nabla} f(y_{k+1}), x \rangle\}. \tag{16}$$

6: Calculate

$$x_{k+1} = y_{k+1} + \rho \frac{\alpha_{k+1}}{A_{k+1}} (u_{k+1} - u_k). \tag{17}$$

7: Set $k = k + 1$.

8: **until** ...

Output: The point x_{k+1}.

2 Randomized Similar Triangles Method

In this section, we introduce our generic Randomized Similar Triangles Method, which is listed as Algorithm 1 below, and prove Theorem 1, which gives its convergence rate. The method is constructed by a modification of Similar Triangles Method (see Dvurechensky et al. [2017]) and, thus, inherits part of its name.

Lemma 1. *Algorithm 1 is correctly defined in the sense that, for all $k \geq 0$, $x_k, y_k \in Q$.*

Proof. The proof is a direct generalization of Lemma 2 in Fercoq and Richtárik [2015]. By definition (16), for all $k \geq 0$, $u_k \in Q$. If we prove that, for all $k \geq 0$, $x_k \in Q$, then, from (15), it follows that, for all $k \geq 0$, $y_k \in Q$. Let us prove that, for all $k \geq 0$, x_k is a convex combination of $u_0 \dots u_k$, namely $x_k = \sum_{l=0}^k \gamma_k^l u_l$, where $\gamma_0^0 = 1$, $\gamma_1^0 = 0$, $\gamma_1^1 = 1$, and for $k \geq 1$,

$$\gamma_{k+1}^l = \begin{cases} \left(1 - \frac{\alpha_{k+1}}{A_{k+1}}\right) \gamma_k^l, & l = 0, \dots, k-1 \\ \frac{\alpha_{k+1}}{A_{k+1}} \left(1 - \rho \frac{\alpha_k}{A_k}\right) + \rho \left(\frac{\alpha_k}{A_k} - \frac{\alpha_{k+1}}{A_{k+1}}\right), & l = k \\ \rho \frac{\alpha_{k+1}}{A_{k+1}}, & l = k+1. \end{cases} \tag{18}$$

Since, $x_0 = u_0$, we have that $\gamma_0^0 = 1$. Next, by (17), we have $x_1 = y_1 + \rho \frac{\alpha_1}{A_1}(u_1 - u_0) = u_0 + \rho \frac{\alpha_1}{A_1}(u_1 - u_0) = (1 - \rho \frac{\alpha_1}{A_1}) u_0 + \rho \frac{\alpha_1}{A_1} u_1$. Solving the equation (14) for $k = 0$, and using the

choice $\alpha_0 = 1 - \frac{1}{\rho}$, we obtain that $\alpha_1 = \frac{1}{\rho}$ and

$$\frac{\alpha_1}{A_1} \overset{(14)}{=} \frac{\alpha_1}{\rho^2 \alpha_1^2} = \frac{1}{\rho}. \tag{19}$$

Hence, $x_1 = u_1$ and $\gamma_1^0 = 0$, $\gamma_1^1 = 1$. Let us now assume that $x_k = \sum_{l=0}^{k} \gamma_k^l u_l$ and prove that x_{k+1} is also a convex combination with coefficients, given by (18). From (15), (17), we have

$$\begin{aligned} x_{k+1} = y_{k+1} + \rho\frac{\alpha_{k+1}}{A_{k+1}}(u_{k+1} - u_k) &= \frac{\alpha_{k+1}u_k + A_k x_k}{A_{k+1}} + \rho\frac{\alpha_{k+1}}{A_{k+1}}(u_{k+1} - u_k) \\ &= \frac{A_k}{A_{k+1}} x_k + \left(\frac{\alpha_{k+1}}{A_{k+1}} - \rho\frac{\alpha_{k+1}}{A_{k+1}}\right) u_k + \rho\frac{\alpha_{k+1}}{A_{k+1}} u_{k+1} \\ = \left(1 - \frac{\alpha_{k+1}}{A_{k+1}}\right) \sum_{l=0}^{k} \gamma_k^l u_l &+ \left(\frac{\alpha_{k+1}}{A_{k+1}} - \rho\frac{\alpha_{k+1}}{A_{k+1}}\right) u_k + \rho\frac{\alpha_{k+1}}{A_{k+1}} u_{k+1}. \end{aligned}$$

Note that all the coefficients sum to 1. Next, we have

$$\begin{aligned} x_{k+1} &= \left(1 - \frac{\alpha_{k+1}}{A_{k+1}}\right) \sum_{l=0}^{k-1} \gamma_k^l u_l + \left(\gamma_k^k \left(1 - \frac{\alpha_{k+1}}{A_{k+1}}\right) + \left(\frac{\alpha_{k+1}}{A_{k+1}} - \rho\frac{\alpha_{k+1}}{A_{k+1}}\right)\right) u_k + \rho\frac{\alpha_{k+1}}{A_{k+1}} u_{k+1} \\ &= \left(1 - \frac{\alpha_{k+1}}{A_{k+1}}\right) \sum_{l=0}^{k-1} \gamma_k^l u_l + \left(\rho\frac{\alpha_k}{A_k} \left(1 - \frac{\alpha_{k+1}}{A_{k+1}}\right) + \left(\frac{\alpha_{k+1}}{A_{k+1}} - \rho\frac{\alpha_{k+1}}{A_{k+1}}\right)\right) u_k + \rho\frac{\alpha_{k+1}}{A_{k+1}} u_{k+1} \\ &= \left(1 - \frac{\alpha_{k+1}}{A_{k+1}}\right) \sum_{l=0}^{k-1} \gamma_k^l u_l + \left(\frac{\alpha_{k+1}}{A_{k+1}} \left(1 - \rho\frac{\alpha_k}{A_k}\right) + \rho\left(\frac{\alpha_k}{A_k} - \frac{\alpha_{k+1}}{A_{k+1}}\right)\right) u_k + \rho\frac{\alpha_{k+1}}{A_{k+1}} u_{k+1}. \end{aligned}$$

So, we see that (18) holds for $k+1$. It remains to show that $\gamma_{k+1}^l \geq 0$, $l = 0, \ldots, k+1$. For γ_{k+1}^l, $l = 0, \ldots, k-1$ и γ_{k+1}^{k+1} it is obvious. From (14), we have

$$\alpha_{k+1} = \frac{1 + \sqrt{1 + 4\rho^2 A_k}}{2\rho^2}.$$

Thus, since $\{A_k\}$, $k \geq 0$ is non-decreasing sequence, $\{\alpha_{k+1}\}$, $k \geq 0$ is also non-decreasing. From (14), we obtain $\frac{\alpha_{k+1}}{A_{k+1}} = \frac{\alpha_{k+1}}{\rho^2 \alpha_{k+1}^2}$, which means that this sequence is non-increasing. Thus, $\frac{\alpha_k}{A_k} \geq \frac{\alpha_{k+1}}{A_{k+1}}$ and $\frac{\alpha_k}{A_k} \leq \frac{\alpha_1}{A_1} \leq \frac{1}{\rho}$ for $k \geq 1$. These inequalities prove that $\gamma_{k+1}^k \geq 0$. □

Lemma 2. *Let the sequences $\{x_k, y_k, u_k, \alpha_k, A_k\}$, $k \geq 0$ be generated by Algorithm 1. Then, for all $u \in Q$, it holds that*

$$\begin{aligned} \alpha_{k+1}\langle \widetilde{\nabla} f(y_{k+1}), u_k - u\rangle \leq & A_{k+1}(f(y_{k+1}) - f(x_{k+1})) + V[u_k](u) - V[u_{k+1}](u) \\ & + \alpha_{k+1}\rho\langle \mathcal{R}_r \xi(y_{k+1}), u_k - u_{k+1}\rangle. \end{aligned} \tag{20}$$

Proof. Using Assumptions 1 and 2 with $\alpha = \alpha_{k+1}$, $y = y_{k+1}$, $u = u_k$, $u_+ = u_{k+1}$, we obtain

$$\begin{aligned}\alpha_{k+1}\langle\widetilde{\nabla} f(y_{k+1}), u_k - u_{k+1}\rangle &\overset{(5)}{=} \alpha_{k+1}\rho\langle\mathcal{R}_r(\mathcal{R}_p^T\nabla f(y_{k+1}) + \xi(y_{k+1})), u_k - u_{k+1}\rangle \\ &\overset{(9)}{=} \alpha_{k+1}\rho\langle\nabla f(y_{k+1}), u_k - u_{k+1}\rangle + \alpha_{k+1}\rho\langle\mathcal{R}_r\xi(y_{k+1}), u_k - u_{k+1}\rangle \\ &\overset{(17)}{=} A_{k+1}\langle\nabla f(y_{k+1}), y_{k+1} - x_{k+1}\rangle + \alpha_{k+1}\rho\langle\mathcal{R}_r\xi(y_{k+1}), u_k - u_{k+1}\rangle.\end{aligned} \tag{21}$$

Note that, from the optimality condition in (16), for any $u \in Q$, we have

$$\langle\nabla V[u_k](u_{k+1}) + \alpha_{k+1}\widetilde{\nabla} f(y_{k+1}), u - u_{k+1}\rangle \geq 0. \tag{22}$$

By the definition of $V[u](x)$, we obtain, for any $u \in Q$,

$$\begin{aligned}V[u_k](u) - V[u_{k+1}](u) - V[u_k](u_{k+1}) =& d(u) - d(u_k) - \langle\nabla d(u_k), u - u_k\rangle \\ &- (d(u) - d(u_{k+1}) - \langle\nabla d(u_{k+1}), u - u_{k+1}\rangle) \\ &- (d(u_{k+1}) - d(u_k) - \langle\nabla d(u_k), u_{k+1} - u_k\rangle) \\ =& \langle\nabla d(u_k) - \nabla d(u_{k+1}), u_{k+1} - u\rangle \\ =& \langle -\nabla V[u_k](u_{k+1}), u_{k+1} - u\rangle.\end{aligned} \tag{23}$$

Further, for any $u \in Q$, by Assumption 3,

$$\begin{aligned}\alpha_{k+1}\langle\widetilde{\nabla} f(y_{k+1}), u_k - u\rangle &= \alpha_{k+1}\langle\widetilde{\nabla} f(y_{k+1}), u_k - u_{k+1}\rangle + \alpha_{k+1}\langle\widetilde{\nabla} f(y_{k+1}), u_{k+1} - u\rangle \\ &\overset{(22)}{\leq} \alpha_{k+1}\langle\widetilde{\nabla} f(y_{k+1}), u_k - u_{k+1}\rangle + \langle -\nabla V[u_k](u_{k+1}), u_{k+1} - u\rangle \\ &\overset{(23)}{=} \alpha_{k+1}\langle\widetilde{\nabla} f(y_{k+1}), u_k - u_{k+1}\rangle + V[u_k](u) - V[u_{k+1}](u) - V[u_k](u_{k+1}) \\ &\overset{(3)}{\leq} \alpha_{k+1}\langle\widetilde{\nabla} f(y_{k+1}), u_k - u_{k+1}\rangle + V[u_k](u) - V[u_{k+1}](u) - \frac{1}{2}\|u_k - u_{k+1}\|_E^2 \\ &\overset{(21),(17)}{=} A_{k+1}\langle\nabla f(y_{k+1}), y_{k+1} - x_{k+1}\rangle + \alpha_{k+1}\rho\langle\mathcal{R}_r\xi(y_{k+1}), u_k - u_{k+1}\rangle + \\ &\quad + V[u_k](u) - V[u_{k+1}](u) - \frac{A_{k+1}^2}{2\rho^2\alpha_{k+1}^2}\|y_{k+1} - x_{k+1}\|_E^2 \\ &\overset{(14)}{=} A_{k+1}\left(\langle\nabla f(y_{k+1}), y_{k+1} - x_{k+1}\rangle - \frac{1}{2}\|y_{k+1} - x_{k+1}\|_E^2\right) + \\ &\quad + \alpha_{k+1}\rho\langle\mathcal{R}_r\xi(y_{k+1}), u_k - u_{k+1}\rangle + V[u_k](u) - V[u_{k+1}](u) \\ &\overset{(17),(13)}{\leq} A_{k+1}\left(f(y_{k+1}) - f(x_{k+1})\right) + V[u_k](u) - V[u_{k+1}](u) + \\ &\quad + \alpha_{k+1}\rho\langle\mathcal{R}_r\xi(y_{k+1}), u_k - u_{k+1}\rangle.\end{aligned}$$

In the last inequality, we used Assumption 3 with $a = \rho\frac{\alpha_{k+1}}{A_{k+1}}$, $x = x_{k+1}$, $y = y_{k+1}$, $u = u_k$, $u_+ = u_{k+1}$. □

Lemma 3. *Let the sequences $\{x_k, y_k, u_k, \alpha_k, A_k\}$, $k \geq 0$ be generated by Algorithm 1. Then, for all $u \in Q$, it holds that*

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(y_{k+1}), u_k - u\rangle \leq & A_{k+1}(f(y_{k+1}) - \mathbb{E}_{k+1} f(x_{k+1})) + V[u_k](u) - \mathbb{E}_{k+1} V[u_{k+1}](u) \\ & + \mathbb{E}_{k+1}\alpha_{k+1}\rho\langle \mathcal{R}_r \xi(y_{k+1}), u - u_{k+1}\rangle, \end{aligned} \tag{24}$$

where $\mathbb{E}_{k+1}$ denotes the expectation conditioned on all the randomness up to step k.

Proof. First, for any $u \in Q$, by Assumption 1,

$$\begin{aligned}\mathbb{E}_{k+1}\alpha_{k+1}\langle \widetilde{\nabla} f(y_{k+1}), u_k - u\rangle &\overset{(5)}{=} \mathbb{E}_{k+1}\alpha_{k+1}\rho\langle \mathcal{R}_r(\mathcal{R}_p^T \nabla f(y_{k+1}) + \xi(y_{k+1})), u_k - u\rangle \\ &\overset{(6)}{=} \alpha_{k+1}\langle \nabla f(y_{k+1}), u_k - u\rangle + \mathbb{E}_{k+1}\alpha_{k+1}\rho\langle \mathcal{R}_r \xi(y_{k+1}), u_k - u\rangle. \end{aligned} \tag{25}$$

Taking conditional expectation $\mathbb{E}_{k+1}$ in (20) of Lemma 2 and using (25), we obtain the statement of the Lemma. □

Lemma 4. *Let the sequences $\{x_k, y_k, u_k, \alpha_k, A_k\}$, $k \geq 0$ be generated by Algorithm 1. Then, for all $u \in Q$, it holds that*

$$\begin{aligned}A_{k+1}\mathbb{E}_{k+1} f(x_{k+1}) - A_k f(x_k) \leq & \alpha_{k+1}\left(f(y_{k+1}) + \langle \nabla f(y_{k+1}), u - y_{k+1}\rangle\right) + V[u_k](u) \\ & - \mathbb{E}_{k+1} V[u_{k+1}](u) + \mathbb{E}_{k+1}\alpha_{k+1}\rho\langle \mathcal{R}_r \xi(y_{k+1}), u - u_{k+1}\rangle. \end{aligned} \tag{26}$$

Proof. For any $u \in Q$,

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(y_{k+1}), y_{k+1} - u\rangle &= \alpha_{k+1}\langle \nabla f(y_{k+1}), y_{k+1} - u_k\rangle + \alpha_{k+1}\langle \nabla f(y_{k+1}), u_k - u\rangle \\ &\overset{(14),(15)}{=} A_k\langle \nabla f(y_{k+1}), x_k - y_{k+1}\rangle + \alpha_{k+1}\langle \nabla f(y_{k+1}), u_k - u\rangle \\ &\overset{\text{conv-ty}}{\leq} A_k\left(f(x_k) - f(y_{k+1})\right) + \alpha_{k+1}\langle \nabla f(y_{k+1}), u_k - u\rangle \\ &\overset{(24)}{\leq} A_k\left(f(x_k) - f(y_{k+1})\right) + A_{k+1}(f(y_{k+1}) - \mathbb{E}_{k+1} f(x_{k+1})) + \\ &\quad + V[u_k](u) - \mathbb{E}_{k+1} V[u_{k+1}](u) + \mathbb{E}_{k+1}\alpha_{k+1}\rho\langle \mathcal{R}_r \xi(y_{k+1}), u - u_{k+1}\rangle \\ &\overset{(14)}{=} \alpha_{k+1} f(y_{k+1}) + A_k f(x_k) - A_{k+1}\mathbb{E}_{k+1} f(x_{k+1}) + V[u_k](u) - \mathbb{E}_{k+1} V[u_{k+1}](u) \\ &\quad + \mathbb{E}_{k+1}\alpha_{k+1}\rho\langle \mathcal{R}_r \xi(y_{k+1}), u - u_{k+1}\rangle. \end{aligned} \tag{27}$$

Rearranging terms, we obtain the statement of the Lemma. □

Theorem 1. *Let the assumptions 1, 2, 3 hold. Let the sequences $\{x_k, y_k, u_k, \alpha_k, A_k\}$, $k \geq 0$ be generated by Algorithm 1. Let f_* be the optimal objective value and x_* be an optimal point in Problem (4). Denote*

$$P_0^2 = A_0(f(x_0) - f_*) + V[u_0](x_*). \tag{28}$$

1. *If the oracle error $\xi(x)$ in (5) can be controlled and, on each iteration, the error level δ in (7) satisfies*

$$\delta \leq \frac{P_0}{4\rho A_k}, \tag{29}$$

then, for all $k \geq 1$,

$$\mathbb{E}f(x_k) - f_* \leq \frac{3P_0^2}{2A_k},$$

where $\mathbb{E}$ denotes the expectation with respect to all the randomness up to step k.

2. *If the oracle error $\xi(x)$ in (5) can not be controlled, then, for all $k \geq 1$,*

$$\mathbb{E}f(x_k) - f_* \leq \frac{2P_0^2}{A_k} + 4A_k\rho^2\delta^2.$$

Proof. Let us change the counter in Lemma 4 from k to i, fix $u = x_*$, take the full expectation in each inequality for $i = 0, ..., k-1$ and sum all the inequalities for $i = 0, ..., k-1$. Then,

$$\begin{aligned} A_k\mathbb{E}f(x_k) - A_0f(x_0) \leq & \sum_{i=0}^{k-1} \alpha_{i+1}\mathbb{E}\left(f(y_{i+1}) + \langle \nabla f(y_{i+1}), x_* - y_{i+1}\rangle\right) + V[u_0](x_*) - \mathbb{E}V[u_k](x_*) \\ & + \sum_{i=0}^{k-1} \mathbb{E}\alpha_{i+1}\rho\langle \mathcal{R}_r\xi(y_{i+1}), x_* - u_{i+1}\rangle \\ \overset{\text{conv-ty},(14),(7)}{\leq} & (A_k - A_0)f(x_*) + V[u_0](x_*) - \mathbb{E}V[u_k](x_*) + \sum_{i=0}^{k-1} \alpha_{i+1}\rho\delta\mathbb{E}\|x_* - u_{i+1}\|_E. \end{aligned}$$

Rearranging terms and using (28), we obtain, for all $k \geq 1$,

$$0 \leq A_k\left(\mathbb{E}f(x_k) - f_*\right) \leq P_0^2 - \mathbb{E}V[u_k](x_*) + \rho\delta\sum_{i=0}^{k-1}\alpha_{i+1}\mathbb{E}R_{i+1}, \tag{30}$$

where we denoted $R_i = \|u_i - x_*\|_E$, $i \geq 0$.

1. We first prove the first statement of the Theorem. We have

$$\frac{1}{2}R_0^2 = \frac{1}{2}\|x_* - u_0\|_E^2 \overset{(3)}{\leq} V[u_0](x_*) \overset{(28)}{\leq} P_0^2. \tag{31}$$

Hence, $\mathbb{E}R_0 = R_0 \leq P_0\sqrt{2} \leq 2P_0$. Let $\mathbb{E}R_i \leq 2P_0$, for all $i = 0, ..., k-1$. Let us prove that $\mathbb{E}R_k \leq 2P_0$. By convexity of square function, we obtain

$$\begin{aligned} \frac{1}{2}\left(\mathbb{E}R_k\right)^2 \leq \frac{1}{2}\mathbb{E}R_k^2 \overset{(3)}{\leq} \mathbb{E}V[u_k](x_*) & \overset{(30)}{\leq} P_0^2 + \rho\delta\sum_{i=0}^{k-2}\alpha_{i+1}2P_0 + \alpha_k\rho\delta\mathbb{E}R_k \\ & \overset{(14)}{=} P_0^2 + 2\rho\delta P_0(A_{k-1} - A_0) + \alpha_k\rho\delta\mathbb{E}R_k \\ & \leq P_0^2 + 2\rho\delta P_0A_k + \alpha_k\rho\delta\mathbb{E}R_k. \end{aligned} \tag{32}$$

Since $\alpha_k \leq A_k$, $k \geq 0$, by the choice of δ (29), we have $2\rho\delta P_0 A_k \leq \frac{P_0^2}{2}$ and $\alpha_k \rho\delta \leq A_k \rho\delta \leq \frac{P_0}{4}$. So, we obtain an inequality for $\mathbb{E}R_k$

$$\frac{1}{2}\left(\mathbb{E}R_k\right)^2 \leq \frac{3P_0^2}{2} + \frac{P_0}{4}\mathbb{E}R_k.$$

Solving this quadratic inequality in $\mathbb{E}R_k$, we obtain

$$\mathbb{E}R_k \leq \frac{P_0}{4} + \sqrt{\frac{P_0^2}{16} + 3P_0^2} = 2P_0.$$

Thus, by induction, we have that, for all $k \geq 0$, $\mathbb{E}R_k \leq 2P_0$. Using the bounds $\mathbb{E}R_i \leq 2P_0$, for all $i = 0, ..., k$, we obtain

$$A_k\left(\mathbb{E}f(x_k) - f_*\right) \overset{(30)}{\leq} P_0^2 + \rho\delta\sum_{i=0}^{k-1}\alpha_{i+1}\mathbb{E}R_i \overset{(14),(29)}{\leq} P_0^2 + \rho\frac{P_0}{4\rho A_k}\cdot\left(A_k - A_0\right)\cdot 2P_0 \leq \frac{3P_0^2}{2}.$$

This finishes the proof of the first statement of the Theorem.

2. Now we prove the second statement of the Theorem. First, from (30) for $k = 1$, we have

$$\frac{1}{2}\left(\mathbb{E}R_1\right)^2 \leq \frac{1}{2}\mathbb{E}R_1^2 \overset{(3)}{\leq} \mathbb{E}V[u_1](x_*) \overset{(30)}{\leq} P_0^2 + \rho\delta\alpha_1\mathbb{E}R_1.$$

Solving this inequality in $\mathbb{E}R_1$, we obtain

$$\mathbb{E}R_1 \leq \rho\delta\alpha_1 + \sqrt{(\rho\delta\alpha_1)^2 + 2P_0^2} \leq 2\rho\delta\alpha_1 + P_0\sqrt{2}, \tag{33}$$

where we used that, for any $a, b \geq 0$, $\sqrt{a^2+b^2} \leq a + b$. Then,

$$P_0^2 + \rho\delta\alpha_1\mathbb{E}R_1 \leq P_0^2 + 2(\rho\delta\alpha_1)^2 + \rho\delta\alpha_1 P_0\sqrt{2} \leq \left(P_0 + \rho\delta\sqrt{2}(A_1 - A_0)\right)^2.$$

Thus, we have proved that the inequality

$$P_0^2 + \rho\delta\sum_{i=0}^{k-2}\alpha_{i+1}\mathbb{E}R_{i+1} \leq \left(P_0 + \rho\delta\sqrt{2}(A_{k-1} - A_0)\right)^2 \tag{34}$$

holds for $k = 2$. Let us assume that it holds for some k and prove that it holds for $k+1$. We have

$$\begin{aligned}\frac{1}{2}\left(\mathbb{E}R_k\right)^2 \leq \frac{1}{2}\mathbb{E}R_k^2 \overset{(3)}{\leq} \mathbb{E}V[u_k](x_*) &\overset{(30)}{\leq} P_0^2 + \rho\delta\sum_{i=0}^{k-2}\alpha_{i+1}\mathbb{E}R_{i+1} + \alpha_k\rho\delta\mathbb{E}R_k \\ &\overset{(34)}{\leq} \left(P_0 + \rho\delta\sqrt{2}(A_{k-1} - A_0)\right)^2 + \alpha_k\rho\delta\mathbb{E}R_k.\end{aligned}$$

Solving this quadratic inequality in $\mathbb{E}R_k$, we obtain

$$\begin{aligned}\mathbb{E}R_k &\leq \alpha_k\rho\delta + \sqrt{(\alpha_k\rho\delta)^2 + 2\left(P_0 + \rho\delta\sqrt{2}(A_{k-1} - A_0)\right)^2} \\ &\leq 2\alpha_k\rho\delta + \left(P_0 + \rho\delta\sqrt{2}(A_{k-1} - A_0)\right)\sqrt{2}, \end{aligned} \tag{35}$$

where we used that, for any $a, b \geq 0$, $\sqrt{a^2 + b^2} \leq a + b$. Further,

$$\begin{aligned} P_0^2 + \rho\delta\sum_{i=0}^{k-1}\alpha_{i+1}\mathbb{E}R_{i+1} &\overset{(34)}{\leq} \left(P_0 + \rho\delta\sqrt{2}(A_{k-1} - A_0)\right)^2 + \rho\delta\alpha_k\mathbb{E}R_k \\ &\overset{(35)}{\leq} \left(P_0 + \rho\delta\sqrt{2}(A_{k-1} - A_0)\right)^2 + 2(\rho\delta\alpha_k)^2 + \rho\delta\alpha_k\left(P_0 + \rho\delta\sqrt{2}(A_{k-1} - A_0)\right)\sqrt{2} \\ &\leq \left(P_0 + \rho\delta\sqrt{2}(A_{k-1} - A_0) + \rho\delta\alpha_k\sqrt{2}\right)^2 = \left(P_0 + \rho\delta\sqrt{2}(A_k - A_0)\right)^2, \end{aligned}$$

which is (34) for $k+1$. Using this inequality, we obtain

$$A_k\left(\mathbb{E}f(x_k) - f_*\right) \overset{(30)}{\leq} P_0^2 + \rho\delta\sum_{i=0}^{k-1}\alpha_{i+1}\mathbb{E}R_{i+1} \leq \left(P_0 + \rho\delta\sqrt{2}(A_k - A_0)\right)^2 \leq 2P_0^2 + 4\rho^2\delta^2A_k^2,$$

which finishes the proof of the Theorem.

□

Let us now estimate the growth rate of the sequence A_k, $k \geq 0$, which will give the rate of convergence for Algorithm 1.

Lemma 5. *Let the sequence $\{A_k\}$, $k \geq 0$ be generated by Algorithm 1. Then, for all $k \geq 1$ it holds that*

$$\frac{(k - 1 + 2\rho)^2}{4\rho^2} \leq A_k \leq \frac{(k - 1 + 2\rho)^2}{\rho^2}. \tag{36}$$

Proof. As we showed in Lemma 1, $\alpha_1 = \frac{1}{\rho}$ and, hence, $A_1 = \alpha_0 + \alpha_1 = 1$. Thus, (36) holds for $k = 1$. Let us assume that (36) holds for some $k \geq 1$ and prove that it holds also for $k+1$. From (14), we have a quadratic equation for α_{k+1}

$$\rho^2\alpha_{k+1}^2 - \alpha_{k+1} - A_k = 0.$$

Since we need to take the largest root, we obtain,

$$\begin{aligned}\alpha_{k+1} &= \frac{1 + \sqrt{1 + 4\rho^2A_k}}{2\rho^2} = \frac{1}{2\rho^2} + \sqrt{\frac{1}{4\rho^4} + \frac{A_k}{\rho^2}} \geq \frac{1}{2\rho^2} + \sqrt{\frac{A_k}{\rho^2}} \\ &\geq \frac{1}{2\rho^2} + \frac{k - 1 + 2\rho}{2\rho^2} = \frac{k + 2\rho}{2\rho^2}, \end{aligned}$$

where we used the induction assumption that (36) holds for k. On the other hand,

$$\alpha_{k+1} = \frac{1}{2\rho^2} + \sqrt{\frac{1}{4\rho^4} + \frac{A_k}{\rho^2}} \leq \frac{1}{\rho^2} + \sqrt{\frac{A_k}{\rho^2}}$$
$$\leq \frac{1}{\rho^2} + \frac{k-1+2\rho}{\rho^2} = \frac{k+2\rho}{\rho^2},$$

where we used inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, $a, b \geq 0$. Using the obtained inequalities for α_{k+1}, from (14) and (36) for k, we get

$$A_{k+1} = A_k + \alpha_{k+1} \geq \frac{(k-1+2\rho)^2}{4\rho^2} + \frac{k+2\rho}{2\rho^2} \geq \frac{(k+2\rho)^2}{4\rho^2}$$

and

$$A_{k+1} = A_k + \alpha_{k+1} \leq \frac{(k-1+2\rho)^2}{\rho^2} + \frac{k+2\rho}{\rho^2} \leq \frac{(k+2\rho)^2}{\rho^2}.$$

In the last inequality we used that $k \geq 1$, $\rho \geq 0$. □

Remark 1. According to Theorem 1, if the desired accuracy of the solution is ε, i.e. the goal is to find such $\hat{x} \in Q$ that $\mathbb{E}f(\hat{x}) - f_* \leq \varepsilon$, then the Algorithm 1 should be stopped when $\frac{3P_0^2}{2A_k} \leq \varepsilon$. Then $\frac{1}{A_k} \leq \frac{2\varepsilon}{3P_0^2}$ and the oracle error level δ should satisfy $\delta \leq \frac{P_0}{4\rho A_k} \leq \frac{\varepsilon}{6\rho P_0}$.

From Lemma 5, we obtain that $\frac{3P_0^2}{2A_k} \leq \varepsilon$ holds when k is the smallest integer satisfying

$$\frac{(k-1+2\rho)^2}{4\rho^2} \geq \frac{3P_0^2}{2\varepsilon}.$$

This means that, to obtain an ε-solution, it is enough to choose

$$k = \max\left\{\left\lceil \rho\sqrt{\frac{6P_0^2}{\varepsilon}} + 1 - 2\rho \right\rceil, 0\right\}.$$

Note that this dependence on ε means that the proposed method is accelerated.

3 Examples of Applications

In this section, we apply our general framework, which consists of assumptions 1, 2, 3, RSTM as listed in Algorithm 1 and convergence rate Theorem 1, to obtain several particular algorithms and their convergence rate. We consider Problem (4) and, for each particular case, introduce a particular setup, which includes properties of the objective function f, available information about this function, properties of the feasible set Q. Based on each setup, we show how the Randomized Inexact Oracle is constructed and check that the assumptions 1, 2, 3 hold. Then, we obtain convergence rate guarantee for each particular algorithm as a corollary of Theorem 1. Our examples include accelerated random directional search

with inexact directional derivative, accelerated random block-coordinate descent with inexact block derivatives, accelerated random derivative-free directional search with inexact function values, accelerated random derivative-free block-coordinate descent with inexact function values. Accelerated random directional search and accelerated random derivative-free directional search were developed in Nesterov and Spokoiny [2017], but for the case of exact directional derivatives and exact function values. Also, in the existing methods, a Gaussian random vector is used for randomization. Accelerated random block-coordinate descent was introduced in Nesterov [2012] and further developed in by several authors (see Introduction for the extended review). Existing methods of this type use exact information on the block derivatives and also only Euclidean proximal setup. In the contrast, our algorithm works with inexact derivatives and is able to work with entropy proximal setup. To the best of our knowledge, our accelerated random derivative-free block-coordinate descent with inexact function values is new. This method also can work with entropy proximal setup.

3.1 Accelerated Random Directional Search

In this subsection, we introduce accelerated random directional search with inexact directional derivative for unconstrained problems with Euclidean proximal setup. We assume that, for all $i = 1, ..., n$, $Q_i = E_i = \mathbb{R}$, $\|x^{(i)}\|_i^2 = (x^{(i)})^2$, $x^{(i)} \in E_i$, $d_i(x^{(i)}) = \frac{1}{2}(x^{(i)})^2$, $x^{(i)} \in E_i$ and, hence, $V_i[z^{(i)}](x^{(i)}) = \frac{1}{2}(x^{(i)} - z^{(i)})^2$, $x^{(i)}, z^{(i)} \in E_i$. Thus, $Q = E = \mathbb{R}^n$. Further, we assume that f in (4) has L-Lipschitz-continuous gradient with respect to Euclidean norm, i.e.

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|_2^2, \quad x, y \in E. \tag{37}$$

We set $\beta_i = L$, $i = 1, ..., n$. Then, by definitions in Subsection 1.1, we have $\|x\|_E^2 = L\|x\|_2^2$, $x \in E$, $d(x) = \frac{L}{2}\|x\|_2^2 = \frac{1}{2}\|x\|_E^2$, $x \in E$, $V[z](x) = \frac{L}{2}\|x - z\|_2^2 = \frac{1}{2}\|x - z\|_E^2$, $x, z \in E$. Also, we have $\|g\|_{E,*}^2 = L^{-1}\|g\|_2^2$, $g \in E^*$.

We assume that, at any point $x \in E$, one can calculate an inexact derivative of f in a direction $e \in E$

$$\tilde{f}'(x, e) = \langle \nabla f(x), e \rangle + \xi(x),$$

where e is a random vector uniformly distributed on the Euclidean sphere of radius 1, i.e. $\mathcal{S}_2(1) := \{s \in \mathbb{R}^n : \|s\|_2 = 1\}$, and the directional derivative error $\xi(x) \in \mathbb{R}$ is uniformly bounded in absolute value by error level Δ, i.e. $|\xi(x)| \leq \Delta$, $x \in E$. Since we are in the Euclidean setting, we consider e also as an element of E^*. We use $n(\langle \nabla f(x), e \rangle + \xi(x))e$ as Randomized Inexact Oracle.

Let us check the assumptions stated in Subsection 1.2.

Randomized Inexact Oracle. In this setting, we have $\rho = n$, $H = \mathbb{R}$, $\mathcal{R}_p^T : E^* \to \mathbb{R}$ is given by $\mathcal{R}_p^T g = \langle g, e \rangle$, $g \in E^*$, $\mathcal{R}_r : \mathbb{R} \to E^*$ is given by $\mathcal{R}_r t = te$, $t \in \mathbb{R}$. Thus,

$$\widetilde{\nabla} f(x) = n(\langle \nabla f(x), e \rangle + \xi(x))e.$$

One can prove that $\mathbb{E}_e n\langle \nabla f(x), e \rangle e = n\mathbb{E}_e e e^T \nabla f(x) = \nabla f(x)$, $x \in E$, and, thus, (6) holds. Also, for all $x \in E$, we have $\|\mathcal{R}_r \xi(x)\|_{E,*} = \frac{1}{\sqrt{L}}\|\xi(x)e\|_2 \leq \frac{\Delta}{\sqrt{L}}$, which proves (7) if we take $\delta = \frac{\Delta}{\sqrt{L}}$.

Regularity of Prox-Mapping. Substituting particular choice of Q, $V[u](x)$, $\widetilde{\nabla} f(x)$ in (8), we obtain

$$u_+ = \arg\min_{x \in \mathbb{R}^n} \left\{ \frac{L}{2}\|x-u\|_2^2 + \alpha\langle n(\langle \nabla f(y), e\rangle + \xi(y))e, x\rangle \right\} = u - \frac{\alpha n}{L}(\langle \nabla f(y), e\rangle + \xi(y))e.$$

Hence, since $\langle e, e\rangle = 1$, we have

$$\begin{aligned}
\langle \mathcal{R}_r \mathcal{R}_p^T \nabla f(y), u - u_+\rangle &= \left\langle \langle \nabla f(y), e\rangle e, \frac{\alpha n}{L}(\langle \nabla f(y), e\rangle + \xi(y))e \right\rangle \\
&= \langle \nabla f(y), e\rangle \langle e, e\rangle \frac{\alpha n}{L}(\langle \nabla f(y), e\rangle + \xi(y)) \\
&= \left\langle \nabla f(y), \frac{\alpha n}{L}(\langle \nabla f(y), e\rangle + \xi(y))e \right\rangle \\
&= \langle \nabla f(y), u - u_+\rangle,
\end{aligned}$$

which proves (9).

Smoothness. By definition of $\|\cdot\|_E$ and (37), we have

$$f(x) \le f(y) + \langle \nabla f(y), x-y\rangle + \frac{L}{2}\|x-y\|_2^2 = f(y) + \langle \nabla f(y), x-y\rangle + \frac{1}{2}\|x-y\|_E^2, \quad x, y \in E$$

and (13) holds.

We have checked that all the assumptions listed in Subsection 1.2 hold. Thus, we can obtain the following convergence rate result for random directional search as a corollary of Theorem 1 and Lemma 5.

Corollary 1. *Let Algorithm 1 with $\widetilde{\nabla} f(x) = n(\langle \nabla f(x), e\rangle + \xi(x))e$, where e is random and uniformly distributed over the Euclidean sphere of radius 1, be applied to Problem (4) in the setting of this subsection. Let f_* be the optimal objective value and x_* be an optimal point in Problem (4). Assume that directional derivative error $\xi(x)$ satisfies $|\xi(x)| \le \Delta$, $x \in E$. Denote*

$$P_0^2 = \left(1 - \frac{1}{n}\right)(f(x_0) - f_*) + \frac{L}{2}\|u_0 - x_*\|_2^2.$$

1. *If the directional derivative error $\xi(x)$ can be controlled and, on each iteration, the error level Δ satisfies*

$$\Delta \le \frac{P_0\sqrt{L}}{4nA_k},$$

then, for all $k \ge 1$,

$$\mathbb{E} f(x_k) - f_* \le \frac{6n^2 P_0^2}{(k-1+2n)^2},$$

where $\mathbb{E}$ denotes the expectation with respect to all the randomness up to step k.

2. *If the directional derivative error $\xi(x)$ can not be controlled, then, for all $k \ge 1$,*

$$\mathbb{E} f(x_k) - f_* \le \frac{8n^2 P_0^2}{(k-1+2n)^2} + \frac{4}{L}(k-1+2n)^2\Delta^2.$$

Remark 2. According to Remark 1 and due to the relation $\delta = \frac{\Delta}{\sqrt{L}}$, we obtain that the error level Δ in the directional derivative should satisfy

$$\Delta \leq \frac{\varepsilon\sqrt{L}}{6nP_0}.$$

At the same time, to obtain an ε-solution for Problem (4), it is enough to choose

$$k = \max\left\{\left\lceil n\sqrt{\frac{6P_0^2}{\varepsilon}} + 1 - 2n\right\rceil, 0\right\}.$$

3.2 Accelerated Random Coordinate Descent

In this subsection, we introduce accelerated random coordinate descent with inexact coordinate derivatives for problems with separable constraints and Euclidean proximal setup. We assume that, for all $i = 1, ..., n$, $E_i = \mathbb{R}$, $Q_i \subseteq E_i$ are closed and convex, $\|x^{(i)}\|_i^2 = (x^{(i)})^2$, $x^{(i)} \in E_i$, $d_i(x^{(i)}) = \frac{1}{2}(x^{(i)})^2$, $x^{(i)} \in Q_i$, and, hence, $V_i[z^{(i)}](x^{(i)}) = \frac{1}{2}(x^{(i)} - z^{(i)})^2$, $x^{(i)}, z^{(i)} \in Q_i$. Thus, $Q = \otimes_{i=1}^n Q_i$ has separable structure.

Let us denote $e_i \in E$ the i-th coordinate vector. Then, for $i = 1, ..., n$, the i-th coordinate derivative of f is $f_i'(x) = \langle \nabla f(x), e_i \rangle$. We assume that the gradient of f in (4) is coordinate-wise Lipschitz continuous with constants L_i, $i = 1, ..., n$, i.e.

$$|f_i'(x + he_i) - f_i'(x)| \leq L_i|h|, \quad h \in \mathbb{R}, \quad i = 1, ..., n, \quad x \in Q. \tag{38}$$

We set $\beta_i = L_i$, $i = 1, ..., n$. Then, by definitions in Subsection 1.1, we have $\|x\|_E^2 = \sum_{i=1}^n L_i (x^{(i)})^2$, $x \in E$, $d(x) = \frac{1}{2}\sum_{i=1}^n L_i (x^{(i)})^2$, $x \in Q$, $V[z](x) = \frac{1}{2}\sum_{i=1}^n L_i (x^{(i)} - z^{(i)})^2$, $x, z \in Q$. Also, we have $\|g\|_{E,*}^2 = \sum_{i=1}^n L_i^{-1} (g^{(i)})^2$, $g \in E^*$.

We assume that, at any point $x \in Q$, one can calculate an inexact coordinate derivative of f

$$\tilde{f}_i'(x) = \langle \nabla f(x), e_i \rangle + \xi(x),$$

where the coordinate i is chosen from $i = 1, ..., n$ at random with uniform probability $\frac{1}{n}$, the coordinate derivative error $\xi(x) \in \mathbb{R}$ is uniformly bounded in absolute value by Δ, i.e. $|\xi(x)| \leq \Delta$, $x \in Q$. Since we are in the Euclidean setting, we consider e_i also as an element of E^*. We use $n(\langle \nabla f(x), e_i \rangle + \xi(x))e_i$ as Randomized Inexact Oracle.

Let us check the assumptions stated in Subsection 1.2.

Randomized Inexact Oracle. In this setting, we have $\rho = n$, $H = E_i = \mathbb{R}$, $\mathcal{R}_p^T : E^* \to \mathbb{R}$ is given by $\mathcal{R}_p^T g = \langle g, e_i \rangle$, $g \in E^*$, $\mathcal{R}_r : \mathbb{R} \to E^*$ is given by $\mathcal{R}_r t = te_i$, $t \in \mathbb{R}$. Thus,

$$\widetilde{\nabla} f(x) = n(\langle \nabla f(x), e_i \rangle + \xi(x))e_i, \quad x \in Q.$$

One can prove that $\mathbb{E}_i n \langle \nabla f(x), e_i \rangle e_i = n\mathbb{E}_i e_i e_i^T \nabla f(x) = \nabla f(x)$, $x \in Q$, and, thus, (6) holds. Also, for all $x \in Q$, we have $\|\mathcal{R}_r \xi(x)\|_{E,*} = \frac{1}{\sqrt{L_i}}|\xi(x)| \leq \frac{\Delta}{\sqrt{L_0}}$, where $L_0 = \min_{i=1,...,n} L_i$. This proves (7) with $\delta = \frac{\Delta}{\sqrt{L_0}}$.

Regularity of Prox-Mapping. Separable structure of Q and $V[u](x)$ means that the problem (8) boils down to n independent problems of the form

$$u_+^{(j)} = \arg\min_{x^{(j)} \in Q_j} \left\{ \frac{L_j}{2}(u^{(j)} - x^{(j)})^2 + \alpha \langle \widetilde{\nabla} f(y), e_j \rangle x^{(j)} \right\}, \quad j = 1, ..., n.$$

Since $\widetilde{\nabla} f(y)$ has only one, i-th, non-zero component, $\langle \widetilde{\nabla} f(y), e_j \rangle$ is zero for all $j \neq i$. Thus, $u - u_+$ has one, i-th, non-zero component and $\langle e_i, u - u_+ \rangle e_i = u - u_+$. Hence,

$$\begin{aligned} \langle \mathcal{R}_r \mathcal{R}_p^T \nabla f(y), u - u_+ \rangle &= \langle \langle \nabla f(y), e_i \rangle e_i, u - u_+ \rangle \\ &= \langle \nabla f(y), e_i \rangle \langle e_i, u - u_+ \rangle \\ &= \langle \nabla f(y), \langle e_i, u - u_+ \rangle e_i \rangle \\ &= \langle \nabla f(y), u - u_+ \rangle, \end{aligned}$$

which proves (9).

Smoothness. By the standard reasoning, using (38), one can prove that, for all $i = 1, ..., n$,

$$f(x + he_i) \leq f(x) + h\langle \nabla f(x), e_i \rangle + \frac{L_i h^2}{2}, \quad h \in \mathbb{R}, \quad x \in Q. \tag{39}$$

Let $u, y \in Q$, $a \in \mathbb{R}$, and $x = y + a(u_+ - u) \in Q$. As we have shown above, $u_+ - u$ has only one, i-th, non-zero component. Hence, there exists $h \in \mathbb{R}$, such that $u_+ - u = he_i$ and $x = y + ahe_i$. Thus, by definition of $\| \cdot \|_E$ and (39), we have

$$\begin{aligned} f(x) = f(y + ahe_i) &\leq f(y) + ah\langle \nabla f(y), e_i \rangle + \frac{L_i}{2}(ah)^2 \\ &= f(y) + \langle \nabla f(y), ahe_i \rangle + \frac{1}{2}\|ahe_i\|_E^2 \\ &= f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}\|x - y\|_E^2. \end{aligned}$$

This proves (13).

We have checked that all the assumptions listed in Subsection 1.2 hold. Thus, we can obtain the following convergence rate result for random coordinate descent as a corollary of Theorem 1 and Lemma 5.

Corollary 2. *Let Algorithm 1 with $\widetilde{\nabla} f(x) = n(\langle \nabla f(x), e_i \rangle + \xi(x))e_i$, where i is uniformly at random chosen from $1, ..., n$, be applied to Problem (4) in the setting of this subsection. Let f_* be the optimal objective value and x_* be an optimal point in Problem (4). Assume that coordinate derivative error $\xi(x)$ satisfies $|\xi(x)| \leq \Delta$, $x \in Q$. Denote*

$$P_0^2 = \left(1 - \frac{1}{n}\right)(f(x_0) - f_*) + \sum_{i=1}^n \frac{L_i}{2}(u_0^{(i)} - x_*^{(i)})^2.$$

1. *If the coordinate derivative error $\xi(x)$ can be controlled and, on each iteration, the error level Δ satisfies*

$$\Delta \leq \frac{P_0\sqrt{L_0}}{4nA_k},$$

then, for all $k \geq 1$,

$$\mathbb{E}f(x_k) - f_* \leq \frac{6n^2P_0^2}{(k-1+2n)^2},$$

where $\mathbb{E}$ denotes the expectation with respect to all the randomness up to step k.

2. *If the coordinate derivative error $\xi(x)$ can not be controlled, then, for all $k \geq 1$,*

$$\mathbb{E}f(x_k) - f_* \leq \frac{8n^2P_0^2}{(k-1+2n)^2} + \frac{4}{L_0}(k-1+2n)^2\Delta^2.$$

Remark 3. According to Remark 1 and due to the relation $\delta = \frac{\Delta}{\sqrt{L_0}}$, we obtain that the error level Δ in the coordinate derivative should satisfy

$$\Delta \leq \frac{\varepsilon\sqrt{L_0}}{6nP_0}.$$

At the same time, to obtain an ε-solution for Problem (4), it is enough to choose

$$k = \max\left\{\left\lceil n\sqrt{\frac{6P_0^2}{\varepsilon}} + 1 - 2n\right\rceil, 0\right\}.$$

3.3 Accelerated Random Block-Coordinate Descent

In this subsection, we consider two block-coordinate settings. The first one is the Euclidean, which is usually used in the literature for accelerated block-coordinate descent. The second one is the entropy, which, to the best of our knowledge, is analyzed in this context for the first time. We develop accelerated random block-coordinate descent with inexact block derivatives for problems with simple constraints in these two settings and their combination.

Euclidean setup. We assume that, for all $i = 1, ..., n$, $E_i = \mathbb{R}^{p_i}$; Q_i is a simple closed convex set; $\|x^{(i)}\|_i^2 = \langle B_i x^{(i)}, x^{(i)}\rangle$, $x^{(i)} \in E_i$, where B_i is symmetric positive semidefinite matrix; $d_i(x^{(i)}) = \frac{1}{2}\|x^{(i)}\|_i^2$, $x^{(i)} \in Q_i$, and, hence, $V_i[z^{(i)}](x^{(i)}) = \frac{1}{2}\|x^{(i)} - z^{(i)}\|_i^2$, $x^{(i)}, z^{(i)} \in Q_i$.

Entropy setup. We assume that, for all $i = 1, ..., n$, $E_i = \mathbb{R}^{p_i}$; Q_i is standard simplex in $\mathbb{R}^{p_i}$, i.e., $Q_i = \{x^{(i)} \in \mathbb{R}_+^{p_i} : \sum_{j=1}^{p_i}[x^{(i)}]_j = 1\}$; $\|x^{(i)}\|_i = \|x^{(i)}\|_1 = \sum_{j=1}^{p_i}|[x^{(i)}]_j|$, $x^{(i)} \in E_i$; $d_i(x^{(i)}) = \sum_{j=1}^{p_i}[x^{(i)}]_j \ln[x^{(i)}]_j$, $x^{(i)} \in Q_i$, and, hence, $V_i[z^{(i)}](x^{(i)}) = \sum_{j=1}^{p_i}[x^{(i)}]_j \ln\frac{[x^{(i)}]_j}{[z^{(i)}]_j}$, $x^{(i)}, z^{(i)} \in Q_i$.

Note that, in each block, one also can choose other proximal setups from Ben-Tal and Nemirovski [2015]. Combination of different setups in different blocks is also possible, i.e., in one block it is possible to choose the Euclidean setup and in an another block one can choose the entropy setup.

Using operators U_i, $i = 1, ..., n$ defined in (1), for each $i = 1, ..., n$, the i-th block derivative of f can be written as $f_i'(x) = U_i^T \nabla f(x)$. We assume that the gradient of f in (4) is block-wise Lipschitz continuous with constants L_i, $i = 1, ..., n$ with respect to chosen norms $\|\cdot\|_i$, i.e.

$$\|f_i'(x + U_i h^{(i)}) - f_i'(x)\|_{i,*} \leq L_i \|h^{(i)}\|_i, \quad h^{(i)} \in E_i, \quad i = 1, ..., n, \quad x \in Q. \tag{40}$$

We set $\beta_i = L_i$, $i = 1, ..., n$. Then, by definitions in Subsection 1.1, we have $\|x\|_E^2 = \sum_{i=1}^n L_i \|x^{(i)}\|_i^2$, $x \in E$, $d(x) = \sum_{i=1}^n L_i d_i(x^{(i)})$, $x \in Q$, $V[z](x) = \sum_{i=1}^n L_i V_i[z^{(i)}](x^{(i)})$, $x, z \in Q$. Also, we have $\|g\|_{E,*}^2 = \sum_{i=1}^n L_i^{-1} \|g^{(i)}\|_{i,*}^2$, $g \in E^*$.

We assume that, at any point $x \in Q$, one can calculate an inexact block derivative of f

$$\tilde{f}_i'(x) = U_i^T \nabla f(x) + \xi(x),$$

where a block number i is chosen from $1, ..., n$ randomly uniformly, the block derivative error $\xi(x) \in E_i^*$ is uniformly bounded in norm by Δ, i.e. $\|\xi(x)\|_{i,*} \leq \Delta$, $x \in Q$, $i = 1, ..., n$. As Randomized Inexact Oracle, we use $n\widetilde{U}_i(U_i^T \nabla f(x) + \xi(x))$, where $\widetilde{U}_i$ is defined in (2).

Let us check the assumptions stated in Subsection 1.2.

Randomized Inexact Oracle. In this setting, we have $\rho = n$, $H = E_i$, $\mathcal{R}_p^T : E^* \to E_i^*$ is given by $\mathcal{R}_p^T g = U_i^T g$, $g \in E^*$, $\mathcal{R}_r : E_i^* \to E^*$ is given by $\mathcal{R}_r g^{(i)} = \widetilde{U}_i g^{(i)}$, $g^{(i)} \in E_i^*$. Thus,

$$\widetilde{\nabla} f(x) = n\widetilde{U}_i(U_i^T \nabla f(x) + \xi(x)), \quad x \in Q.$$

Since $i \in R[1, n]$, one can prove that $\mathbb{E}_i n\widetilde{U}_i U_i^T \nabla f(x) = \nabla f(x)$, $x \in Q$, and, thus, (6) holds. Also, for all $x \in Q$, we have $\|\mathcal{R}_r \xi(x)\|_{E,*} = \|\widetilde{U}_i \xi(x)\|_{E,*} = \frac{1}{\sqrt{L_i}} \|\xi(x)\|_{i,*} \leq \frac{\Delta}{\sqrt{L_0}}$, where $L_0 = \min_{i=1,...,n} L_i$. This proves (7) with $\delta = \frac{\Delta}{\sqrt{L_0}}$.

Regularity of Prox-Mapping. Separable structure of Q and $V[u](x)$ means that the problem (8) boils down to n independent problems of the form

$$u_+^{(j)} = \arg \min_{x^{(j)} \in Q_j} \left\{ L_j V[u^{(j)}](x^{(j)}) + \alpha \langle U_j^T \widetilde{\nabla} f(y), x^{(j)} \rangle \right\}, \quad j = 1, ..., n.$$

Since $\widetilde{\nabla} f(y)$ has non-zero components only in the block i, $U_j^T \widetilde{\nabla} f(y)$ is zero for all $j \neq i$. Thus, $u - u_+$ has non-zero components only in the block i and $U_i \widetilde{U}_i^T (u - u_+) = u - u_+$. Hence,

$$\begin{aligned} \langle \mathcal{R}_r \mathcal{R}_p^T \nabla f(y), u - u_+ \rangle &= \langle \widetilde{U}_i U_i^T \nabla f(y), u - u_+ \rangle \\ &= \langle \nabla f(y), U_i \widetilde{U}_i^T (u - u_+) \rangle \\ &= \langle \nabla f(y), u - u_+ \rangle, \end{aligned}$$

which proves (9).

Smoothness. By the standard reasoning, using (40), one can prove that, for all $i = 1, ..., n$,

$$f(x + U_i h^{(i)}) \leq f(x) + \langle U_i^T \nabla f(x), h^{(i)} \rangle + \frac{L_i}{2} \|h^{(i)}\|_i^2, \quad h^{(i)} \in E_i, \quad x \in Q. \tag{41}$$

Let $u, y \in Q$, $a \in \mathbb{R}$, and $x = y + a(u_+ - u) \in Q$. As we have shown above, $u_+ - u$ has non-zero components only in the block i. Hence, there exists $h^{(i)} \in E_i$, such that $u_+ - u = U_i h^{(i)}$ and $x = y + aU_i h^{(i)}$. Thus, by definition of $\|\cdot\|_E$ and (41), we have

$$\begin{aligned} f(x) = f(y + aU_i h^{(i)}) &\leq f(y) + \langle U_i^T \nabla f(y), a h^{(i)} \rangle + \frac{L_i}{2} \|a h^{(i)}\|_i^2 \\ &= f(y) + \langle \nabla f(y), a U_i h^{(i)} \rangle + \frac{1}{2} \|a U_i h^{(i)}\|_E^2 \\ &= f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2} \|x - y\|_E^2. \end{aligned}$$

This proves (13).

We have checked that all the assumptions listed in Subsection 1.2 hold. Thus, we can obtain the following convergence rate result for random block-coordinate descent as a corollary of Theorem 1 and Lemma 5.

Corollary 3. *Let Algorithm 1 with $\widetilde{\nabla} f(x) = n\widetilde{U}_i(U_i^T \nabla f(x) + \xi(x))$, where i is uniformly at random chosen from $1, ..., n$, be applied to Problem (4) in the setting of this subsection. Let f_* be the optimal objective value and x_* be an optimal point in Problem (4). Assume that block derivative error $\xi(x)$ satisfies $|\xi(x)| \leq \Delta$, $x \in Q$. Denote*

$$P_0^2 = \left(1 - \frac{1}{n}\right)(f(x_0) - f_*) + V[u_0](x_*).$$

1. *If the block derivative error $\xi(x)$ can be controlled and, on each iteration, the error level Δ satisfies*

$$\Delta \leq \frac{P_0 \sqrt{L_0}}{4nA_k},$$

then, for all $k \geq 1$,

$$\mathbb{E} f(x_k) - f_* \leq \frac{6n^2 P_0^2}{(k - 1 + 2n)^2},$$

where $\mathbb{E}$ denotes the expectation with respect to all the randomness up to step k.

2. *If the block derivative error $\xi(x)$ can not be controlled, then, for all $k \geq 1$,*

$$\mathbb{E} f(x_k) - f_* \leq \frac{8n^2 P_0^2}{(k - 1 + 2n)^2} + \frac{4}{L_0}(k - 1 + 2n)^2 \Delta^2.$$

Remark 4. According to Remark 1 and due to the relation $\delta = \frac{\Delta}{\sqrt{L_0}}$, we obtain that the block derivative error Δ should satisfy

$$\Delta \leq \frac{\varepsilon \sqrt{L_0}}{6nP_0}.$$

At the same time, to obtain an ε-solution for Problem (4), it is enough to choose

$$k = \max\left\{ \left\lceil n\sqrt{\frac{6P_0^2}{\varepsilon}} + 1 - 2n \right\rceil, 0 \right\}.$$

3.4 Accelerated Random Derivative-Free Directional Search

In this subsection, we consider the same setting as in Subsection 3.1, except for Randomized Inexact Oracle. Instead of directional derivative, we use here its finite-difference approximation. We assume that, for all $i = 1, ..., n$, $Q_i = E_i = \mathbb{R}$, $\|x^{(i)}\|_i = (x^{(i)})^2$, $x^{(i)} \in E_i$, $d_i(x^{(i)}) = \frac{1}{2}(x^{(i)})^2$, $x^{(i)} \in E_i$, and, hence, $V_i[z^{(i)}](x^{(i)}) = \frac{1}{2}(x^{(i)} - z^{(i)})^2$, $x^{(i)}, z^{(i)} \in E_i$. Thus, $Q = E = \mathbb{R}^n$. Further, we assume that f in (4) has L-Lipschitz-continuous gradient with respect to Euclidean norm, i.e.

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|_2^2, \quad x, y \in E. \tag{42}$$

We set $\beta_i = L$, $i = 1, ..., n$. Then, by definitions in Subsection 1.1, we have $\|x\|_E^2 = L\|x\|_2^2$, $x \in E$, $d(x) = \frac{L}{2}\|x\|_2^2 = \frac{1}{2}\|x\|_E^2$, $x \in E$, $V[z](x) = \frac{L}{2}\|x - z\|_2^2 = \frac{1}{2}\|x - z\|_E^2$, $x, z \in E$. Also, we have $\|g\|_{E,*}^2 = L^{-1}\|g\|_2^2$, $g \in E^*$.

We assume that, at any point $x \in E$, one can calculate an inexact value $\tilde{f}(x)$ of the function f, s.t. $|\tilde{f}(x) - f(x)| \leq \Delta$, $x \in E$. To approximate the gradient of f, we use

$$\widetilde{\nabla} f(x) = n\frac{\tilde{f}(x + \tau e) - \tilde{f}(x)}{\tau} e,$$

where $\tau > 0$ is small parameter, which will be chosen later, $e \in E$ is a random vector uniformly distributed on the Euclidean sphere of radius 1, i.e. on $\mathcal{S}_2(1) := \{s \in \mathbb{R}^n : \|s\|_2 = 1\}$. Since, we are in the Euclidean setting, we consider e also as an element of E^*.

Let us check the assumptions stated in Subsection 1.2.

Randomized Inexact Oracle. First, let us show that the finite-difference approximation for the gradient of f can be expressed in the form of (5). We have

$$\widetilde{\nabla} f(x) = n\frac{\tilde{f}(x + \tau e) - \tilde{f}(x)}{\tau} e = n\left(\langle \nabla f(x), e \rangle + \frac{1}{\tau}(\tilde{f}(x + \tau e) - \tilde{f}(x) - \tau\langle \nabla f(x), e \rangle)\right) e.$$

Taking $\rho = n$, $H = \mathbb{R}$, $\mathcal{R}_p^T : E^* \to \mathbb{R}$ be given by $\mathcal{R}_p^T g = \langle g, e \rangle$, $g \in E^*$, $\mathcal{R}_r : \mathbb{R} \to E^*$ be given by $\mathcal{R}_r t = te$, $t \in \mathbb{R}$, we obtain

$$\widetilde{\nabla} f(x) = n(\langle \nabla f(x), e \rangle + \xi(x))e,$$

where $\xi(x) = \frac{1}{\tau}(\tilde{f}(x + \tau e) - \tilde{f}(x) - \tau\langle \nabla f(x), e \rangle)$. One can prove that $\mathbb{E}_e n\langle \nabla f(x), e \rangle e = n\mathbb{E}_e e e^T \nabla f(x) = \nabla f(x)$, $x \in E$, and, thus, (6) holds. It remains to prove (7), i.e., find δ s.t. for all $x \in E$, we have $\|\mathcal{R}_r \xi(x)\|_{E,*} \leq \delta$.

$$\begin{aligned}
\|\mathcal{R}_r \xi(x)\|_{E,*} &= \frac{1}{\sqrt{L}}\|\xi(x)e\|_2 = \frac{1}{\sqrt{L}}\left\|\frac{1}{\tau}(\tilde{f}(x + \tau e) - \tilde{f}(x) - \tau\langle \nabla f(x), e \rangle)e\right\|_2 \\
&= \frac{1}{\sqrt{L}}\left\|\frac{1}{\tau}(\tilde{f}(x + \tau e) - f(x + \tau e) - (\tilde{f}(x) - f(x))\right\|_2 \\
&\quad + \frac{1}{\sqrt{L}}\left\|(f(x + \tau e) - f(x) - \tau\langle \nabla f(x), e \rangle))e\right\|_2 \\
&\leq \frac{2\Delta}{\tau\sqrt{L}} + \frac{\tau\sqrt{L}}{2}.
\end{aligned}$$

Here we used that $|\tilde{f}(x) - f(x)| \leq \Delta$, $x \in E$ and (42). So, we have that (7) holds with $\delta = \frac{2\Delta}{\tau\sqrt{L}} + \frac{\tau\sqrt{L}}{2}$. To balance both terms, we choose $\tau = 2\sqrt{\frac{\Delta}{L}}$, which leads to equality $\delta = 2\sqrt{\Delta}$.

Regularity of Prox-Mapping. This assumption can be checked in the same way as in Subsection 3.1.

Smoothness. This assumption can be checked in the same way as in Subsection 3.1.

We have checked that all the assumptions listed in Subsection 1.2 hold. Thus, we can obtain the following convergence rate result for random derivative-free directional search as a corollary of Theorem 1 and Lemma 5.

Corollary 4. *Let Algorithm 1 with $\widetilde{\nabla} f(x) = n\frac{\tilde{f}(x+\tau e)-\tilde{f}(x)}{\tau} e$, where e is random and uniformly distributed over the Euclidean sphere of radius 1, be applied to Problem (4) in the setting of this subsection. Let f_* be the optimal objective value and x_* be an optimal point in Problem (4). Assume that function value error $\tilde{f}(x) - f(x)$ satisfies $|\tilde{f}(x) - f(x)| \leq \Delta$, $x \in E$. Denote*

$$P_0^2 = \left(1 - \frac{1}{n}\right)(f(x_0) - f_*) + \frac{L}{2}\|u_0 - x_*\|_2^2.$$

1. *If the error in the value of the objective f can be controlled and, on each iteration, the error level Δ satisfies*

$$\Delta \leq \frac{P_0^2}{64n^2 A_k^2},$$

and $\tau = 2\sqrt{\frac{\Delta}{L}}$ then, for all $k \geq 1$,

$$\mathbb{E} f(x_k) - f_* \leq \frac{6n^2 P_0^2}{(k - 1 + 2n)^2},$$

where $\mathbb{E}$ denotes the expectation with respect to all the randomness up to step k.

2. *If the error in the value of the objective f can not be controlled and $\tau = 2\sqrt{\frac{\Delta}{L}}$, then, for all $k \geq 1$,*

$$\mathbb{E} f(x_k) - f_* \leq \frac{8n^2 P_0^2}{(k - 1 + 2n)^2} + 16(k - 1 + 2n)^2 L\Delta.$$

Remark 5. According to Remark 1 and due to the relation $\delta = 2\sqrt{\Delta}$, we obtain that the error level in the function value should satisfy

$$\Delta \leq \frac{\varepsilon^2}{144n^2 P_0^2}.$$

The parameter τ should satisfy

$$\tau \leq \frac{\varepsilon}{6nP_0\sqrt{L}}.$$

At the same time, to obtain an ε-solution for Problem (4), it is enough to choose

$$k = \max\left\{\left\lceil n\sqrt{\frac{6P_0^2}{\varepsilon}} + 1 - 2n\right\rceil, 0\right\}.$$

3.5 Accelerated Random Derivative-Free Coordinate Descent

In this subsection, we consider the same setting as in Subsection 3.2, except for Randomized Inexact Oracle. Instead of coordinate derivative, we use here its finite-difference approximation. We assume that, for all $i = 1, ..., n$, $E_i = \mathbb{R}$, $Q_i \subseteq E_i$ are closed and convex, $\|x^{(i)}\|_i = (x^{(i)})^2$, $x^{(i)} \in E_i$, $d_i(x^{(i)}) = \frac{1}{2}(x^{(i)})^2$, $x^{(i)} \in Q_i$, and, hence, $V_i[z^{(i)}](x^{(i)}) = \frac{1}{2}(x^{(i)} - z^{(i)})^2$, $x^{(i)}, z^{(i)} \in Q_i$. Thus, $Q = \otimes_{i=1}^n Q_i$ has separable structure.

Let us denote $e_i \in E$ the i-th coordinate vector. Then, for $i = 1, ..., n$, the i-th coordinate derivative of f is $f_i'(x) = \langle \nabla f(x), e_i \rangle$. We assume that the gradient of f in (4) is coordinate-wise Lipschitz continuous with constants L_i, $i = 1, ..., n$, i.e.

$$|f_i'(x + he_i) - f_i'(x)| \leq L_i|h|, \quad h \in \mathbb{R}, \quad i = 1, ..., n, \quad x \in Q. \tag{43}$$

We set $\beta_i = L_i$, $i = 1, ..., n$. Then, by definitions in Subsection 1.1, we have $\|x\|_E^2 = \sum_{i=1}^n L_i(x^{(i)})^2$, $x \in E$, $d(x) = \frac{1}{2}\sum_{i=1}^n L_i(x^{(i)})^2$, $x \in Q$, $V[z](x) = \frac{1}{2}\sum_{i=1}^n L_i(x^{(i)} - z^{(i)})^2$, $x, z \in Q$. Also, we have $\|g\|_{E,*}^2 = \sum_{i=1}^n L_i^{-1}(g^{(i)})^2$, $g \in E^*$.

We assume that, at any point x in a small vicinity $\bar{Q}$ of the set Q, one can calculate an inexact value $\tilde{f}(x)$ of the function f, s.t. $|\tilde{f}(x) - f(x)| \leq \Delta$, $x \in \bar{Q}$. To approximate the gradient of f, we use

$$\widetilde{\nabla} f(x) = n\frac{\tilde{f}(x + \tau e_i) - \tilde{f}(x)}{\tau} e_i,$$

where $\tau > 0$ is small parameter, which will be chosen later, and the coordinate i is chosen from $i = 1, ..., n$ randomly with uniform probability $\frac{1}{n}$. Since, we are in the Euclidean setting, we consider e_i also as an element of E^*.

Let us check the assumptions stated in Subsection 1.2.

Randomized Inexact Oracle. First, let us show that the finite-difference approximation for the gradient of f can be expressed in the form of (5). We have

$$\widetilde{\nabla} f(x) = n\frac{\tilde{f}(x + \tau e_i) - \tilde{f}(x)}{\tau} e_i = n\left(\langle \nabla f(x), e_i \rangle + \frac{1}{\tau}(\tilde{f}(x + \tau e_i) - \tilde{f}(x) - \tau\langle \nabla f(x), e_i \rangle)\right) e_i.$$

Taking $\rho = n$, $H = \mathbb{R}$, $\mathcal{R}_p^T : E^* \to \mathbb{R}$ is given by $\mathcal{R}_p^T g = \langle g, e_i \rangle$, $g \in E^*$, $\mathcal{R}_r : \mathbb{R} \to E^*$ is given by $\mathcal{R}_r t = te_i$, $t \in \mathbb{R}$, we obtain

$$\widetilde{\nabla} f(x) = n(\langle \nabla f(x), e_i \rangle + \xi(x))e_i,$$

where $\xi(x) = \frac{1}{\tau}(\tilde{f}(x + \tau e_i) - \tilde{f}(x) - \tau\langle \nabla f(x), e_i \rangle)$. One can prove that $\mathbb{E}_i n\langle \nabla f(x), e_i \rangle e_i = n\mathbb{E}_i e_i e_i^T \nabla f(x) = \nabla f(x)$, $x \in Q$, and, thus, (6) holds. It remains to prove (7), i.e., find δ s.t.

for all $x \in Q$, we have $\|\mathcal{R}_r\xi(x)\|_{E,*} \leq \delta$.

$$\begin{aligned}\|\mathcal{R}_r\xi(x)\|_{E,*} &= \frac{1}{\sqrt{L_i}}|\xi(x)| = \frac{1}{\sqrt{L_i}}\left|\frac{1}{\tau}(\tilde{f}(x+\tau e_i) - \tilde{f}(x) - \tau\langle\nabla f(x), e_i\rangle)\right| \\ &= \frac{1}{\tau\sqrt{L_i}}\left|\tilde{f}(x+\tau e_i) - f(x+\tau e_i) - (\tilde{f}(x) - f(x))\right| \\ &\quad + \frac{1}{\tau\sqrt{L_i}}\left|f(x+\tau e_i) - f(x) - \tau\langle\nabla f(x), e_i\rangle\right| \\ &\leq \frac{2\Delta}{\sqrt{L_i}\tau} + \frac{\tau\sqrt{L_i}}{2}.\end{aligned}$$

Here we used that $|\tilde{f}(x) - f(x)| \leq \Delta$, $x \in \bar{Q}$ and (39), which follows from (43). So, we obtain that (7) holds with $\delta = \frac{2\Delta}{\sqrt{L_i}\tau} + \frac{\sqrt{L_i}\tau}{2}$. To balance both terms, we choose $\tau = 2\sqrt{\frac{\Delta}{L_i}} \leq 2\sqrt{\frac{\Delta}{L_0}}$, where $L_0 = \min_{i=1,\ldots,n} L_i$. This leads to equality $\delta = 2\sqrt{\Delta}$.

Regularity of Prox-Mapping. This assumption can be checked in the same way as in Subsection 3.2.

Smoothness. This assumption can be checked in the same way as in Subsection 3.2.

We have checked that all the assumptions listed in Subsection 1.2 hold. Thus, we can obtain the following convergence rate result for random derivative-free coordinate descent as a corollary of Theorem 1 and Lemma 5.

Corollary 5. *Let Algorithm 1 with* $\widetilde{\nabla} f(x) = n\frac{\tilde{f}(x+\tau e_i) - \tilde{f}(x)}{\tau} e_i$, *where* i *is random and uniformly distributed in* $1, \ldots, n$, *be applied to Problem* (4) *in the setting of this subsection. Let* f_* *be the optimal objective value and* x_* *be an optimal point in Problem* (4). *Assume that function value error* $\tilde{f}(x) - f(x)$ *satisfies* $|\tilde{f}(x) - f(x)| \leq \Delta$, $x \in \bar{Q}$. *Denote*

$$P_0^2 = \left(1 - \frac{1}{n}\right)(f(x_0) - f_*) + \sum_{i=1}^{n} \frac{L_i}{2}(u_0^{(i)} - x_*^{(i)})^2.$$

1. *If the error in the value of the objective* f *can be controlled and, on each iteration, the error level* Δ *satisfies*

$$\Delta \leq \frac{P_0^2}{64n^2A_k^2},$$

and $\tau = 2\sqrt{\frac{\Delta}{L_0}}$ *then, for all* $k \geq 1$,

$$\mathbb{E}f(x_k) - f_* \leq \frac{6n^2P_0^2}{(k-1+2n)^2},$$

where $\mathbb{E}$ *denotes the expectation with respect to all the randomness up to step* k.

2. *If the error in the value of the objective* f *can not be controlled and* $\tau = 2\sqrt{\frac{\Delta}{L_0}}$, *then, for all* $k \geq 1$,

$$\mathbb{E}f(x_k) - f_* \leq \frac{8n^2P_0^2}{(k-1+2n)^2} + 16(k-1+2n)^2\Delta.$$

Remark 6. According to Remark 1 and due to the relation $\delta = 2\sqrt{\Delta}$, we obtain that the error level in the function value should satisfy

$$\Delta \leq \frac{\varepsilon^2}{144 n^2 P_0^2}.$$

The parameter τ should satisfy

$$\tau \leq \frac{\varepsilon}{6 n P_0 \sqrt{L_0}}.$$

At the same time, to obtain an ε-solution for Problem (4), it is enough to choose

$$k = \max\left\{\left\lceil n\sqrt{\frac{6P_0^2}{\varepsilon}} + 1 - 2n \right\rceil, 0\right\}.$$

3.6 Accelerated Random Derivative-Free Block-Coordinate Descent

In this subsection, we consider the same setting as in Subsection 3.3, except for Randomized Inexact Oracle. Instead of block derivative, we use here its finite-difference approximation. As in Subsection 3.3, we consider Euclidean setup and entropy setup.

Euclidean setup. We assume that, for all $i = 1, ..., n$, $E_i = \mathbb{R}^{p_i}$; Q_i is a simple closed convex set; $\|x^{(i)}\|_i^2 = \langle B_i x^{(i)}, x^{(i)} \rangle$, $x^{(i)} \in E_i$, where B_i is symmetric positive semidefinite matrix; $d_i(x^{(i)}) = \frac{1}{2}\|x^{(i)}\|_i^2$, $x^{(i)} \in Q_i$, and, hence, $V_i[z^{(i)}](x^{(i)}) = \frac{1}{2}\|x^{(i)} - z^{(i)}\|_i^2$, $x^{(i)}, z^{(i)} \in Q_i$.

Entropy setup. We assume that, for all $i = 1, ..., n$, $E_i = \mathbb{R}^{p_i}$; Q_i is standard simplex in $\mathbb{R}^{p_i}$, i.e., $Q_i = \{x^{(i)} \in \mathbb{R}_+^{p_i} : \sum_{j=1}^{p_i} [x^{(i)}]_j = 1\}$; $\|x^{(i)}\|_i = \|x^{(i)}\|_1 = \sum_{j=1}^{p_i} |[x^{(i)}]_j|$, $x^{(i)} \in E_i$; $d_i(x^{(i)}) = \sum_{j=1}^{p_i} [x^{(i)}]_j \ln [x^{(i)}]_j$, $x^{(i)} \in Q_i$, and, hence, $V_i[z^{(i)}](x^{(i)}) = \sum_{j=1}^{p_i} [x^{(i)}]_j \ln \frac{[x^{(i)}]_j}{[z^{(i)}]_j}$, $x^{(i)}, z^{(i)} \in Q_i$.

Note that, in each block, one also can choose other proximal setups from Ben-Tal and Nemirovski [2015]. Combination of different setups in different blocks is also possible, i.e., in one block it is possible to choose the Euclidean setup and in an another block one can choose the entropy setup.

Using operators U_i, $i = 1, ..., n$ defined in (1), for each $i = 1, ..., n$, the i-th block derivative of f can be written as $f_i'(x) = U_i^T \nabla f(x)$. We assume that the gradient of f in (4) is block-wise Lipschitz continuous with constants L_i, $i = 1, ..., n$ with respect to chosen norms $\|\cdot\|_i$, i.e.,

$$\|f_i'(x + U_i h^{(i)}) - f_i'(x)\|_{i,*} \leq L_i \|h^{(i)}\|_i, \quad h^{(i)} \in E_i, \quad i = 1, ..., n \quad x \in Q. \tag{44}$$

We set $\beta_i = L_i$, $i = 1, ..., n$. Then, by definitions in Subsection 1.1, we have $\|x\|_E^2 = \sum_{i=1}^n L_i \|x^{(i)}\|_2^2$, $x \in E$, $d(x) = \frac{1}{2}\sum_{i=1}^n L_i \|x^{(i)}\|_2^2$, $x \in Q$, $V[z](x) = \frac{1}{2}\sum_{i=1}^n L_i \|x^{(i)} - z^{(i)}\|_2^2$, $x, z \in Q$. Also, we have $\|g\|_{E,*}^2 = \sum_{i=1}^n L_i^{-1} \|g^{(i)}\|_2^2$, $g \in E^*$.

We assume that, at any point x in a small vicinity $\bar{Q}$ of the set Q, one can calculate an inexact value $\tilde{f}(x)$ of the function f, s.t. $|\tilde{f}(x) - f(x)| \leq \Delta$, $x \in \bar{Q}$. To approximate the

gradient of f, we use

$$\widetilde{\nabla} f(x) = n\widetilde{U}_i \left(\frac{\tilde{f}(x+\tau U_i e_1) - \tilde{f}(x)}{\tau}, ..., \frac{\tilde{f}(x+\tau U_i e_{p_i}) - \tilde{f}(x)}{\tau} \right)^T, \tag{45}$$

where $\tau > 0$ is small parameter, which will be chosen later, a block number i is chosen from $i = 1, ..., n$ randomly with uniform probability $\frac{1}{n}$, $e_1, ..., e_{p_i}$ are coordinate vectors in E_i, U_i is defined in (1), $\widetilde{U}_i$ is defined in (2).

Let us check the assumptions stated in Subsection 1.2.

Randomized Inexact Oracle. First, let us show that the random derivative-free block-coordinate approximation for the gradient of f can be expressed in the form of (5). Denote $\tilde{g}_i = \frac{1}{\tau} \left(\tilde{f}(x+\tau U_i e_1) - \tilde{f}(x), ..., \tilde{f}(x+\tau U_i e_{p_i}) - \tilde{f}(x) \right)^T \in E_i$, $i = 1, ..., n$. We have

$$\widetilde{\nabla} f(x) = n\widetilde{U}_i \tilde{g}_i = n\widetilde{U}_i \left(U_i^T \nabla f(x) + (\tilde{g}_i - U_i^T \nabla f(x)) \right).$$

Taking $\rho = n$, $H = E_i$, $\mathcal{R}_p^T : E^* \to E_i^*$ be given by $\mathcal{R}_p^T g = U_i^T g$, $g \in E^*$ and $\mathcal{R}_r : E_i^* \to E^*$ be given by $\mathcal{R}_r g^{(i)} = \widetilde{U}_i g^{(i)}$, $g^{(i)} \in E_i^*$, we obtain

$$\widetilde{\nabla} f(x) = n\widetilde{U}_i (U_i^T \nabla f(x) + \xi(x)),$$

where $\xi(x) = \tilde{g}_i - U_i^T \nabla f(x)$. Since $i \in R[1, n]$, one can prove that $\mathbb{E}_i n\widetilde{U}_i U_i^T \nabla f(x) = \nabla f(x)$, $x \in Q$, and, thus, (6) holds. It remains to prove (7), i.e., find δ s.t. for all $x \in Q$, we have $\|\mathcal{R}_r \xi(x)\|_{E,*} \leq \delta$. Let us fix any i from $1, ..., n$ and j from $1, ..., p_i$. Then, for any $x \in \bar{Q}$, the j-th coordinate of $\xi(x) = \tilde{g}_i - U_i^T \nabla f(x)$ can be estimated as follows

$$\begin{aligned}
|[\xi(x)]_j| &= \left| \frac{1}{\tau} (\tilde{f}(x+\tau U_i e_j) - \tilde{f}(x) - \tau \langle U_i^T \nabla f(x), e_j \rangle \right| \\
&= \frac{1}{\tau} \left| \tilde{f}(x+\tau U_i e_j) - f(x+\tau U_i e_j) - (\tilde{f}(x) - f(x)) \right| \\
&\quad + \frac{1}{\tau} \left| (f(x+\tau U_i e_j) - f(x) - \tau \langle U_i^T \nabla f(x), e_j \rangle) \right| \\
&\leq \frac{2\Delta}{\tau} + \frac{\tau L_i}{2}.
\end{aligned} \tag{46}$$

Here we used that $|\tilde{f}(x) - f(x)| \leq \Delta$, $x \in \bar{Q}$, (41), which follows from (44). In our setting, for any $i = 1, ..., n$, $\|\cdot\|_{i,*}$ is either max-norm (for the entropy case) or Euclidean norm (for the Euclidean case). Thus, in the worst case of Euclidean norm

$$\|\mathcal{R}_r \xi(x)\|_{E,*} = \|\widetilde{U}_i \xi(x)\|_{E,*} = \frac{1}{\sqrt{L_i}} \|\xi(x)\|_{i,*} \overset{(46)}{\leq} \frac{\sqrt{p_i}}{\sqrt{L_i}} \left(\frac{2\Delta}{\tau} + \frac{\tau L_i}{2} \right) \leq \sqrt{p_{\max}} \left(\frac{2\Delta}{\tau\sqrt{L_i}} + \frac{\tau\sqrt{L_i}}{2} \right),$$

where $p_{\max} = \max_{i=1,...,n} p_i$. So, we obtain that (7) holds with $\delta = \sqrt{p_{\max}} \left(\frac{2\Delta}{\tau\sqrt{L_i}} + \frac{\tau\sqrt{L_i}}{2} \right)$. To balance both terms we choose $\tau = 2\sqrt{\frac{\Delta}{L_i}} \leq 2\sqrt{\frac{\Delta}{L_0}}$, where $L_0 = \min_{i=1,...,n} L_i$. This leads to equality $\delta = 2\sqrt{p_{\max}\Delta}$.

Regularity of Prox-Mapping. This assumption can be checked in the same way as in Subsection 3.3.

Smoothness. This assumption can be checked in the same way as in Subsection 3.3.

We have checked that all the assumptions listed in Subsection 1.2 hold. Thus, we can obtain the following convergence rate result for random derivative-free block-coordinate descent as a corollary of Theorem 1 and Lemma 5.

Corollary 6. *Let Algorithm 1 with $\widetilde{\nabla} f(x)$ defined in (45), be applied to Problem (4) in the setting of this subsection. Let f_* be the optimal objective value and x_* be an optimal point in Problem (4). Assume that function value error $\tilde{f}(x) - f(x)$ satisfies $|\tilde{f}(x) - f(x)| \leq \Delta$, $x \in \bar{Q}$. Denote*

$$P_0^2 = \left(1 - \frac{1}{n}\right)(f(x_0) - f_*) + V[u_0](x_*).$$

1. *If the error in the value of the objective f can be controlled and, on each iteration, the error level Δ satisfies*

$$\Delta \leq \frac{P_0^2}{64 n^2 p_{\max} A_k^2},$$

and $\tau = 2\sqrt{\frac{\Delta}{L_0}}$ then, for all $k \geq 1$,

$$\mathbb{E} f(x_k) - f_* \leq \frac{6 n^2 P_0^2}{(k - 1 + 2n)^2},$$

where $\mathbb{E}$ denotes the expectation with respect to all the randomness up to step k.

2. *If the error in the value of the objective f can not be controlled and $\tau = 2\sqrt{\frac{\Delta}{L_0}}$, then, for all $k \geq 1$,*

$$\mathbb{E} f(x_k) - f_* \leq \frac{8 n^2 P_0^2}{(k - 1 + 2n)^2} + 16 (k - 1 + 2n)^2 p_{\max} \Delta.$$

Remark 7. According to Remark 1 and due to the relation $\delta = 2\sqrt{p_{\max} \Delta}$, we obtain that the error level in the function value should satisfy

$$\Delta \leq \frac{\varepsilon^2}{144 n^2 p_{\max} P_0^2}.$$

The parameter τ should satisfy

$$\tau \leq \frac{\varepsilon}{6 n P_0 \sqrt{L_0}}.$$

At the same time, to obtain an ε-solution for Problem (4), it is enough to choose

$$k = \max\left\{\left\lceil n\sqrt{\frac{6P_0^2}{\varepsilon}} + 1 - 2n \right\rceil, 0\right\}.$$

3.7 Accelerated Random Derivative-Free Block-Coordinate Descent with Random Approximations for Block Derivatives

In this subsection, we combine random block-coordinate descent of Subsection 3.3 with random derivative-free directional search described in Subsection 3.4 and random derivative-free coordinate descent described in Subsection 3.5. We construct randomized approximations for block derivatives based on finite-difference approximation of directional derivatives. Unlike Subsection 3.3, we consider only Euclidean setup. We assume that, for all $i = 1, ..., n$, $E_i = \mathbb{R}^{p_i}$; $\|x^{(i)}\|_i^2 = \|x^{(i)}\|_2^2$, $x^{(i)} \in E_i$; Q_i is either E_i, or $\otimes_{j=1}^{p_i} Q_{ij}$, where $Q_{ij} \subseteq \mathbb{R}$ are closed convex sets; $d_i(x^{(i)}) = \frac{1}{2}\|x^{(i)}\|_i^2$, $x^{(i)} \in Q_i$ and, hence, $V_i[z^{(i)}](x^{(i)}) = \frac{1}{2}\|x^{(i)} - z^{(i)}\|_i^2$, , $x^{(i)}, z^{(i)} \in Q_i$. For the case, $Q_i = E_i$, we consider randomization on the Euclidean sphere of radius 1, as in Subsection 3.4. For the case, $Q_i = \otimes_{j=1}^{p_i} Q_{ij}$, we consider coordinate-wise randomization, as in Subsection 3.5.

Using operators U_i, $i = 1, ..., n$ defined in (1), for each $i = 1, ..., n$, the i-th block derivative of f can be written as $f_i'(x) = U_i^T \nabla f(x)$. We assume that the gradient of f in (4) is block-wise Lipschitz continuous with constants L_i, $i = 1, ..., n$ with respect to chosen norms $\|\cdot\|_i$, i.e.,

$$\|f_i'(x + U_i h^{(i)}) - f_i'(x)\|_{i,*} \leq L_i \|h^{(i)}\|_i, \quad h^{(i)} \in E_i, \quad i = 1, ..., n \quad x \in Q. \tag{47}$$

We set $\beta_i = L_i$, $i = 1, ..., n$. Then, by definitions in Subsection 1.1, we have $\|x\|_E^2 = \sum_{i=1}^n L_i \|x^{(i)}\|_2^2$, $x \in E$, $d(x) = \frac{1}{2}\sum_{i=1}^n L_i \|x^{(i)}\|_2^2$, $x \in Q$, $V[z](x) = \frac{1}{2}\sum_{i=1}^n L_i \|x^{(i)} - z^{(i)}\|_2^2$, $x, z \in Q$. Also, we have $\|g\|_{E,*}^2 = \sum_{i=1}^n L_i^{-1} \|g^{(i)}\|_2^2$, $g \in E^*$.

We assume that, at any point x in a small vicinity $\bar{Q}$ of the set Q, one can calculate an inexact value $\tilde{f}(x)$ of the function f, s.t. $|\tilde{f}(x) - f(x)| \leq \Delta$, $x \in \bar{Q}$. To approximate the gradient of f, we first randomly choose a block $i \in 1, ..., n$ with probability p_i/p, where $p = \sum_{i=1}^n p_i$. Then we use one the following types of random directions $e \in E_i$ to approximate the block derivative $f_i'(x)$ by a finite-difference.

1. If $Q_i = E_i$, we take $e \in E_i$ to be random vector uniformly distributed on the Euclidean sphere of radius 1, i.e. $\mathcal{S}_2(1) := \{s \in \mathbb{R}^{p_i} : \|s\|_2 = 1\}$. We call this *unconstrained case*.

2. If $Q_i = \otimes_{j=1}^{p_i} Q_{ij}$, we take e to be random uniformly chosen from $1, ..., p_i$ coordinate vector, i.e. $e = e_j \in E_i$ with probability $\frac{1}{p_i}$. We call this *separable case*.

Based on these randomizations and inexact function values, our randomized approximation for the gradient of f is

$$\widetilde{\nabla} f(x) = p\widetilde{U}_i \frac{\tilde{f}(x + \tau U_i e) - \tilde{f}(x)}{\tau} e,$$

where $\tau > 0$ is small parameter, which will be chosen later, U_i is defined in (1) and $\widetilde{U}_i$ is defined in (2).

Let us check the assumptions stated in Subsection 1.2.

Randomized Inexact Oracle. First, let us show that the random derivative-free block-coordinate approximation for the gradient of f can be expressed in the form of (5). We have

$$\begin{aligned}\widetilde{\nabla} f(x) &= p\widetilde{U}_i \frac{\tilde{f}(x+\tau U_i e) - \tilde{f}(x)}{\tau} e \\ &= p\widetilde{U}_i \left(\langle U_i^T \nabla f(x), e\rangle e + \frac{1}{\tau}(\tilde{f}(x+\tau U_i e) - \tilde{f}(x) - \tau\langle U_i^T \nabla f(x), e\rangle) e \right). \end{aligned} \tag{48}$$

Taking $\rho = p$, $H = E_i$, $\mathcal{R}_p^T : E^* \to E_i^*$ be given by $\mathcal{R}_p^T g = \langle U_i^T g, e\rangle e$, $g \in E^*$ and $\mathcal{R}_r : E_i^* \to E^*$ be given by $\mathcal{R}_r g^{(i)} = \widetilde{U}_i g^{(i)}$, $g^{(i)} \in E_i^*$, we obtain

$$\widetilde{\nabla} f(x) = p\widetilde{U}_i(\langle U_i^T \nabla f(x), e\rangle e + \xi(x)),$$

where $\xi(x) = \frac{1}{\tau}(\tilde{f}(x+\tau U_i e) - \tilde{f}(x) - \tau\langle U_i^T \nabla f(x), e\rangle) e$. By the choice of probability distributions for i and e and their independence, we have, for all $x \in Q$,

$$\begin{aligned}\mathbb{E}_{i,e} p\widetilde{U}_i \langle U_i^T \nabla f(x), e\rangle e &= p\mathbb{E}_{i,e} \widetilde{U}_i e e^T U_i^T \nabla f(x) = p\mathbb{E}_i \widetilde{U}_i \left(E_e e e^T\right) U_i^T \nabla f(x) \\ &= p\mathbb{E}_i \frac{1}{p_i} \widetilde{U}_i U_i^T \nabla f(x) = \nabla f(x)\end{aligned}$$

and, thus, (6) holds. It remains to prove (7), i.e., find δ s.t. for all $x \in Q$, we have $\|\mathcal{R}_r \xi(x)\|_{E,*} \le \delta$. We have

$$\begin{aligned}&\|\mathcal{R}_r \xi(x)\|_{E,*} = \|\widetilde{U}_i \xi(x)\|_{E,*} = \frac{1}{\sqrt{L_i}} \left\| \frac{1}{\tau}(\tilde{f}(x+\tau U_i e) - \tilde{f}(x) - \tau\langle U_i^T \nabla f(x), e\rangle) e \right\|_{i,*} \\ &= \frac{\|e\|_{i,*}}{\tau\sqrt{L_i}} \left| \tilde{f}(x+\tau U_i e) - f(x+\tau U_i e) - (\tilde{f}(x) - f(x)) + (f(x+\tau U_i e) - f(x) - \tau\langle U_i^T \nabla f(x), e\rangle) \right| \\ &\le \frac{2\Delta \|e\|_{i,*}}{\tau\sqrt{L_i}} + \frac{\tau \|e\|_{i,*} \|e\|_i^2 \sqrt{L_i}}{2} \\ &= \frac{2\Delta}{\tau\sqrt{L_i}} + \frac{\tau\sqrt{L_i}}{2}.\end{aligned}$$

Here we used that $|\tilde{f}(x) - f(x)| \le \Delta$, $x \in \bar{Q}$, (41), which follows from (47), and that the norms $\|\cdot\|_i$, $\|\cdot\|_{i,*}$ are standard Euclidean. So, we obtain that (7) holds with $\delta = \frac{2\Delta}{\tau\sqrt{L_i}} + \frac{\tau\sqrt{L_i}}{2}$. To balance both terms we choose $\tau = 2\sqrt{\frac{\Delta}{L_i}} \le 2\sqrt{\frac{\Delta}{L_0}}$, where $L_0 = \min_{i=1,\dots,n} L_i$. This leads to equality $\delta = 2\sqrt{\Delta}$.

Regularity of Prox-Mapping. Separable structure of Q and $V[u](x)$ means that the problem (8) boils down to n independent problems of the form

$$u_+^{(l)} = \arg\min_{x^{(l)} \in Q_l} \left\{ \frac{L_l}{2} \|u^{(l)} - x^{(l)}\|_2^2 + \alpha\langle U_l^T \widetilde{\nabla} f(y), x^{(l)}\rangle \right\}, \quad l = 1, \dots, n.$$

Since $\widetilde{\nabla} f(y)$ has non-zero components only in the block i, $U_l^T \widetilde{\nabla} f(y)$ is zero for all $l \neq i$. Thus, $u - u_+$ has non-zero components only in the block i and $U_i(u^{(i)} - u_+^{(i)}) = u - u_+$.

In the unconstrained case, similarly to Subsection 3.1, we obtain that $u^{(i)} - u_+^{(i)} = \gamma e$, where γ is some constant. Using these two facts, we obtain

$$\begin{aligned}
\langle \mathcal{R}_r \mathcal{R}_p^T \nabla f(y), u - u_+ \rangle &= \langle \widetilde{U}_i \langle U_i^T \nabla f(y), e \rangle e, u - u_+ \rangle \\
&= \langle \langle U_i^T \nabla f(y), e \rangle e, \widetilde{U}_i^T (u - u_+) \rangle \\
&= \langle \langle U_i^T \nabla f(y), e \rangle e, u^{(i)} - u_+^{(i)} \rangle \\
&= \langle \langle U_i^T \nabla f(y), e \rangle e, \gamma e \rangle \\
&= \langle U_i^T \nabla f(y), \gamma e \rangle \langle e, e \rangle \\
&= \langle U_i^T \nabla f(y), u^{(i)} - u_+^{(i)} \rangle \\
&= \langle \nabla f(y), U_i(u^{(i)} - u_+^{(i)}) \rangle \\
&= \langle \nabla f(y), u - u_+ \rangle,
\end{aligned}$$

which proves (9) for the unconstrained case.

In the separable case, similarly to Subsection 3.2, we obtain that $u^{(i)} - u_+^{(i)}$ has only one j-th non-zero coordinate, where $j \in 1, ..., p_i$. Hence, $\langle e_j, u^{(i)} - u_+^{(i)} \rangle e_j = u^{(i)} - u_+^{(i)}$. So, we get,

$$\begin{aligned}
\langle \mathcal{R}_r \mathcal{R}_p^T \nabla f(y), u - u_+ \rangle &= \langle \widetilde{U}_i \langle U_i^T \nabla f(y), e_j \rangle e_j, u - u_+ \rangle \\
&= \langle \langle U_i^T \nabla f(y), e_j \rangle e_j, \widetilde{U}_i^T (u - u_+) \rangle \\
&= \langle \langle U_i^T \nabla f(y), e_j \rangle e_j, u^{(i)} - u_+^{(i)} \rangle \\
&= \langle U_i^T \nabla f(y), e_j \rangle \langle e_j, u^{(i)} - u_+^{(i)} \rangle \\
&= \langle U_i^T \nabla f(y), \langle e_j, u^{(i)} - u_+^{(i)} \rangle e_j \rangle \\
&= \langle U_i^T \nabla f(y), u^{(i)} - u_+^{(i)} \rangle \\
&= \langle \nabla f(y), U_i(u^{(i)} - u_+^{(i)}) \rangle \\
&= \langle \nabla f(y), u - u_+ \rangle,
\end{aligned}$$

which proves (9) for the separable case.

Smoothness. This assumption can be checked in the same way as in Subsection 3.3.

We have checked that all the assumptions listed in Subsection 1.2 hold. Thus, we can obtain the following convergence rate result for random derivative-free block-coordinate descent with random approximations for block derivatives as a corollary of Theorem 1 and Lemma 5.

Corollary 7. *Let Algorithm 1 with $\widetilde{\nabla} f(x)$ defined in* (48)*, be applied to Problem* (4) *in the setting of this subsection. Let f_* be the optimal objective value and x_* be an optimal point in Problem* (4)*. Assume that function value error $\tilde{f}(x) - f(x)$ satisfies $|\tilde{f}(x) - f(x)| \leq \Delta$,*

$x \in \bar{Q}$. *Denote*

$$P_0^2 = \left(1 - \frac{1}{p}\right)(f(x_0) - f_*) + \sum_{i=1}^{n} \frac{L_i}{2} \|u_0^{(i)} - x_*^{(i)}\|_2^2.$$

1. *If the error in the value of the objective f can be controlled and, on each iteration, the error level Δ satisfies*

$$\Delta \leq \frac{P_0^2}{64p^2A_k^2},$$

and $\tau = 2\sqrt{\frac{\Delta}{L_0}}$ then, for all $k \geq 1$,

$$\mathbb{E}f(x_k) - f_* \leq \frac{6p^2P_0^2}{(k-1+2p)^2},$$

where $\mathbb{E}$ denotes the expectation with respect to all the randomness up to step k.

2. *If the error in the value of the objective f can not be controlled and $\tau = 2\sqrt{\frac{\Delta}{L_0}}$, then, for all $k \geq 1$,*

$$\mathbb{E}f(x_k) - f_* \leq \frac{8p^2P_0^2}{(k-1+2p)^2} + 16(k-1+2p)^2\Delta.$$

Remark 8. According to Remark 1 and due to the relation $\delta = 2\sqrt{\Delta}$, we obtain that the error level in the function value should satisfy

$$\Delta \leq \frac{\varepsilon^2}{144p^2P_0^2}.$$

The parameter τ should satisfy

$$\tau \leq \frac{\varepsilon}{6pP_0\sqrt{L_0}}.$$

At the same time, to obtain an ε-solution for Problem (4), it is enough to choose

$$k = \max\left\{\left\lceil p\sqrt{\frac{6P_0^2}{\varepsilon}} + 1 - 2p\right\rceil, 0\right\}.$$

Conclusion

In this paper, we introduce a unifying framework, which allows to construct different types of accelerated randomized methods for smooth convex optimization problems and to prove convergence rate theorems for these methods. As we show, our framework is rather flexible and allows to reproduce known results as well as obtain new methods with convergence rate analysis. At the moment randomized methods for empirical risk minimization problems are not directly covered by our framework. It seems to be a n interesting direction for

further research. Another directions, in which we actually work, include generalization of our framework for strongly convex problems based on well-known restart technique. Another direction of our work is connected to non-uniform probabilities for sampling of coordinate blocks and composite optimization problems.

Acknowledgments. The authors are very grateful to Yu. Nesterov for fruitful discussions.

References

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 1200–1205, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4528-6. doi: 10.1145/3055399.3055448. URL http://doi.acm.org/10.1145/3055399.3055448. arXiv:1603.05953.

Zeyuan Allen-Zhu, Zheng Qu, Peter Richtarik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1110–1119, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/allen-zhuc16.html. First appeared in arXiv:1512.09103.

Aaron Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization (Lecture Notes)*. Personal web-page of A. Nemirovski, 2015. URL http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf.

Lev Bogolubsky, Pavel Dvurechensky, Alexander Gasnikov, Gleb Gusev, Yurii Nesterov, Andrei M Raigorodskii, Aleksey Tikhonov, and Maksim Zhukovskii. Learning supervised pagerank with gradient-based and gradient-free optimization methods. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4914–4922. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6565-learning-supervised-pagerank-with-gradient-based-and-gradient-free-optimization-pdf.

Cong D. Dang and Guanghui Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM J. on Optimization*, 25(2):856–881, April 2015. ISSN 1052-6234. doi: 10.1137/130936361. URL https://doi.org/10.1137/130936361.

Pavel Dvurechensky, Sergey Omelchenko, and Alexander Tiurin. Adaptive similar triangles method: a stable alternative to sinkhorn's algorithm for regularized optimal transport. *arXiv preprint arXiv:1706.07622*, 2017.

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015. First appeared in arXiv:1312.5799.

Roy Frostig, Rong Ge, Sham Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2540–2548, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/frostig15.html.

A. V. Gasnikov, A. A. Lagunovskaya, I. N. Usmanova, and F. A. Fedorenko. Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control*, 77(11):2018–2034, Nov 2016a. ISSN 1608-3032. doi: 10.1134/S0005117916110114. URL http://dx.doi.org/10.1134/S0005117916110114.

A. V. Gasnikov, E. A. Krymova, A. A. Lagunovskaya, I. N. Usmanova, and F. A. Fedorenko. Stochastic online optimization. single-point and multi-point non-linear multi-armed bandits. convex and strongly-convex case. *Automation and Remote Control*, 78(2):224–234, Feb 2017. ISSN 1608-3032. doi: 10.1134/S0005117917020035. URL http://dx.doi.org/10.1134/S0005117917020035.

Alexander Gasnikov, Pavel Dvurechensky, and Yurii Nesterov. Stochastic gradient methods with inexact oracle. *Proceedings of Moscow Institute of Physics and Technology*, 8(1): 41–91, 2016b. In Russian.

Alexander Gasnikov, Pavel Dvurechensky, and Ilnura Usmanova. On accelerated randomized methods. *Proceedings of Moscow Institute of Physics and Technology*, 8(2):67–100, 2016c. In Russian, first appeared in arXiv:1508.02182.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, Jun 2017. ISSN 1436-4646. doi: 10.1007/s10107-017-1173-0. URL http://dx.doi.org/10.1007/s10107-017-1173-0.

Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, FOCS '13, pages 147–156, Washington, DC, USA, 2013. IEEE Computer Society. ISBN 978-0-7695-5135-7. doi: 10.1109/FOCS.2013.24. URL http://dx.doi.org/10.1109/FOCS.2013.24. First appeared in arXiv:1305.1922.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 3384–3392, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969442.2969617.

Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated proximal coordinate gradient method. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and

K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3059–3067. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5356-an-accelerated-proximal-coordinate-gradient-method.pdf. First appeared in arXiv:1407.1296.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. doi: 10.1137/100802001. URL https://doi.org/10.1137/100802001. First appeared in 2010 as CORE discussion paper 2010/2.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, April 2017. ISSN 1615-3375. doi: 10.1007/s10208-015-9296-2. URL https://doi.org/10.1007/s10208-015-9296-2. First appeared in 2011 as CORE discussion paper 2011/16.

Yurii Nesterov and Sebastian U. Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017. doi: 10.1137/16M1060182. URL https://doi.org/10.1137/16M1060182. First presented in May 2015 http://www.mathnet.ru:8080/PresentFiles/11909/7_nesterov.pdf.

Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 64–72, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/shalev-shwartz14.html. First appeared in arXiv:1309.2375.

Rachael Tappenden, Peter Richtárik, and Jacek Gondzio. Inexact coordinate descent: Complexity and preconditioning. *Journal of Optimization Theory and Applications*, 170(1):144–176, Jul 2016. ISSN 1573-2878. doi: 10.1007/s10957-016-0867-4. URL http://dx.doi.org/10.1007/s10957-016-0867-4. First appeared in arXiv:1304.5530.

Alexander Tyurin. Mirror version of similar triangles method for constrained optimization problems. *arXiv preprint arXiv:1705.09809*, 2017.

Yuchen Zhang and Xiao Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 353–361, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/zhanga15.html.