European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

An accelerated directional derivative method for smooth stochastic convex optimization

Pavel Dvurechensky[a,b,c,*], Eduard Gorbunov[d,e], Alexander Gasnikov[f,g,h]

[a] Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany
[b] Institute for Information Transmission Problems RAS, Bolshoy Karetny per. 19, build.1, Moscow 127051, Russia
[c] National Research University Higher School of Economics, 11, Pokrovsky boulevard, Moscow 109028, Russia
[d] Moscow Institute of Physics and Technology, 9 Institutskiy per., 141700 Dolgoprudny, Moscow Region, Russia
[e] National Research University Higher School of Economics, 11, Pokrovsky boulevard, Moscow 109028, Russia
[f] Moscow Institute of Physics and Technology, 9 Institutskiy per., 141700 Dolgoprudny, Moscow Region, Russia
[g] National Research University Higher School of Economics, 11, Pokrovsky boulevard, 109028 Moscow, Russia
[h] Institute for Information Transmission Problems RAS Bolshoy Karetny per. 19, build.1, Moscow 127051, Russia

ARTICLE INFO

Article history:
Received 1 March 2018
Accepted 13 August 2020
Available online xxx

Keywords:
Stochastic programming
Convex programming
Acceleration
Derivative-free optimization
Zero-order methods

ABSTRACT

We consider smooth stochastic convex optimization problems in the context of algorithms which are based on directional derivatives of the objective function. This context can be considered as an intermediate one between derivative-free optimization and gradient-based optimization. We assume that at any given point and for any given direction, a stochastic approximation for the directional derivative of the objective function at this point and in this direction is available with some additive noise. The noise is assumed to be of an unknown nature, but bounded in the absolute value. We underline that we consider directional derivatives in *any* direction, as opposed to coordinate descent methods which use only derivatives in coordinate directions. For this setting, we propose a non-accelerated and an accelerated directional derivative method and provide their complexity bounds. Our non-accelerated algorithm has a complexity bound which is similar to the gradient-based algorithm, that is, without any dimension-dependent factor. Our accelerated algorithm has a complexity bound which coincides with the complexity bound of the accelerated gradient-based algorithm up to a factor of square root of the problem dimension. We extend these results to strongly convex problems.

1. Introduction

Zero-order or derivative-free optimization considers problems of minimization of a function using only, possibly noisy, observations of its values. This area of optimization has a long history, starting as early as in 1960 (Rosenbrock, 1960; Fabian, 1967), see also (Brent, 1973; Spall, 2003; Conn, Scheinberg, and Vicente, 2009). Even an older area of optimization, which started in 19th century Cauchy (1847), considers first-order methods which use the information about the gradient of the objective function. In this paper, we choose an intermediate class of problems. Namely, we assume that at any given point and for any given direction, a noisy stochastic approximation for the directional derivative of the objective function at this point in this direction is available. We underline that we consider directional derivatives in *any* direction, as opposed to coordinate descent methods which rely only on derivatives in coordinate directions. We refer to the class of optimization methods, which use directional derivatives of the objective function, as *directional derivative methods*. Unlike well developed areas of derivative-free and first-order stochastic optimization methods, the area of directional derivative optimization methods for stochastic optimization problems is not sufficiently covered in the literature. This class of optimization methods can be motivated by at least three situations.

The first one is connected to Automatic Differentiation Wengert (1964). Assume that the objective function is given as a computer program, which performs elementary arithmetic operations and elementary functions evaluations. Automatic Differentiation allows to calculate the gradient of this objective function and the additional computational cost is no more than five times larger than the cost of the evaluation of the objective value. The drawback of this approach is that it requires to store in memory the result of all the intermediate operations, which can require large memory

* Corresponding author at: Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany

E-mail addresses: pavel.dvurechensky@wias-berlin.de (P. Dvurechensky), eduard.gorbunov@phystech.edu (E. Gorbunov), gasnikov@yandex.ru (A. Gasnikov).

https://doi.org/10.1016/j.ejor.2020.08.027

amount. On the contrary, calculation of the directional derivative is easier than the calculation of the full gradient and requires the same memory amount as the calculation of the value of the objective Kim, Nesterov, Skokov, and Cherkasskii (1984). Since a random vector can be a part of the program input or some randomness can be used during the program execution, stochastic optimization problems can also be considered.

Importantly, automatic calculation of the directional derivative does not require the objective function to be smooth. This fact motivates the study of directional derivative methods in connection to Deep Learning. Indeed, learning problem is often stated as a problem of minimization of a loss function. A non-smooth activation function, called rectifier, is frequently used in Deep Learning as a building block for the loss function. Formally speaking, this non-smoothness does not allow to use Automatic Differentiation in the form of backpropagation to calculate the gradient of the objective function. At the same time, directional derivatives can be calculated by properly modified backpropagation.

The second motivating situation is connected to quasi-variational inequalities, which are used in modelling of different phenomena, such as sandpile formation and growth Prigozhin (1996), determination of lakes and river networks Barrett and Prigozhin (2014), and superconductivity Barrett and Prigozhin (2010). It happens that directional derivatives can be calculated for such problems Mordukhovich and Outrata (2007) as a solution to some auxiliary problem. Since this subproblem can not always be solved exactly, the noise in the directional derivative naturally arises. If the considered physical phenomenon takes place in some random media, stochastic optimization can be a natural approach to use.

The third motivating situation is connected to derivative-free stochastic optimization. In this situation a gradient approximation, based on the difference of stochastic approximations for the values of the objective in two close points, can be considered as a noisy directional derivative in the direction given by the difference of these two points Dvurechensky, Gasnikov, and Tiurin (2017). In this case, derivative-free stochastic optimization can be considered as a particular case of directional derivative stochastic optimization.

Motivated by potential presence of non-stochastic noise in the problem, we assume that the noise in the directional derivative consists of two parts. Similar to stochastic optimization problems, the first part is of a stochastic nature. On the opposite, the second part is an additive noise of an unknown nature, but bounded in the absolute value. More precisely, we consider the following optimization problem

$$\min_{x\in\mathbb{R}^n}\left\{f(x):=\mathbb{E}_\xi[F(x,\xi)]=\int_{\mathcal{X}}F(x,\xi)dP(x)\right\}, \tag{1}$$

where ξ is a random vector with probability distribution $P(\xi)$, $\xi\in\mathcal{X}$, and for P-almost every $\xi\in\mathcal{X}$, the function $F(x,\xi)$ is closed and convex. Moreover, we assume that, for P almost every ξ, the function $F(x,\xi)$ has gradient $g(x,\xi)$, which is $L(\xi)$-Lipschitz continuous with respect to the Euclidean norm and there exists $L_2\geq 0$ such that $\sqrt{\mathbb{E}_\xi L(\xi)^2}\leqslant L_2<+\infty$. Under this assumptions, $\mathbb{E}_\xi g(x,\xi)=\nabla f(x)$ and f has L_2-Lipschitz continuous gradient with respect to the Euclidean norm. Also we assume that

$$\mathbb{E}_\xi[\|g(x,\xi)-\nabla f(x)\|_2^2]\leqslant\sigma^2, \tag{2}$$

where $\|\cdot\|_2$ is the Euclidean norm.

Finally, we assume that an optimization procedure, given a point $x\in\mathbb{R}^n$, direction $e\in S_2(1)$ and ξ independently drawn from P, can obtain a noisy stochastic approximation $\widetilde{f}'(x,\xi,e)$ for the directional derivative $\langle g(x,\xi),e\rangle$:

$$\widetilde{f}'(x,\xi,e)=\langle g(x,\xi),e\rangle+\zeta(x,\xi,e)+\eta(x,\xi,e),$$
$$\mathbb{E}_\xi(\zeta(x,\xi,e))^2\leqslant\Delta_\zeta,\ \forall x\in\mathbb{R}^n,\forall e\in S_2(1),$$
$$|\eta(x,\xi,e)|\leqslant\Delta_\eta,\ \forall x\in\mathbb{R}^n,\forall e\in S_2(1),\ \text{a.s. in } \xi, \tag{3}$$

where $S_2(1)$ is the Euclidean sphere or radius one with the center at the point zero and the values Δ_ζ, Δ_η are controlled and can be made as small as it is desired. Note that we use the smoothness of $F(\cdot,\xi)$ to write the directional derivative as $\langle g(x,\xi),e\rangle$, but we *do not assume* that the whole stochastic gradient $g(x,\xi)$ is available.

It is well-known (Lan, 2012; Devolder, 2011; Dvurechensky and Gasnikov, 2016; Gasnikov and Dvurechensky, 2016) that, if the stochastic approximation $g(x,\xi)$ for the gradient of f is available, an accelerated gradient method has complexity bound $O\left(\max\left\{\sqrt{L_2/\varepsilon},\sigma^2/\varepsilon^2\right\}\right)$, where ε is the target optimization error. The question, to which we give a positive answer in this paper, is as follows.

Is it possible to solve a smooth stochastic optimization problem with the same ε-dependence in the complexity and only noisy observations of the directional derivative?

1.1. Related work

We first consider the related work on directional derivative optimization methods and, then, a closely related class of derivative-free methods with two-point feedback, the latter meaning that an optimization method uses two function value evaluations on each iteration. Since all the considered methods are randomized, we compare oracle complexity bounds in terms of expectation, that is, a number of directional derivatives or function values evaluations which is sufficient to achieve an error ε in the expected optimization error $\mathbb{E}f(\hat{x})-f^*$, where $\hat{x}$ is the output of an algorithm and f^* is the optimal value of f.

1.1.1. Directional derivative methods

Deterministic smooth optimization problems. In Nesterov and Spokoiny (2017), the authors consider the Euclidean case and propose a non-accelerated and an accelerated directional derivative method for smooth convex problems with complexity bounds $O(nL_2/\varepsilon)$ and $O(n\sqrt{L_2/\varepsilon})$ respectively. Also they propose a non-accelerated and an accelerated method for problems with μ-strongly convex objective and prove complexity bounds $O(nL_2/\mu\log_2(1/\varepsilon))$ and $O(n\sqrt{L_2/\mu}\log_2(1/\varepsilon))$ respectively. For a more general case of problems with additional bounded noise in directional derivatives, but also for the Euclidean case, an accelerated directional derivative method was proposed in Dvurechensky et al. (2017) and a bound $O(n\sqrt{L_2/\varepsilon})$ was proved.

We also should mention coordinate descent methods. In the seminal paper Nesterov (2012), a random coordinate descent for smooth convex and μ-strongly convex optimization problems were proposed and $O(L/\varepsilon)$ and $O(L/\mu\log_2(1/\varepsilon))$ complexity bounds were proved, where L is an effective Lipschitz constant of the gradient varying from n to some average over coordinates coordinate-wise Lipschitz constant. In the same paper, an accelerated version of random coordinate descent was proposed for convex problems and $O(n\sqrt{L/\varepsilon})$ complexity bound was proved. Papers (Fercoq & Richtárik, 2015; Lee & Sidford, 2013; Lin, Lu, & Xiao, 2014; Shalev-Shwartz & Zhang, 2014) generalize accelerated random coordinate descent for different settings, including μ-strongly convex problems, and (Allen-Zhu, Qu, Richtarik, & Yuan, 2016; Gasnikov, Dvurechensky, & Usmanova, 2016a; Nesterov & Stich, 2017) provide a $O(\sqrt{L/\varepsilon})$ and $O(\sqrt{L/\mu}\log_2(1/\varepsilon))$ complexity bounds, where L is an effective Lipschitz constant of the gradient varying from n to some average over coordinates coordinate-wise Lipschitz constant, and, in the best case, is dimension-independent. An accelerated random coordinate descent with inexact coordinate-wise derivatives was proposed in Dvurechensky et al. (2017) with $O(n\sqrt{L/\varepsilon})$ complexity bound and also a unified view on directional derivative methods, coordinate descent and derivative-free methods.

Stochastic optimization problems. A directional derivative method for non-smooth stochastic convex optimization problems was introduced in Nesterov and Spokoiny (2017) with a complexity bound $O(n^2/\varepsilon^2)$. A random coordinate descent method for non-smooth stochastic convex and μ- strongly convex optimization problems were introduced in Dang and Lan (2015) with complexity bounds $O(n/\varepsilon^2)$ and $O(n/\mu\varepsilon)$ respectively.

1.1.2. Derivative-free methods

Deterministic smooth optimization problems. A non-accelerated and an accelerated derivative-free method for this type of problems were proposed in Nesterov and Spokoiny (2017) for the Euclidean case with the bounds $O(nL_2/\varepsilon)$ and $O(n\sqrt{L_2/\varepsilon})$ respectively. The same paper proposed a non-accelerated and an accelerated method for μ-strongly convex problems with complexity bounds $O(nL_2/\mu \log_2(1/\varepsilon))$ and $O(n\sqrt{L_2/\mu}\log_2(1/\varepsilon))$ respectively. A non-accelerated derivative-free method for deterministic problems with additional bounded noise in function values was proposed in Bogolubsky et al. (2016) together with $O(nL_2/\varepsilon)$ bound and application to learning parameter of a parametric PageRank model, see also (Gasnikov, Gasnikova, Dvurechensky, Mohammed, and Chernousova, 2017a; Gasnikov et al., 2018). Deterministic problems with additional bounded noise in function values were also considered in Dvurechensky et al. (2017), where several accelerated derivative-free methods, including Derivative-Free Block-Coordinate Descent, were proposed and a bound $O(n\sqrt{L/\varepsilon})$ was proved, where L depends on the method and, in some sense, characterizes the average over blocks of coordinates Lipschitz constant of the derivative in the block. Mixed first-order/zero-order setting is considered in Beznosikov, Gorbunov, and Gasnikov (2020a). After our paper appeared as a preprint, the papers Berahas, Byrd, and Nocedal (2019a); Bollapragada and Wild (2019) studied derivative-free quasi-Newton methods for problems with noisy function values, and the paper Berahas, Cao, Choromanski, and Scheinberg (2019b) reported theoretical and empirical comparison of different gradient approximations for zero-order methods.

Stochastic optimization problems. Most of the authors in this group solve a more general problem of bandit convex optimization and obtain bounds on the so-called regret. It is well known Cesa-bianchi, Conconi, and Gentile (2002) that a bound on the regret can be converted to a bound on the expected optimization error. Non-smooth stochastic optimization problems were considered in Nesterov and Spokoiny (2017), where an $O(n^2/\varepsilon^2)$ complexity bound was proved for a derivative-free method. This bound was improved by (Duchi, Jordan, Wainwright, and Wibisono, 2015; Gasnikov, Lagunovskaya, Usmanova, and Fedorenko, 2016b; Gasnikov, Krymova, Lagunovskaya, Usmanova, and Fedorenko, 2017b; Shamir, 2017; Bayandina, Gasnikov, and Lagunovskaya, 2018b; Hu, L.A., Gyrgy, and Szepesvari, 2016) to[1] $\tilde{O}(n^{2/q}R_p^2/\varepsilon^2)$, where $p \in \{1, 2\}$, $\frac{1}{p}+\frac{1}{q}=1$ and R_p is the radius of the feasible set in the p-norm $\|\cdot\|_p$. For non-smooth μ_p-strongly convex w.r.t. to p-norm problems, the authors of Bayandina et al. (2018b); Gasnikov et al. (2017b) proved a bound $\tilde{O}(n^{2/q}/(\mu_p\varepsilon))$. A version of these methods for non-smooth saddle-point problems is developed in Beznosikov, Sadiev, and Gasnikov (2020b).

Intermediate, partially smooth problems with a restrictive assumption of boundedness of $\mathbb{E}\ \|g(x,\xi)\|^2$, were considered in Duchi et al. (2015), where it was proved that a proper modification of Mirror Descent algorithm with derivative-free approximation of the gradient gives a bound $O(n^{2/q}R_p^2/\varepsilon^2)$ for convex problems, improving upon the bound $\tilde{O}(n^2/\varepsilon^2)$ of Agarwal, Dekel, and Xiao (2010). For strongly convex w.r.t 2-norm problems, the authors of Agarwal et al. (2010) obtained a bound $\tilde{O}(n^2/\varepsilon)$, which was later extended for μ_p-strongly convex problems and improved to $\tilde{O}(n^{2/q}/(\mu_p\varepsilon))$ in Gasnikov et al. (2017b).

In the fully smooth case, without the assumption that $\mathbb{E}\|g(x,\xi)\|^2 < +\infty$, papers Ghadimi and Lan (2013); Ghadimi, Lan, and Zhang (2016) proposed a derivative-free algorithm for the Euclidean case with the bound

$$\tilde{O}\left(\max\left\{\frac{nL_2R_2}{\varepsilon}, \frac{n\sigma^2}{\varepsilon^2}\right\}\right).$$

In Gorbunov, Dvurechensky, and Gasnikov (2018), the authors proposed a non-accelerated and an accelerated derivative-free method with the bounds

$$\tilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}}L_2R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2R_p^2}{\varepsilon^2}\right\}\right), \quad \tilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2R_p^2}{\varepsilon}}, \frac{n^{\frac{2}{q}}\sigma^2R_p^2}{\varepsilon^2}\right\}\right)$$

respectively, where R_p characterizes the distance in p-norm between the starting point of the algorithm and a solution to (1), $p \in \{1, 2\}$ and $q \in \{2, \infty\}$ is the conjugate to p, given by the identity $\frac{1}{p}+\frac{1}{q}=1$.

The authors of Chen, Orvieto, and Lucchi (2020) combine accelerated derivative-free optimization with accelerated variance reduction technique for finite-sum convex problems in the Euclidean setup.

Other works. For a recent review of derivative-free optimization see Larson, Menickelly, and Wild (2019) and for a review of stochastic optimization, including derivative-free optimization, see Powell (2019).

1.2. Our contributions

As we have seen above, only two results on directional derivative methods for non-smooth stochastic convex optimization are available in the literature, and, to the best of our knowledge, nothing is known about directional derivative methods for smooth stochastic convex optimization, even in the well-developed area of random coordinate descent methods. Our main contribution consists in closing this gap in the theory of directional derivative methods for stochastic optimization and considering even more general setting with additional noise of an unknown nature in the directional derivative.

Our methods are based on two proximal setups Ben-Tal and Nemirovski (2015) characterized by the value[2] $p \in \{1, 2\}$ and its conjugate $q \in \{2, \infty\}$, given by the identity $\frac{1}{p}+\frac{1}{q}=1$. The case $p=1$ corresponds to the choice of 1-norm in $\mathbb{R}^n$ and corresponding prox-function which is strongly convex with respect to this norm (we provide the details below). The case $p=2$ corresponds to the choice of the Euclidean 2-norm in $\mathbb{R}^n$ and squared Euclidean norm as the prox-function. As our main contribution, we propose an Accelerated Randomized Directional Derivative (ARDD) algorithm for smooth stochastic optimization based on noisy observations of directional derivative of the objective. Our method has the complexity bound

$$\tilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2R_p^2}{\varepsilon}}, \frac{n^{\frac{2}{q}}\sigma^2R_p^2}{\varepsilon^2}\right\}\right), \quad (4)$$

where R_p characterizes the distance in p-norm between the starting point of the algorithm and a solution to (1).

As our second contribution, we propose a non-accelerated Randomized Directional Derivative (RDD) algorithm with the complex-

[1] $\tilde{O}$ hides polylogarithmic factors $(\ln n)^c$, $c > 0$.

[2] Strictly speaking, we are able to consider all the intermediate cases $p \in [1, 2]$, but we are not aware of any proximal setup which is compatible with $p \notin \{1, 2\}$

Please cite this article as: P. Dvurechensky, E. Gorbunov and A. Gasnikov, An accelerated directional derivative method for smooth stochastic convex optimization, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2020.08.027

ity bound

$$\widetilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}}L_2R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2R_p^2}{\varepsilon^2}\right\}\right). \quad (5)$$

Interestingly, for this method when $p=1$ and $q=\infty$, we obtain complexity bound which depends on the dimension n only logarithmically despite we use only noisy directional derivative observations. Let us comment on the comparison between the accelerated and non-accelerated method. In the regime of small variance σ^2 in both bounds the dominating term is the first one. If $p=1$, $q=\infty$ and $L_2R_p^2 < n\varepsilon$, then the bound for the non-accelerated method is smaller than that of for the accelerated. In this regime it is preferred to use the non-accelerated method.

Note that, in the case of (1) having a sparse solution, our bounds for $p=1$ allow to gain a factor of $\sqrt{n}$ in the complexity of the accelerated method and a factor of n in the complexity of the non-accelerated method in comparison to the Euclidean case $p=2$. Indeed, sparsity of a solution x^* means that $\|x^*\|_1 = O(1)\cdot\|x^*\|_2$ and, if the starting point is zero, we obtain $R_1^2 = \|x^*\|_1^2 = O(1)\cdot\|x^*\|_2^2 = O(1)R_2^2$. Hence, the bounds for $p=1$ and $p=2$ can be compared only based on the corresponding powers of n, the latter being smaller for the case $p=1$, $q=\infty$.

We underline here that our methods are based on random directions drawn from the uniform distribution on the unit Euclidean sphere and our results for $p=1$ can not be obtained by random coordinate descent.

As our third contribution, we extend the above results to the case when the objective function is additionally known to be μ_p-strongly convex w.r.t. p-norm. For this case, we propose an accelerated and a non-accelerated algorithm which respectively have complexity bounds

$$\widetilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2}{\mu_p}}\log_2\frac{\mu_pR_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2}{\mu_p\varepsilon}\right\}\right),$$
$$\widetilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}}L_2}{\mu_p}\log_2\frac{\mu_pR_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2}{\mu_p\varepsilon}\right\}\right). \quad (6)$$

In the regime of small variance σ^2 in both bounds the dominating term is the first one. If $p=1$, $q=\infty$ and $\frac{L_2}{\mu_p} < n$, then the bound for the non-accelerated method is smaller than that of for the accelerated. In this regime of relatively well-conditioned problems it is preferred to use the non-accelerated method.

As our final contribution, we consider derivative-free smooth stochastic convex optimization with inexact values of the stochastic approximations for the function values as a particular case of optimization using noisy directional derivatives. This allows us to obtain the complexity bounds of Gorbunov et al. (2018) as a straightforward corollary of our results in this paper. At the same time we obtain new complexity bounds for the strongly convex case which, to the best of our knowledge, were not known in the literature.

Note that our results for accelerated and non-accelerated methods are somewhat similar to the finite-sum minimization problems of the form

$$\min_{x\in\mathbb{R}^n}\sum_{i=1}^m f_i(x),$$

where f_i are convex smooth functions. For such problems accelerated methods have complexity $\widetilde{O}(m+\sqrt{mL/\varepsilon})$ and non-accelerated methods have complexity $\widetilde{O}(m+L/\varepsilon)$ (see, e.g. Allen-Zhu, 2017 for a nice review on the topic). As we see, acceleration allows to take the square root of the second term but for the price of $\sqrt{m}$ and the two bounds can not be directly compared without additional assumptions on the value of $m\varepsilon$.

Special note on (Gorbunov et al., 2018; Vorontsova, Gasnikov, Gorbunov, and Dvurechenskii, 2019). One of the novelties and insights in the approach of this paper in comparison to (Gorbunov et al., 2018; Vorontsova et al., 2019) is to realize that gradient-free methods are a particular case of directional derivative methods with inexact oracle. Unlike these papers, in the current paper we need to account for two types of inexactness. One is stochastic with bounded second moment and the second is bounded a.s. This is a more complicated assumption than the one in (Gorbunov et al., 2018; Vorontsova et al., 2019) and we have to assume that the error values can be controlled, unlike (Gorbunov et al., 2018; Vorontsova et al., 2019). Moreover, since the oracle returns different information, we have to construct our stochastic approximation of the gradient differently, which also changes the proof technique. We also analyze in this paper the case of strongly convex objective values, which was not done in (Gorbunov et al., 2018; Vorontsova et al., 2019).

1.3. Paper organization

The rest of the paper is organized as follows. In Section 2, both for convex and strongly convex problems, we introduce our algorithms, state their convergence rate theorems and corresponding complexity bounds. Section 3 is devoted to proof of the convergence rate theorem for our accelerated method and convex objective functions. Section 4 is devoted to proof of the convergence rate theorem for our non-accelerated method and convex objective functions. In Section 5 we provide the proofs for the case of strongly convex objective function. Finally, in Section 6 we provide numerical experiments with two types of objective functions: worst case functions for first-order methods Nesterov (2004) and least squares problem.

2. Algorithms and main results

In this section, we provide our non-accelerated and accelerated directional derivative methods both for convex and strongly convex problems together with convergence theorems and corresponding complexity bounds. The proofs are rather technical and postponed to next sections.

2.1. Preliminaries

We start by introducing necessary objects and technical results.

Proximal setup. Let $p\in[1,2]$ and $\|x\|_p$ be the p-norm in $\mathbb{R}^n$ defined as

$$\|x\|_p^p = \sum_{i=1}^n |x_i|^p, \quad x\in\mathbb{R}^n,$$

$\|\cdot\|_q$ be its dual, defined by $\|g\|_q = \max_x\left\{\langle g,x\rangle, \|x\|_p\le 1\right\}$, where $q\in[2,\infty]$ is the conjugate number to p, given by $\frac{1}{p}+\frac{1}{q}=1$, and, for $q=\infty$, by definition $\|x\|_\infty = \max_{i=1,\dots,n}|x_i|$.

We choose a *prox-function* $d(x)$ which is continuous, convex on $\mathbb{R}^n$ and is 1-strongly convex on $\mathbb{R}^n$ with respect to $\|\cdot\|_p$, i.e., for any $x,y\in\mathbb{R}^n$ $d(y)-d(x)-\langle\nabla d(x),y-x\rangle \ge \frac{1}{2}\|y-x\|_p^2$. Without loss of generality, we assume that $\min_{x\in\mathbb{R}^n} d(x)=0$. We define also the corresponding *Bregman divergence* $V[z](x) = d(x)-d(z)-\langle\nabla d(z),x-z\rangle$, $x,z\in\mathbb{R}^n$. Note that, by the strong convexity of d,

$$V[z](x) \ge \frac{1}{2}\|x-z\|_p^2, \quad x,z\in\mathbb{R}^n. \quad (7)$$

For the case $p=1$, we choose the following prox-function Ben-Tal and Nemirovski (2015)

$$d(x) = \frac{en^{(\kappa-1)(2-\kappa)/\kappa}\ln n}{2}\|x\|_\kappa^2, \quad \kappa = 1+\frac{1}{\ln n} \quad (8)$$

and, for the case $p=2$, we choose the prox-function to be the squared Euclidean norm

$$d(x) = \frac{1}{2}\|x\|_2^2. \tag{9}$$

Main technical lemma. In our proofs of complexity bounds, we rely on the following lemma. The proof is rather technical and is provided in the appendix.

Lemma 1. *Let $e \in RS_2(1)$, i.e be a random vector uniformly distributed on the surface of the unit Euclidean sphere in $\mathbb{R}^n$, $p \in [1, 2]$ and q be given by $\frac{1}{p}+\frac{1}{q}=1$. Then, for $n \geq 8$ and $\rho_n = \min\{q-1, 16\ln n - 8\}n^{\frac{2}{q}-1}$,*

$$\mathbb{E}_e\|e\|_q^2 \leq \rho_n, \tag{10}$$

$$\mathbb{E}_e\left(\langle s, e\rangle^2\|e\|_q^2\right) \leq \frac{6\rho_n}{n}\|s\|_2^2, \quad \forall s \in \mathbb{R}^n. \tag{11}$$

Stochastic approximation of the gradient. Based on the noisy stochastic observations (3) of the directional derivative, we form the following stochastic approximation of $\nabla f(x)$

$$\widetilde{\nabla}^m f(x) = \frac{1}{m}\sum_{i=1}^m \widetilde{f}'(x, \xi_i, e)e, \tag{12}$$

where $e \in RS_2(1)$, ξ_i, $i=1,\dots,m$ are independent realizations of ξ, m is the *batch size*.

2.2. Algorithms and main results for convex problems

Our Accelerated Randomized Directional Derivative (ARDD) method is listed as Algorithm 1.

Algorithm 1 Accelerated Randomized Directional Derivative (ARDD) method.

Input: x_0—starting point; $N \geqslant 1$— number of iterations; $m \geqslant 1$— batch size.
Output: point y_N.
1: $y_0 \leftarrow x_0$, $z_0 \leftarrow x_0$.
2: **for** $k=0,\dots,N-1$. **do**
3: $\alpha_{k+1} \leftarrow \frac{k+2}{96n^2\rho_n L_2}$, $\tau_k \leftarrow \frac{1}{48\alpha_{k+1}n^2\rho_n L_2} = \frac{2}{k+2}$.
4: Generate $e_{k+1} \in RS_2(1)$ independently from previous iterations and ξ_i, $i=1,\dots,m$ – independent realizations of ξ.
5: Calculate
$$\widetilde{\nabla}^m f(x_{k+1}) = \frac{1}{m}\sum_{i=1}^m \widetilde{f}'(x_{k+1}, \xi_i, e)e.$$
6: $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k)y_k$.
7: $y_{k+1} \leftarrow x_{k+1} - \frac{1}{2L_2}\widetilde{\nabla}^m f(x_{k+1})$.
8: $z_{k+1} \leftarrow \operatorname*{argmin}_{z\in\mathbb{R}^n}\left\{\alpha_{k+1}n\left\langle\widetilde{\nabla}^m f(x_{k+1}), z-z_k\right\rangle + V[z_k](z)\right\}$.
9: **end for**
10: **return** y_N

Theorem 1. *Let ARDD method be applied to solve problem (1). Then*

$$\mathbb{E}[f(y_N)] - f(x^*) \leqslant \frac{384\Theta_p n^2\rho_n L_2}{N^2} + \frac{4N}{nL_2}\cdot\frac{\sigma^2}{m} + \frac{61N}{24L_2}\Delta_\zeta + \frac{122N}{3L_2}\Delta_\eta^2 + \frac{12\sqrt{2n\Theta_p}}{N^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2, \tag{13}$$

where $\Theta_p = V[z_0](x^*)$ is defined by the chosen proximal setup and $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\dots,e_N,\xi_{1,1},\dots,\xi_{N,m}}[\cdot]$.

Before we proceed to the non-accelerated method, we give the appropriate choice of the ARDD method parameters N, m, and accuracy of the directional derivative evaluation Δ_ζ, Δ_η. These values are chosen such that the r.h.s. of (13) is smaller than ε. For simplicity we omit numerical constants and summarize the obtained values of the algorithm parameters in Table 1 below. The last row represents the total number Nm of oracle calls, that is, the number of directional derivative evaluations, which was advertised in (4). Note that the bound (13) allows also to choose the accuracy of the directional derivative evaluation Δ_ζ, Δ_η decreasing with N. This is done by making each term with Δ_ζ or Δ_η in the r.h.s. to be of the same order as the first term.

Our Randomized Directional Derivative (RDD) method is listed as Algorithm 2.

Algorithm 2 Randomized Directional Derivative (RDD) method.

Input: x_0—starting point; $N \geqslant 1$— number of iterations; $m \geqslant 1$— batch size.
Output: point $\bar{x}_N$.
1: **for** $k=0,\dots,N-1$. **do**
2: $\alpha \leftarrow \frac{1}{48n\rho_n L_2}$.
3: Generate $e_{k+1} \in RS_2(1)$ independently from previous iterations and ξ_i, $i=1,\dots,m$ – independent realizations of ξ.
4: Calculate
$$\widetilde{\nabla}^m f(x_k) = \frac{1}{m}\sum_{i=1}^m \widetilde{f}'(x_k, \xi_i, e)e.$$
5: $x_{k+1} \leftarrow \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\alpha n\left\langle\widetilde{\nabla}^m f(x_k), x-x_k\right\rangle + V[x_k](x)\right\}$.
6: **end for**
7: **return** $\bar{x}_N \leftarrow \frac{1}{N}\sum_{k=0}^{N-1} x_k$

Theorem 2. *Let RDD method be applied to solve problem (1). Then*

$$\mathbb{E}[f(\bar{x}_N)] - f(x_*) \leqslant \frac{384n\rho_n L_2\Theta_p}{N} + \frac{2}{L_2}\frac{\sigma^2}{m} + \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 + \frac{8\sqrt{2n\Theta_p}}{N}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2, \tag{14}$$

where $\Theta_p = V[z_0](x^*)$ is defined by the chosen proximal setup and $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\dots,e_N,\xi_{1,1},\dots,\xi_{N,m}}[\cdot]$.

Before we proceed, we give the appropriate choice of the RDD method parameters N, m, and accuracy of the directional derivative evaluation Δ_ζ, Δ_η. These values are chosen such that the r.h.s. of (14) is smaller than ε. For simplicity we omit numerical constants and summarize the obtained values of the algorithm parameters in Table 2 below. The last row represents the total number Nm of oracle calls, that is, the number of directional derivative evaluations, which was advertised in (5). Note that the bound (14) allows also to choose the accuracy of the directional derivative evaluation Δ_ζ, Δ_η decreasing with N. This is done by making each term with Δ_ζ or Δ_η in the r.h.s. to be of the same order as the first term.

2.3. Extensions for strongly convex problems

In this section, we assume additionally that f is μ_p-strongly convex w.r.t. p-norm. Our algorithms and proofs rely on the following fact. Let x_* be some fixed point and x be a random point such that $\mathbb{E}_x\left[\|x-x_*\|_p^2\right] \leqslant R_p^2$, then

$$\mathbb{E}_x d\left(\frac{x-x_*}{R_p}\right) \leqslant \frac{\Omega_p}{2}, \tag{15}$$

Table 1
Algorithm 1 parameters for the cases $p=1$ and $p=2$.

	$p=1$	$p=2$
N	$O\left(\sqrt{\frac{n\ln nL_2\Theta_1}{\varepsilon}}\right)$	$O\left(\sqrt{\frac{n^2L_2\Theta_2}{\varepsilon}}\right)$
m	$O\left(\max\left\{1, \sqrt{\frac{\ln n}{n}}\cdot\frac{\sigma^2}{\varepsilon^{3/2}}\cdot\sqrt{\frac{\Theta_1}{L_2}}\right\}\right)$	$O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon^{3/2}}\cdot\sqrt{\frac{\Theta_2}{L_2}}\right\}\right)$
Δ_ζ	$O\left(\min\left\{n(\ln n)^2L_2^2\Theta_1, \frac{\varepsilon^2}{n\Theta_1}, \frac{\varepsilon^{\frac{3}{2}}}{\sqrt{n\ln n}}\cdot\sqrt{\frac{L_2}{\Theta_1}}\right\}\right)$	$O\left(\min\left\{n^3L_2^2\Theta_2, \frac{\varepsilon^2}{n\Theta_2}, \frac{\varepsilon^{\frac{3}{2}}}{n}\cdot\sqrt{\frac{L_2}{\Theta_2}}\right\}\right)$
Δ_η	$O\left(\min\left\{\sqrt{n}\ln nL_2\sqrt{\Theta_1}, \frac{\varepsilon}{\sqrt{n\Theta_1}}, \frac{\varepsilon^{\frac{3}{4}}}{\sqrt[4]{n\ln n}}\cdot\sqrt[4]{\frac{L_2}{\Theta_1}}\right\}\right)$	$O\left(\min\left\{n^{\frac{3}{2}}L_2\sqrt{\Theta_2}, \frac{\varepsilon}{\sqrt{n\Theta_2}}, \frac{\varepsilon^{\frac{3}{4}}}{\sqrt{n}}\cdot\sqrt[4]{\frac{L_2}{\Theta_2}}\right\}\right)$
O-le calls	$O\left(\max\left\{\sqrt{\frac{n\ln nL_2\Theta_1}{\varepsilon}}, \frac{\sigma^2\Theta_1\ln n}{\varepsilon^2}\right\}\right)$	$O\left(\max\left\{\sqrt{\frac{n^2L_2\Theta_2}{\varepsilon}}, \frac{\sigma^2\Theta_2 n}{\varepsilon^2}\right\}\right)$

Table 2
Algorithm 2 parameters for the cases $p=1$ and $p=2$.

	$p=1$	$p=2$
N	$O\left(\frac{L_2\Theta_1\ln n}{\varepsilon}\right)$	$O\left(\frac{nL_2\Theta_2}{\varepsilon}\right)$
m	$O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}\right)$	$O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}\right)$
Δ_ζ	$O\left(\min\left\{\frac{(\ln n)^2}{n}L_2^2\Theta_1, \frac{\varepsilon^2}{n\Theta_1}, \frac{\varepsilon L_2}{n}\right\}\right)$	$O\left(\min\left\{nL_2^2\Theta_2, \frac{\varepsilon^2}{n\Theta_2}, \frac{\varepsilon L_2}{n}\right\}\right)$
Δ_η	$O\left(\min\left\{\frac{\ln n}{\sqrt{n}}L_2\sqrt{\Theta_1}, \frac{\varepsilon}{\sqrt{n\Theta_1}}, \sqrt{\frac{\varepsilon L_2}{n}}\right\}\right)$	$O\left(\min\left\{\sqrt{n}L_2\sqrt{\Theta_2}, \frac{\varepsilon}{\sqrt{n\Theta_2}}, \sqrt{\frac{\varepsilon L_2}{n}}\right\}\right)$
O-le calls	$O\left(\max\left\{\frac{L_2\Theta_1\ln n}{\varepsilon}, \frac{\sigma^2\Theta_1\ln n}{\varepsilon^2}\right\}\right)$	$O\left(\max\left\{\frac{nL_2\Theta_2}{\varepsilon}, \frac{n\sigma^2\Theta_2}{\varepsilon^2}\right\}\right)$

where $\mathbb{E}_x$ denotes the expectation with respect to random vector x and Ω_p is defined as follows. For $p=1$ and our choice of the prox-function (8), $\Omega_p = en^{(\kappa-1)(2-\kappa)/\kappa}\ln n = O(\ln n)$ for our choice of $\kappa = 1+\frac{1}{\ln n}$, see Nemirovsky and Yudin (1983),Juditsky and Nesterov (2014). For $p=2$ and our choice of the prox-function (9), $\Omega_p = 1$. Our Accelerated Randomized Directional Derivative method for strongly convex problems (ARDDsc) is listed as Algorithm 3.

Algorithm 3 Accelerated Randomized Directional Derivative method for strongly convex functions (ARDDsc).

Input: x_0—starting point s.t. $\|x_0 - x_*\|_p^2 \le R_p^2$; $K \geqslant 1$— number of iterations; μ_p – strong convexity parameter.
Output: point u_K.

Set
$$N_0 = \left\lceil \sqrt{\frac{8aL_2\Omega_p}{\mu_p}} \right\rceil, \tag{16}$$
where $a = 384n^2\rho_n$.

for $k = 0, \dots, K-1$ **do**

Set
$$m_k := \max\left\{1, \left\lceil \frac{8b\sigma^2 N_0 2^k}{L_2\mu_p R_p^2} \right\rceil\right\},\ R_k^2 := R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1-2^{-k}\right), \tag{17}$$
where $b = \frac{4}{n}$.

Set $d_k(x) = R_k^2 d\left(\frac{x-u_k}{R_k}\right)$.

Run ARDD with starting point u_k and prox-function $d_k(x)$ for N_0 steps with batch size m_k.

Set $u_{k+1} = y_{N_0}$, $k = k+1$.

end for

return u_K

Theorem 3. *Let f in problem (1) be μ_p-strongly convex and ARDDsc method be applied to solve this problem. Then*

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2}\cdot 2^{-K} + 2\Delta. \tag{18}$$

where $\Delta = \frac{61N_0}{24L_2}\Delta_\zeta + \frac{122N_0}{3L_2}\Delta_\eta^2 + \frac{12\sqrt{2nR_p^2\Omega_p}}{N_0^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2$. *Moreover, under an appropriate choice of Δ_ζ and Δ_η s.t. $2\Delta \le \varepsilon/2$, the oracle complexity to achieve ε-accuracy of the solution is*

$$\tilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2\Omega_p}{\mu_p}}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right).$$

Despite we have linear convergence in terms of the iterations number, the number of the oracle evaluations corresponds to sublinear convergence. The reason is that we consider general stochastic optimization problem, rather than finite-sum problems for which the linear convergence rate is achievable in terms of the oracle evaluations Allen-Zhu (2017). Our oracle complexity corresponds to the lower complexity bounds Nemirovsky and Yudin (1983) for general stochastic convex optimization.

Before we proceed to the non-accelerated method, we give the appropriate choice of the accuracy of the directional derivative evaluation Δ_ζ, Δ_η for ARDDsc to achieve an accuracy ε of the solution. These values are chosen such that the r.h.s. of (18) is smaller than ε. For simplicity we omit numerical constants and summarize the obtained values of the algorithm parameters in Table 3 below. The last row represents the total number of oracle calls, that is, the number of directional derivative evaluations, which was stated in (6).

Our Randomized Directional Derivative method for strongly convex problems (RDDsc) is listed as Algorithm 4.

Theorem 4. *Let f in problem (1) be μ_p-strongly convex and RDDsc method be applied to solve this problem. Then*

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2}\cdot 2^{-K} + 2\Delta. \tag{21}$$

where $\Delta = \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 + \frac{8\sqrt{2nR_p^2\Omega_p}}{N_0}\left(\frac{\sqrt{\Delta_\zeta}}{2}+2\Delta_\eta\right)+\frac{N_0}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2}+2\Delta_\eta\right)^2$. *Moreover, under an appropriate choice of Δ_ζ and Δ_η s.t. $2\Delta \le \varepsilon/2$, the oracle complexity to achieve ε-accuracy of the solution*

Table 3
Algorithm 3 parameters for the cases $p=1$ and $p=2$.

	$p=1$	$p=2$
Δ_ζ	$O\left(\min\left\{\varepsilon\sqrt{\frac{L_2\mu_1}{n\ln n\Omega_1}},\ \varepsilon^2\frac{n(\ln n)^2L_2^2\Omega_1}{R_1^2\mu_1^2},\ \varepsilon\cdot\frac{\mu_1}{n\Omega_1}\right\}\right)$	$O\left(\min\left\{\varepsilon\sqrt{\frac{L_2\mu_2}{n^2\Omega_2}},\ \varepsilon^2\frac{n^3L_2^2\Omega_2}{R_2^2\mu_2^2},\ \varepsilon\cdot\frac{\mu_2}{n\Omega_2}\right\}\right)$
Δ_η	$O\left(\min\left\{\sqrt{\varepsilon}\sqrt[4]{\frac{L_2\mu_1}{n\ln n\Omega_1}},\ \varepsilon\frac{\sqrt{n}\ln nL_2\sqrt{\Omega_1}}{R_1\mu_1},\ \sqrt{\varepsilon}\cdot\sqrt{\frac{\mu_1}{n\Omega_1}}\right\}\right)$	$O\left(\min\left\{\sqrt{\varepsilon}\sqrt[4]{\frac{L_2\mu_2}{n^2\Omega_2}},\ \varepsilon\frac{\sqrt{n^3}L_2\sqrt{\Omega_2}}{R_2\mu_2},\ \sqrt{\varepsilon}\cdot\sqrt{\frac{\mu_2}{n\Omega_2}}\right\}\right)$
O-le calls	$\widetilde{O}\left(\max\left\{\sqrt{\frac{n\ln nL_2\Omega_1}{\mu_1}}\log_2\frac{\mu_1R_1^2}{\varepsilon},\ \frac{\sigma^2\Omega_1\ln n}{\mu_1\varepsilon}\right\}\right)$	$\widetilde{O}\left(\max\left\{n\sqrt{\frac{L_2\Omega_2}{\mu_2}}\log_2\frac{\mu_2R_2^2}{\varepsilon},\ \frac{n\sigma^2\Omega_2}{\mu_2\varepsilon}\right\}\right)$

Table 4
Algorithm 4 parameters for the cases $p=1$ and $p=2$.

	$p=1$	$p=2$
Δ_ζ	$O\left(\min\left\{\frac{\varepsilon L_2}{n},\ \varepsilon^2\frac{(\ln n)^2L_2^2}{nR_1^2\mu_1^2},\ \varepsilon\frac{\mu_1}{n\Omega_1}\right\}\right)$	$O\left(\min\left\{\frac{\varepsilon L_2}{n},\ \varepsilon^2\frac{nL_2^2}{R_2^2\mu_2^2},\ \varepsilon\frac{\mu_2}{n\Omega_2}\right\}\right)$
Δ_η	$O\left(\min\left\{\sqrt{\frac{\varepsilon L_2}{n}},\ \varepsilon\frac{\ln nL_2}{\sqrt{n}R_1\mu_1},\ \sqrt{\varepsilon\frac{\mu_1}{n\Omega_1}}\right\}\right)$	$O\left(\min\left\{\sqrt{\frac{\varepsilon L_2}{n}},\ \varepsilon\frac{\sqrt{n}L_2}{R_2\mu_2},\ \sqrt{\varepsilon\frac{\mu_2}{n\Omega_2}}\right\}\right)$
O-le calls	$\widetilde{O}\left(\max\left\{\frac{L_2\Omega_1\ln n}{\mu_1}\log_2\frac{\mu_1R_1^2}{\varepsilon},\ \frac{\sigma^2\Omega_1}{\mu_1\varepsilon}\right\}\right)$	$\widetilde{O}\left(\max\left\{\frac{nL_2\Omega_2}{\mu_2}\log_2\frac{\mu_2R_2^2}{\varepsilon},\ \frac{n\sigma^2\Omega_2}{\mu_2\varepsilon}\right\}\right)$

Algorithm 4 Randomized Directional Derivative method for strongly convex functions (RDDsc).

Input: x_0—starting point s.t. $\|x_0-x_*\|_p^2\le R_p^2$; $K\geqslant 1$— number of iterations; μ_p – strong convexity parameter.
Output: point u_K.
1: Set
$$N_0=\left\lceil\frac{8aL_2\Omega_p}{\mu_p}\right\rceil, \tag{19}$$
where $a=384n\rho_n$.
2: **for** $k=0,\dots,K-1$ **do**
3: Set
$$m_k:=\max\left\{1,\left\lceil\frac{8b\sigma^2 2^k}{L_2\mu_pR_p^2}\right\rceil\right\},\quad R_k^2:=R_p^2 2^{-k}+\frac{4\Delta}{\mu_p}\left(1-2^{-k}\right), \tag{20}$$
where $b=2$
4: Set $d_k(x)=R_k^2 d\left(\frac{x-u_k}{R_k}\right)$.
5: Run RDD with starting point u_k and prox-function $d_k(x)$ for N_0 steps with batch size m_k.
6: Set $u_{k+1}=y_{N_0}$, $k=k+1$.
7: **end for**
8: **return** u_K

is

$$\widetilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}}L_2\Omega_p}{\mu_p}\log_2\frac{\mu_pR_p^2}{\varepsilon},\frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right).$$

Despite we have linear convergence in terms of the iterations number, the number of the oracle evaluations corresponds to sublinear convergence. The reason is that we consider general stochastic optimization problem, rather than finite-sum problems for which the linear convergence rate is achievable in terms of the oracle evaluations Allen-Zhu (2017). Our oracle complexity corresponds to the lower complexity bounds Nemirovsky and Yudin (1983) for general stochastic convex optimization.

Before we proceed, we give the appropriate choice of the accuracy of the directional derivative evaluation Δ_ζ, Δ_η for RDDsc to achieve an accuracy ε of the solution. These values are chosen such that the r.h.s. of (21) is smaller than ε. For simplicity we omit numerical constants and summarize the obtained values of the algorithm parameters in Table 4 below. The last row represents the total number of oracle calls, that is, the number of directional derivative evaluations, which was stated in (6).

2.4. Corollaries for derivative-free optimization

In this section, following Gorbunov et al. (2018), we consider derivative-free smooth stochastic optimization in the two-point feedback situation. We assume that an optimization procedure, given a pair of points $(x,y)\in\mathbb{R}^{2n}$, can obtain a pair of noisy stochastic realizations $(\widetilde{f}(x,\xi),\widetilde{f}(y,\xi))$ of the objective value f, where

$$\widetilde{f}(x,\xi)=F(x,\xi)+\Xi(x,\xi),\quad |\Xi(x,\xi)|\leqslant\Delta,\ \forall x\in\mathbb{R}^n,\ \text{a.s. in }\xi, \tag{22}$$

and ξ is independently drawn from P.

Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$\begin{aligned}\widetilde{\nabla}^m f^t(x)&=\frac{1}{m}\sum_{i=1}^m\frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e\\&=\left(\langle g^m(x,\vec{\xi_m}),e\rangle+\frac{1}{m}\sum_{i=1}^m(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{23}$$

where $e\in RS_2(1)$, ξ_i, $i=1,\dots,m$ are independent realizations of ξ, m is the *batch size*, t is some small positive parameter which we call *smoothing parameter*, $g^m(x,\vec{\xi_m}):=\frac{1}{m}\sum_{i=1}^m g(x,\xi_i)$, and

$$\zeta(x,\xi_i,e)=\frac{F(x+te,\xi_i)-F(x,\xi_i)}{t}-\langle g(x,\xi_i),e\rangle,$$
$$\eta(x,\xi_i,e)=\frac{\Xi(x+te,\xi_i)-\Xi(x,\xi_i)}{t},\quad i=1,\dots,m.$$

By Lipschitz smoothness of $F(\cdot,\xi)$, we have $|\zeta(x,\xi,e)|\leqslant\frac{L(\xi)t}{2}$ for all $x\in\mathbb{R}^n$ and $e\in S_2(1)$. Hence, $\mathbb{E}_\xi(\zeta(x,\xi,e))^2\leqslant\frac{L_2^2t^2}{4}$ for all $x\in\mathbb{R}^n$ and $e\in S_2(1)$. At the same time, from (22), we have that $|\eta(x,\xi,e)|\leqslant\frac{2\Delta}{t}$ for all $x\in\mathbb{R}^n$, $e\in S_2(1)$ and a.s. in ξ. Applying Theorem 1 and Theorem 2 with $\Delta_\zeta=\frac{L_2^2t^2}{4}$ and $\Delta_\eta=\frac{2\Delta}{t}$, we reproduce respectively the result of Theorems 2 and 3 in Gorbunov et al. (2018). Applying Theorems 3 and 4 with $\Delta_\zeta=\frac{L_2^2t^2}{4}$ and $\Delta_\eta=\frac{2\Delta}{t}$, we obtain also complexity bounds (6) for derivative-free smooth stochastic strongly convex optimization, which was not yet done in the literature.

3. Proof of main result for ARDD method

We divide the proof of Theorem 1 into two large steps. First, to simplify the derivations, we prove this theorem assuming two additional inequalities which connect noisy stochastic approximation of the gradient (12) with the true gradient and function values. This result is stated as Lemma 2. Then, in Lemma 3, we show that our approximation of the gradient (12) indeed satisfies these two inequalities.

Lemma 2. *Let $\{x_k, y_k, z_k\}$, $k \geq 0$ be generated by ARDD method. Assume that there exist numbers $\delta_1 > 0$, $\delta_2 > 0$ such that, for all $k \geq 0$*

$$\mathbb{E}\left[\langle\widetilde{\nabla}^m f(x_{k+1}),\, z_k - x_*\rangle\right] \geqslant \frac{1}{n}\mathbb{E}\left[\langle\nabla f(x_{k+1}),\, z_k - x_*\rangle\right] - \delta_1\mathbb{E}[\|z_k - x_*\|] \tag{24}$$

and

$$\mathbb{E}\left[\|\widetilde{\nabla}^m f(x_{k+1})\|_q^2\right] \leqslant 96\rho_n L_2(\mathbb{E}[f(x_{k+1})] - \mathbb{E}[f(y_{k+1})]) + \delta_2, \tag{25}$$

where expectation is taken w.r.t. all randomness and x^* is a solution to (1). Then

$$\mathbb{E}[f(y_N)] - f(x^*) \leqslant \frac{384\Theta_p n^2\rho_n L_2}{N^2} + \frac{12n\sqrt{2\Theta_p}}{N^2}\delta_1 + \frac{N}{24\rho_n L_2}\delta_2 + \frac{N^2}{12\rho_n L_2}\delta_1^2, \tag{26}$$

where $\Theta_p = V[z_0](x^*)$ is defined by the chosen proximal setup and the expectation is taken w.r.t. all randomness.

This result is proved below in Section 3.1.

Lemma 3. *Let $\{x_k, y_k, z_k\}$, $k \geq 0$ be generated by ARDD method. Then (24) and (25) hold with*

$$\delta_1 = \frac{\sqrt{\Delta_\zeta}}{2\sqrt{n}} + \frac{2\Delta_\eta}{\sqrt{n}} \tag{27}$$

and

$$\delta_2 = \frac{96\rho_n}{n}\cdot\frac{\sigma^2}{m} + 61\rho_n\Delta_\zeta + 976\rho_n\Delta_\eta^2. \tag{28}$$

This result is proved below in Section 3.2.

Proof of Theorem 1. Combining Lemmas 2 and 3, we obtain (13). □

3.1. Proof Lemma 2

The following lemma estimates the progress in step 8 of ARDD method (and in step 5 of RDD method), which is a Mirror Descent step.

Lemma 4. *Assume that $z_+ = \operatorname{argmin}_{v\in\mathbb{R}^n}\left\{\alpha n\langle\widetilde{\nabla}^m f(x),\, v - z\rangle + V[z](v)\right\}$. Then, for any fixed $u \in \mathbb{R}^n$,*

$$\alpha n\mathbb{E}\left[\langle\widetilde{\nabla}^m f(x),\, z - u\rangle\right] \leqslant \frac{\alpha^2 n^2}{2}\mathbb{E}\left[\|\widetilde{\nabla}^m f(x)\|_q^2\right] + \mathbb{E}[V[z](u)] - \mathbb{E}[V[z_+](u)], \tag{29}$$

where expectation is taken w.r.t. all randomness.

Proof. For all $u \in \mathbb{R}^n$, we have

$$\begin{aligned}\alpha n\langle\widetilde{\nabla}^m f(x),\, z - u\rangle &= \alpha n\langle\widetilde{\nabla}^m f(x),\, z - z_+\rangle + \alpha n\langle\widetilde{\nabla}^m f(x),\, z_+ - u\rangle\\ &\overset{①}{\leqslant} \alpha n\langle\widetilde{\nabla}^m f(x),\, z - z_+\rangle + \langle -\nabla V[z](z_+),\, z_+ - u\rangle\\ &\overset{②}{=} \alpha n\langle\widetilde{\nabla}^m f(x),\, z - z_+\rangle + V[z](u) - V[z_+](u) - V[z](z_+)\\ &\overset{③}{\leqslant} \left(\alpha n\langle\widetilde{\nabla}^m f(x),\, z - z_+\rangle - \frac{1}{2}\|z - z_+\|_p^2\right)\\ &\quad + V[z](u) - V[z_+](u) \overset{④}{\leqslant} \frac{\alpha^2 n^2}{2}\|\widetilde{\nabla}^m f(x)\|_q^2 + V[z](u) - V[z_+](u),\end{aligned} \tag{30}$$

where ① follows from the definition of z_+, whence $\langle\nabla V[z](z_+) + \alpha n\widetilde{\nabla}^m f^t(x),\, u - z_+\rangle \geqslant 0$ for all $u \in \mathbb{R}^n$; ② follows from the "'magic identity"' Fact 5.3.3 in Ben-Tal and Nemirovski (2015) for the Bregman divergence; ③ follows from (7); and ④ follows from the Fenchel inequality $\zeta\langle s, z\rangle - \frac{1}{2}\|z\|_p^2 \leq \frac{\zeta^2}{2}\|s\|_q^2$. Taking full expectation we get (29). □

Now we prove the following lemma which estimates the one-iteration progress of the whole algorithm.

Lemma 5. *Let $\{x_k, y_k, z_k, \alpha_k, \tau_k\}$, $k \geq 0$ be generated by ARDD method. Then, under assumptions of Lemma 2,*

$$\begin{aligned}&48n^2\rho_n L_2\alpha_{k+1}^2\mathbb{E}[f(y_{k+1})] - (48n^2\rho_n L_2\alpha_{k+1}^2 - \alpha_{k+1})\mathbb{E}[f(y_k)]\\ &\quad - \mathbb{E}[V[z_k](x_*)] + \mathbb{E}[V[z_{k+1}](x_*)] - \alpha_{k+1}\delta_1 n\mathbb{E}[\|z_k - x_*\|_p]\\ &\quad - \frac{\alpha_{k+1}^2 n^2}{2}\delta_2 \leqslant \alpha_{k+1} f(x_*),\end{aligned} \tag{31}$$

where expectation is taken w.r.t. all randomness, x^* is a solution to (1).

Proof. Combining (24)–(29), we obtain

$$\begin{aligned}\alpha_{k+1}\mathbb{E}[\langle\nabla f(x_{k+1}), z_k - x_*\rangle] &\leqslant 48\alpha^2 n^2\rho_n L_2(\mathbb{E}[f(x_{k+1})] - \mathbb{E}[f(y_{k+1})])\\ &\quad + \mathbb{E}\left[V_{z_k}(x_*)\right] - \mathbb{E}[V[z_{k+1}](x_*)] + \alpha_{k+1}\delta_1 n\mathbb{E}[\|z_k - x_*\|_p] + \frac{\alpha_{k+1}^2 n^2}{2}\delta_2.\end{aligned} \tag{32}$$

Further,

$$\begin{aligned}&\alpha_{k+1}(\mathbb{E}[f(x_{k+1})] - f(x_*)) \leqslant \alpha_{k+1}\mathbb{E}[\langle\nabla f(x_{k+1}),\, x_{k+1} - x_*\rangle]\\ &= \alpha_{k+1}\mathbb{E}[\langle\nabla f(x_{k+1}),\, x_{k+1} - z_k\rangle] + \alpha_{k+1}\mathbb{E}[\langle\nabla f(x_{k+1}),\, z_k - x_*\rangle]\\ &\overset{①}{=} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\mathbb{E}[\langle\nabla f(x_{k+1}),\, y_k - x_{k+1}\rangle] + \alpha_{k+1}\mathbb{E}[\langle\nabla f(x_{k+1}),\, z_k - x_*\rangle]\\ &\overset{②}{\leqslant} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}(\mathbb{E}[f(y_k)] - \mathbb{E}[f(x_{k+1})]) + \alpha_{k+1}\mathbb{E}[\langle\nabla f(x_{k+1}),\, z_k - x_*\rangle]\\ &\overset{(32)}{\leqslant} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}(\mathbb{E}[f(y_k)] - \mathbb{E}[f(x_{k+1})]) + 48\alpha^2 n^2\rho_n L_2(\mathbb{E}[f(x_{k+1})]\\ &\quad - \mathbb{E}[f(y_{k+1})]) + \mathbb{E}\left[V_{z_k}(x_*)\right] - \mathbb{E}[V[z_{k+1}](x_*)] + \alpha_{k+1}\delta_1 n\mathbb{E}[\|z_k - x_*\|_p]\\ &\quad + \frac{\alpha_{k+1}^2 n^2}{2}\delta_2 \overset{③}{=} (48\alpha_{k+1}^2 n^2\rho_n L_2 - \alpha_{k+1})\mathbb{E}[f(y_k)]\\ &\quad - 48\alpha_{k+1}^2 n^2\rho_n L_2\mathbb{E}[f(y_{k+1})] + \alpha_{k+1}\mathbb{E}[f(x_{k+1})] + \mathbb{E}\left[V_{z_k}(x_*)\right]\\ &\quad - \mathbb{E}[V[z_{k+1}](x_*)] + \alpha_{k+1}\delta_1 n\mathbb{E}[\|z_k - x_*\|_p] + \frac{\alpha_{k+1}^2 n^2}{2}\delta_2.\end{aligned}$$

Here ① is since $x_{k+1} := \tau_k z_k + (1-\tau_k)y_k \Leftrightarrow \tau_k(x_{k+1} - z_k) = (1-\tau_k)(y_k - x_{k+1})$, ② follows from the convexity of f and the inequality $1 - \tau_k \geqslant 0$ and ③ is since $\tau_k = \frac{1}{48\alpha_{k+1}n^2\rho_n L_2}$. Rearranging the terms, we obtain the statement of the lemma. □

We are now ready to finish the proof of Lemma 2.

Proof of Lemma 3. Note that $48n^2\rho_n L_2\alpha_{k+1}^2 - \alpha_{k+1} + \frac{1}{192n^2\rho_n L_2} = 48n^2\rho_n L_2\alpha_k^2$. That is,

$$\begin{aligned}48n^2\rho_n L_2\alpha_{k+1}^2 - \alpha_{k+1} + \frac{1}{192n^2\rho_n L_2} &= \frac{(k+2)^2}{192n^2\rho_n L_2}\\ - \frac{k+2}{96n^2\rho_n L_2} + \frac{1}{192n^2\rho_n L_2} &= \frac{k^2 + 4k + 4 - 2k - 4 + 1}{192n^2\rho_n L_2}\\ &= \frac{(k+1)^2}{192n^2\rho_n L_2} = 48n^2\rho_n L_2\alpha_k^2.\end{aligned}$$

Telescoping (31) for $k = 0, 1, 2, \ldots, l-1$ for $l \leq N$ we have[3]

$$48n^2\rho_n L_2\alpha_l^2\mathbb{E}[f(y_l)] + \sum_{k=1}^{l-1}\frac{1}{192n^2\rho_n L_2}\mathbb{E}[f(y_k)]$$

[3] Note that $\alpha_1 = \frac{2}{96n^2\rho_n L_2} = \frac{1}{48n^2\rho_n L_2}$ and therefore $48n^2\rho_n L_2\alpha_1^2 - \alpha_1 = 0$.

$$
\begin{aligned}
&- V[z_0](x_*) + \mathbb{E}[V[z_l](x_*)] - \zeta_1 \sum_{k=0}^{l-1} \alpha_{k+1} \mathbb{E}[\|u - z_k\|_p] \\
&- \zeta_2 \sum_{k=0}^{l-1} \alpha_{k+1}^2 \leqslant \sum_{k=0}^{l-1} \alpha_{k+1} f(u),
\end{aligned} \tag{33}
$$

where we denoted

$$
\zeta_1 := \delta_1 n, \quad \zeta_2 := \frac{n^2}{2}\delta_2. \tag{34}
$$

We define $\Theta := V[z_0](x^*)$, $R_k := \mathbb{E}[\|x^* - z_k\|_p]$. Also, from (7), we have that $\zeta_1 \alpha_1 R_0 \leq \frac{\sqrt{2\Theta}\zeta_1}{48n^2\rho_n L_2}$. To simplify the notation, we define $B_l := \zeta_2 \sum_{k=0}^{l-1} \alpha_{k+1}^2 + \Theta + \frac{\sqrt{2\Theta}\zeta_1}{48n^2\rho_n L_2}$. Since $\sum_{k=0}^{l-1} \alpha_{k+1} = \frac{l(l+3)}{192n^2\rho_n L_2}$ and, for all $i = 1, \ldots, N$, $f(y_i) \leq f(x^*)$, we obtain from (33)

$$
\begin{aligned}
&\frac{(l+1)^2}{192n^2\rho_n L_2}\mathbb{E}[f(y_l)] \leqslant f(x^*)\left(\frac{(l+3)l}{192n^2\rho_n L_2} - \frac{l-1}{192n^2\rho_n L_2}\right) \\
&+ B_l - \mathbb{E}[V[z_l](x^*)] + \zeta_1 \sum_{k=1}^{l-1} \alpha_{k+1} R_k, \\
&0 \leqslant \frac{(l+1)^2}{192n^2\rho_n L_2}(\mathbb{E}[f(y_l)] - f(x^*)) \leqslant B_l - \mathbb{E}[V[z_l](x^*)] \\
&+ \zeta_1 \sum_{k=1}^{l-1} \alpha_{k+1} R_k,
\end{aligned} \tag{35}
$$

which gives

$$
\mathbb{E}[V[z_l](x^*)] \leqslant B_l + \zeta_1 \sum_{k=1}^{l-1} \alpha_{k+1} R_k. \tag{36}
$$

Moreover,

$$
\begin{aligned}
&\frac{1}{2}(\mathbb{E}[\|z_l - x^*\|_p])^2 \leqslant \frac{1}{2}\mathbb{E}[\|z_l - x^*\|_p^2] \leqslant \mathbb{E}[V[z_l](x^*)] \\
&\overset{(36)}{\leqslant} B_l + \zeta_1 \sum_{k=1}^{l-1} \alpha_{k+1} R_k,
\end{aligned} \tag{37}
$$

whence,

$$
R_l \leqslant \sqrt{2} \cdot \sqrt{B_l + \zeta_1 \sum_{k=1}^{l-1} \alpha_{k+1} R_k}. \tag{38}
$$

Applying Lemma 12 for $a_0 = \zeta_2 \alpha_1^2 + \Theta + \frac{\sqrt{2\Theta}\zeta_1}{48n^2\rho_n L_2}$, $a_k = \zeta_2 \alpha_{k+1}^2$, $b = \zeta_1$ for $k = 1, \ldots, N-1$, we obtain

$$
B_l + \zeta_1 \sum_{k=1}^{l-1} \alpha_{k+1} R_k \leqslant \left(\sqrt{B_l} + \sqrt{2}\zeta_1 \cdot \frac{l^2}{96n^2\rho_n L_2}\right)^2, \; l = 1, \ldots, N \tag{39}
$$

Since $V[z](x^*) \geq 0$, by inequality (35) for $l = N$ and the definition of B_l, we have

$$
\begin{aligned}
&\frac{(N+1)^2}{192n^2\rho_n L_2}(\mathbb{E}[f(y_N)] - f(x^*)) \leqslant \left(\sqrt{B_N} + \sqrt{2}\zeta_1 \cdot \frac{N^2}{96n^2\rho_n L_2}\right)^2 \\
&\overset{①}{\leqslant} 2B_N + 4\zeta_1^2 \cdot \frac{N^4}{(96n^2\rho_n L_2)^2} = 2\zeta_2 \sum_{k=0}^{l-1} \alpha_{k+1}^2 + 2\Theta \\
&+ \frac{\sqrt{2\Theta}\zeta_1}{24n^2\rho_n L_2} + 4\zeta_1^2 \cdot \frac{N^4}{(96n^2\rho_n L_2)^2} \\
&\overset{②}{\leqslant} 2\Theta + \frac{\sqrt{2\Theta}\zeta_1}{24n^2\rho_n L_2} + \frac{2\zeta_2 (N+1)^3}{(96n^2\rho_n L_2)^2} + 4\zeta_1^2 \cdot \frac{N^4}{(96n^2\rho_n L_2)^2}
\end{aligned} \tag{40}
$$

where ① is due to the fact that $\forall a, b \in \mathbb{R} \quad (a+b)^2 \leqslant 2a^2 + 2b^2$ and ② is because $\sum_{k=0}^{N-1} \alpha_{k+1}^2 = \frac{1}{(96n^2\rho_n L_2)^2} \sum_{k=2}^{N+1} k^2 \leqslant \frac{1}{(96n^2\rho_n L_2)^2} \cdot \frac{(N+1)(N+2)(2N+3)}{6} \leqslant \frac{1}{(96n^2\rho_n L_2)^2} \cdot \frac{(N+1)2(N+1)3(N+1)}{6} = \frac{(N+1)^3}{(96n^2\rho_n L_2)^2}$. Dividing (40) by $\frac{(N+1)^2}{192n^2\rho_n L_2}$ and substituting ζ_1, ζ_2 from (34), we obtain

$$
\begin{aligned}
\mathbb{E}[f(y_N)] - f(x^*) &\leqslant \frac{384\Theta n^2\rho_n L_2}{(N+1)^2} + \frac{12\sqrt{2\Theta}}{(N+1)^2}\zeta_1 + \frac{(N+1)\zeta_2}{24n^2\rho_n L_2} \\
&+ \frac{N^4\zeta_1^2}{12n^2\rho_n L_2 (N+1)^2} \\
&\leqslant \frac{384\Theta n^2\rho_n L_2}{N^2} + \frac{12n\sqrt{2\Theta}}{N^2}\delta_1 + \frac{N}{24\rho_n L_2}\delta_2 + \frac{N^2}{12\rho_n L_2}\delta_1^2.
\end{aligned}
$$

□

3.2. Proof Lemma 3

We start with the following technical result which connects our noisy approximation (12) of the stochastic gradient with the stochastic gradient itself and also with ∇f.

Lemma 6. *For all $x, s \in \mathbb{R}^n$, we have*

$$
\mathbb{E}_e \|\widetilde{\nabla}^m f(x)\|_q^2 \leqslant \tfrac{12\rho_n}{n}\|g^m(x, \vec{\xi_m})\|_2^2 + \tfrac{\rho_n}{m}\sum_{i=1}^m \zeta(x, \xi_i)^2 + 16\rho_n \Delta_\eta^2, \tag{41}
$$

$$
\mathbb{E}_e \|\widetilde{\nabla}^m f(x)\|_2^2 \geqslant \frac{1}{2n}\|g^m(x, \vec{\xi_m})\|_2^2 - \frac{1}{2m}\sum_{i=1}^m \zeta(x, \xi_i)^2 - 8\Delta_\eta^2, \tag{42}
$$

$$
\mathbb{E}_e \langle \widetilde{\nabla}^m f(x), s\rangle \geqslant \frac{1}{n}\langle g^m(x, \vec{\xi_m}), s\rangle - \frac{\|s\|_p}{2m\sqrt{n}}\sum_{i=1}^m |\zeta(x, \xi_i)| - \frac{2\Delta_\eta \|s\|_p}{\sqrt{n}}, \tag{43}
$$

$$
\begin{aligned}
&\mathbb{E}_e \|\langle \nabla f(x), e\rangle e - \widetilde{\nabla}^m f(x)\|_2^2 \leqslant \frac{2}{n}\|\nabla f(x) - g^m(x, \vec{\xi_m})\|_2^2 \\
&+ \frac{1}{m}\sum_{i=1}^m \zeta(x, \xi_i)^2 + 16\Delta_\eta^2,
\end{aligned} \tag{44}
$$

where $g^m(x, \vec{\xi_m}) := \frac{1}{m}\sum_{i=1}^m g(x, \xi_i)$, $\zeta(x, \xi_i)$ and Δ_η are defined in (3).

Proof. First of all, we rewrite $\widetilde{\nabla}^m f(x)$ as follows

$$
\widetilde{\nabla}^m f(x) = \left(\left\langle g^m(x, \vec{\xi_m}), e\right\rangle + \frac{1}{m}\sum_{i=1}^m \theta(x, \xi_i, e)\right)e,
$$

where

$$
\theta(x, \xi_i, e) = \zeta(x, \xi_i) + \eta(x, \xi_i, e), \quad i = 1, \ldots, m.
$$

By (3), we have

$$
|\theta(x, \xi_i, e)| \leq |\zeta(x, \xi_i)| + \Delta_\eta. \tag{45}
$$

Proof of (41).

$$
\begin{aligned}
&\mathbb{E}_e \|\widetilde{\nabla}^m f(x)\|_q^2 = \mathbb{E}_e \left\|\left(\left\langle g^m(x, \vec{\xi_m}), e\right\rangle + \frac{1}{m}\sum_{i=1}^m \theta(x, \xi_i, e)\right)e\right\|_q^2 \\
&\overset{①}{\leqslant} 2\mathbb{E}_e \|\langle g^m(x, \vec{\xi_m}), e\rangle e\|_q^2 + 2\mathbb{E}_e \left\|\frac{1}{m}\sum_{i=1}^m \theta(x, \xi_i, e)e\right\|_q^2 \\
&\overset{②}{\leqslant} \frac{12\rho_n}{n}\|g^m(x, \vec{\xi_m})\|_2^2 + \frac{2\rho_n}{m}\sum_{i=1}^m \left(|\zeta(x, \xi_i)| + \Delta_\eta\right)^2
\end{aligned}
$$

$$\leqslant \frac{12\rho_n}{n}\|g^m(x,\vec{\xi_m})\|_2^2 + \frac{\rho_n}{m}\sum_{i=1}^m \zeta(x,\xi_i)^2 + 16\rho_n\Delta_\eta^2, \quad (46)$$

where ① holds since $\|x+y\|_q^2 \leqslant 2\|x\|_q^2 + 2\|y\|_q^2, \forall x,y\in\mathbb{R}^n$; ② follows from inequalities (10),(11), (45) and the fact that, for any $a_1, a_2, \dots, a_m > 0$, it holds that $\left(\sum_{i=1}^m a_i\right)^2 \leqslant m\sum_{i=1}^m a_i^2$.

Proof of (42).

$$\mathbb{E}_e\|\widetilde{\nabla}^m f(x)\|_2^2 = \mathbb{E}_e\left\|\left(\langle g^m(x,\vec{\xi_m}),e\rangle + \frac{1}{m}\sum_{i=1}^m \theta(x,\xi_i,e)\right)e\right\|_2^2$$

$$\overset{①}{\geqslant} \frac{1}{2}\mathbb{E}_e\|\langle g^m(x,\vec{\xi_m}),e\rangle e\|_2^2 - \frac{1}{m}\sum_{i=1}^m\left(|\zeta(x,\xi_i)| + \Delta_\eta\right)^2$$

$$\overset{②}{\geqslant} \frac{1}{2n}\|g^m(x,\vec{\xi_m})\|_2^2 - \frac{1}{2m}\sum_{i=1}^m \zeta(x,\xi_i)^2 - 8\Delta_\eta^2, \quad (47)$$

where ① follows from (45) and inequality $\|x+y\|_2^2 \geqslant \frac{1}{2}\|x\|_2^2 - \|y\|_2^2, \forall x,y\in\mathbb{R}^n$; ② follows from $e\in S_2(1)$ and Lemma B.10 in Bogolubsky et al. (2016), stating that, for any $s\in\mathbb{R}^n$, $\mathbb{E}\langle s,e\rangle^2 = \frac{1}{n}\|s\|_2^2$.

Proof of (43).

$$\mathbb{E}_e\langle\widetilde{\nabla}^m f(x),s\rangle = \mathbb{E}_e\langle\langle g^m(x,\vec{\xi_m}),e\rangle e,s\rangle + \mathbb{E}_e\frac{1}{m}\sum_{i=1}^m \theta(x,\xi_i,e)\langle e,s\rangle$$

$$\overset{①}{\geqslant} \frac{1}{n}\langle g^m(x,\vec{\xi_m}),s\rangle - \frac{1}{m}\sum_{i=1}^m\left(|\zeta(x,\xi_i)| + \Delta_\eta\right)\mathbb{E}_e|\langle e,s\rangle|$$

$$\overset{②}{\geqslant} \frac{1}{n}\langle g^m(x,\vec{\xi_m}),s\rangle - \frac{\|s\|_p}{2m\sqrt{n}}\sum_{i=1}^m|\zeta(x,\xi_i)| - \frac{2\Delta_\eta\|s\|_p}{\sqrt{n}} \quad (48)$$

where ① follows from $\mathbb{E}_e[n\langle g,e\rangle e] = g, \forall g\in\mathbb{R}^n$ and (45); ② follows from Lemma B.10 in Bogolubsky et al. (2016), since $\mathbb{E}|\langle s,e\rangle| \leq \sqrt{\mathbb{E}\langle s,e\rangle^2}$, and the fact that $\|x\|_2 \leq \|x\|_p$ for $p\leq 2$.

Proof of (44).

$$\mathbb{E}_e\|\langle\nabla f(x),e\rangle e - \widetilde{\nabla}^m f(x)\|_2^2$$

$$= \mathbb{E}_e\left\|\langle\nabla f(x),e\rangle e - \langle g^m(x,\vec{\xi_m}),e\rangle e - \frac{1}{m}\sum_{i=1}^m\theta(x,\xi_i,e)e\right\|_2^2$$

$$\overset{①}{\leqslant} 2\mathbb{E}_e\left\|\langle\nabla f(x) - g^m(x,\vec{\xi_m}),e\rangle e\right\|_2^2 + 2\mathbb{E}_e\left\|\frac{1}{m}\sum_{i=1}^m\theta(x,\xi_i,e)e\right\|_2^2$$

$$\overset{②}{\leqslant} \frac{2}{n}\|\nabla f(x) - g^m(x,\vec{\xi_m})\|_2^2 + \frac{1}{m}\sum_{i=1}^m\zeta(x,\xi_i)^2 + 16\Delta_\eta^2, \quad (49)$$

where ① holds since $\|x+y\|_2^2 \leqslant 2\|x\|_2^2 + 2\|y\|_2^2, \forall x,y\in\mathbb{R}^n$; ② follows from $e\in S_2(1)$ and Lemma B.10 in Bogolubsky et al. (2016), and (45). □

We continue by proving the following lemma which estimates the progress in step 7 of ARDD, which is a gradient step.

Lemma 7. *Assume that* $y = x - \frac{1}{2L_2}\widetilde{\nabla}^m f(x)$. *Then,*

$$\|g^m(x,\vec{\xi_m})\|_2^2 \leq 8nL_2(f(x) - \mathbb{E}_e f(y)) + 8\|\nabla f(x) - g^m(x,\vec{\xi_m})\|_2^2 + \frac{5n}{m}\sum_{i=1}^m\zeta(x,\xi_i)^2 + 80n\Delta_\eta^2, \quad (50)$$

where $g^m(x,\vec{\xi_m})$ *is defined in* Lemma 6, $\zeta(x,\xi_i)$ *and* Δ_η *are defined in* (3).

Proof. Since $\widetilde{\nabla}^m f(x)$ is collinear to e, we have that, for some $\gamma\in\mathbb{R}$, $y - x = \gamma e$. Then, since $\|e\|_2 = 1$,

$$\langle\nabla f(x), y-x\rangle = \langle\nabla f(x),e\rangle\gamma = \langle\nabla f(x),e\rangle\langle e,y-x\rangle = \langle\langle\nabla f(x),e\rangle e,\ y-x\rangle.$$

From this and L_2-smoothness of f we obtain

$$f(y) \leqslant f(x) + \langle\langle\nabla f(x),e\rangle e,\ y-x\rangle + \frac{L_2}{2}||y-x||_2^2$$
$$\leqslant f(x) + \langle\widetilde{\nabla}^m f(x),\ y-x\rangle + L_2||y-x||_2^2 + \langle\langle\nabla f(x),e\rangle e - \widetilde{\nabla}^m f(x),\ y-x\rangle - \frac{L_2}{2}||y-x||_2^2$$
$$\overset{①}{\leqslant} f(x) + \langle\widetilde{\nabla}^m f(x),\ y-x\rangle + L_2||y-x||_2^2 + \frac{1}{2L_2}\|\langle\nabla f(x),e\rangle e - \widetilde{\nabla}^m f(x)\|_2^2,$$

where ① follows form the Fenchel inequality $\langle s,z\rangle - \frac{\zeta}{2}\|z\|_2^2 \leq \frac{1}{2\zeta}\|s\|_2^2$. Using $y = x - \frac{1}{2L_2}\widetilde{\nabla}^m f(x)$, we get

$$\frac{1}{4L_2}\|\widetilde{\nabla}^m f(x)\|_2^2 \leqslant f(x) - f(y) + \frac{1}{2L_2}\|\langle\nabla f(x),e\rangle e - \widetilde{\nabla}^m f(x)\|_2^2$$

Taking the expectation in e and applying (42), (44), we obtain

$$\frac{1}{4L_2}\left(\frac{1}{2n}\|g^m(x,\vec{\xi_m})\|_2^2 - \frac{1}{2m}\sum_{i=1}^m\zeta(x,\xi_i)^2 - 8\Delta_\eta^2\right)$$
$$\leqslant \frac{1}{4L_2}\mathbb{E}_e\|\widetilde{\nabla}^m f(x)\|_2^2$$
$$\leqslant f(x) - \mathbb{E}_e f(y) + \frac{1}{2L_2}\mathbb{E}_e\|\langle\nabla f(x),e\rangle e - \widetilde{\nabla}^m f(x)\|_2^2$$
$$\leqslant f(x) - \mathbb{E}_e f(y) + \frac{1}{2L_2}\left(\frac{2}{n}\|\nabla f(x) - g^m(x,\vec{\xi_m})\|_2^2 + \frac{t^2}{m}\sum_{i=1}^m\zeta(x,\xi_i)^2 + 16\Delta_\eta^2\right),$$

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

We are now ready to finish the proof of Lemma 3.

Proof of Lemma 3**.** Taking the expectation w.r.t. all randomness[4] of (43) and using inequality

$$\mathbb{E}[|\zeta(x,\xi_i)|] \leqslant \sqrt{\mathbb{E}[|\zeta(x,\xi_i)|^2]} \overset{(3)}{\leqslant} \sqrt{\Delta_\zeta},$$

we obtain inequality (24) with $\delta_1 = \frac{\sqrt{\Delta_\zeta}}{2\sqrt{n}} + \frac{2\Delta_\eta}{\sqrt{n}}$. Combining (41) and (50), taking the full expectation and using $\mathbb{E}[\|\nabla f(x) - g^m(x,\xi)\|_2^2] \leqslant \frac{\sigma^2}{m}$, which follows from (2), we obtain (25) with $\delta_2 = \frac{96\rho_n}{n}\cdot\frac{\sigma^2}{m} + 61\rho_n\Delta_\zeta + 976\rho_n\Delta_\eta^2$. □

4. Proof of main result for RDD method

As in the previous section, we divide the proof of Theorem 2 into large steps. First, to simplify the derivations, we prove this theorem assuming two additional inequalities which connect or noisy stochastic approximation of the gradient (12) with the true gradient and function values. Then we show that our approximation of the gradient (12) indeed satisfies these two inequalities.

Lemma 8. *Let* $\{x_k, y_k, z_k\}$, $k\geq 0$ *be generated by RDD method. Assume that there exist numbers* $\delta_1 > 0$, $\delta_2 > 0$ *such that, for all* $k\geq 0$

[4] Note that we use $s = z_k - x_*$ which does not depend on $\xi_1, \xi_2, \dots, \xi_m$ from the $(k+1)$-th iterate and it does not depend on e_{k+1}. Therefore we can use tower property of mathematical expectation and take firstly conditional expectation w.r.t. $\xi_1, \dots, \xi_m$ and after that take full expectation.

$$\mathbb{E}\left[\langle\widetilde{\nabla}^m f(x_k), x_k - x_*\rangle\right] \geqslant \frac{1}{n}\mathbb{E}[\langle\nabla f(x_k),\, x_k - x_*\rangle] - \delta_1\mathbb{E}[\|x_k - x_*\|_p] \quad (51)$$

$$\mathbb{E}\left[\|\widetilde{\nabla}^m f(x_k)\|_q^2\right] \leqslant \frac{48\rho_n L_2}{n}(\mathbb{E}[f(x_k)] - f(x_*)) + \delta_2, \quad (52)$$

where expectation is taken w.r.t. all randomness and x_* is a solution to (1). Then

$$\mathbb{E}[f(\bar{x}_N)] - f(x_*) \leqslant \frac{384n\rho_n L_2\Theta_p}{N} + \frac{n}{12\rho_n L_2}\delta_2 + \frac{8n\sqrt{2\Theta_p}}{N}\delta_1 + \frac{nN}{3L_2\rho_n}\delta_1^2, \quad (53)$$

where $\Theta_p = V[z_0](x^*)$ is defined by the chosen proximal setup and the expectation is taken w.r.t. all randomness.

This result is proved below in Section 4.1.

Lemma 9. *Let $\{x_k, y_k, z_k\}$, $k \geq 0$ be generated by RDD method. Then (51) and (52) hold with*

$$\delta_1 = \frac{\sqrt{\Delta_\zeta}}{2\sqrt{n}} + \frac{2\Delta_\eta}{\sqrt{n}} \quad (54)$$

and

$$\delta_2 = \frac{24\rho_n}{n}\cdot\frac{\sigma^2}{m} + \rho_n\Delta_\zeta + 16\rho_n\Delta_\eta^2. \quad (55)$$

This result is proved below in Section 4.2.

Proof of Theorem 2. Combining Lemmas 8 and 9, we obtain (14). □

4.1. Proof Lemma 8

Combining (29), (51) and (52) we get

$$\alpha\mathbb{E}[\langle\nabla f(x_k),\, x_k - x_*\rangle] \leqslant 24\alpha^2 n\rho_n L_2(\mathbb{E}[f(x_k)] - f(x_*)) + \alpha\delta_1 n\mathbb{E}[\|x_k - x_*\|_p] + \frac{\alpha^2 n^2}{2}\delta_2 + \mathbb{E}[V[x_k](x_*)] - \mathbb{E}[V[x_{k+1}](x_*)],$$

whence due to convexity of f we have

$$\underbrace{(\alpha - 24\alpha^2 n\rho_n L_2)}_{\frac{\alpha}{4}}(\mathbb{E}[f(x_k)] - f(x_*)) \leqslant \alpha\delta_1 n\mathbb{E}[\|x_k - x_*\|_p] + \frac{\alpha^2 n^2}{2}\delta_2 + \mathbb{E}[V[x_k](x_*)] - \mathbb{E}[V[x_{k+1}](x_*)], \quad (56)$$

because $\alpha = \frac{1}{48n\rho_n L_2}$. Summing (56) for $k = 0, \ldots, l-1$, where $l \leq N$ we get

$$0 \leqslant \frac{N\alpha}{4}(\mathbb{E}[f(\bar{x}_l)] - f(x_*)) \leqslant \frac{\alpha^2 n^2 l}{2}\delta_2 + \alpha\delta_1 n\sum_{k=0}^{l-1}\mathbb{E}[\|x_k - x_*\|_p] + \underbrace{V[x_0](x_*)}_{\Theta_p} - \mathbb{E}[V[x_l](x^*)], \quad (57)$$

where $\bar{x}_l \overset{\text{def}}{=} \frac{1}{l}\sum_{k=0}^{l-1}x_k$. From the previous inequality we get

$$\frac{1}{2}(\mathbb{E}[\|x_l - x_*\|_p])^2 \leqslant \frac{1}{2}\mathbb{E}[\|x_l - x_*\|_p^2] \leqslant \mathbb{E}[V[x_l](x_*)] \leqslant \Theta_p + l\cdot\frac{\alpha^2 n^2}{2}\delta_2 + \alpha\delta_1 n\sum_{k=0}^{l-1}\mathbb{E}[\|x_k - x_*\|_p], \quad (58)$$

whence $\forall l \leq N$ we obtain

$$\mathbb{E}[\|x_k - x_*\|_p] \leqslant \sqrt{2}\sqrt{\Theta_p + l\cdot\frac{\alpha^2 n^2}{2}\delta_2 + \alpha\delta_1 n\sum_{k=0}^{l-1}\mathbb{E}[\|x_k - x_*\|_p]}. \quad (59)$$

Denote $R_k = \mathbb{E}[\|x^* - x_k\|_p]$ for $k = 0, \ldots, N$. Applying Lemma 13 for $a_0 = \Theta_p + \alpha\delta_1 n\mathbb{E}[\|x_0 - x_*\|_p] \leqslant \Theta_p + \alpha n\sqrt{2\Theta_p}\delta_1$, $a_k = \frac{\alpha^2 n^2}{2}\delta_2$, $b = n\delta_1$ for $k = 1, \ldots, N-1$ we have for $l = N$

$$\frac{N\alpha}{4}(\mathbb{E}[f(\bar{x}_N)] - f(x_*)) \leqslant \left(\sqrt{\Theta_p + N\cdot\frac{\alpha^2 n^2}{2}\delta_2 + \alpha n\sqrt{2\Theta_p}\delta_1} + \sqrt{2}n\delta_1\alpha N\right)^2 \overset{①}{\leqslant} 2\Theta_p + N\alpha^2 n^2\delta_2 + 2\alpha n\sqrt{2\Theta_p}\delta_1 + 4n^2\delta_1^2\alpha^2 N^2,$$

whence

$$\mathbb{E}[f(\bar{x}_N)] - f(x_*) \leqslant \frac{384n\rho_n L_2\Theta_p}{N} + \frac{n}{12\rho_n L_2}\delta_2 + \frac{8n\sqrt{2\Theta_p}}{N}\delta_1 + \frac{nN}{3L_2\rho_n}\delta_1^2,$$

because $\alpha = \frac{1}{48n\rho_n L_2}$.

4.2. Proof Lemma 9

Taking mathematical expectation w.r.t. all randomness from the (43) we obtain[5] inequality (51) with $\delta_1 = \frac{\sqrt{\Delta_\zeta}}{2\sqrt{n}} + \frac{2\Delta_\eta}{\sqrt{n}}$, because $\mathbb{E}[|\zeta(x, \xi_i)|] \leqslant \sqrt{\mathbb{E}[|\zeta(x, \xi_i)|^2]} \overset{(3)}{\leqslant} \sqrt{\Delta_\zeta}$. Combining (41) and

$$\|g^m(x, \vec{\xi}_m)\|_2^2 \leqslant 2\|\nabla f(x)\|_2^2 + 2\|\nabla f(x) - g^m(x, \vec{\xi}_m)\|_2^2 \leqslant 4L_2(\mathbb{E}[f(x)] - f(x_*)) + 2\|\nabla f(x) - g^m(x, \vec{\xi}_m)\|_2^2,$$

$$\times\ \mathbb{E}[\|\nabla f(x) - g^m(x, \vec{\xi}_m)\|_2^2] \leqslant \frac{\sigma^2}{m}$$

and taking full mathematical expectation we obtain (52) with $\delta_2 = \frac{24\rho_n}{n}\cdot\frac{\sigma^2}{m} + \rho_n\Delta_\zeta + 16\rho_n\Delta_\eta^2$.

5. Proofs for strongly convex problems

5.1. Accelerated algorithm

Lemma 10. *Assume that we start ARDD Algorithm 1 from a random point x_0 such that $\mathbb{E}_{x_0}\|x^* - x_0\|_p^2 \leqslant R_p^2$, use the function $R_p^2\, d\left(\frac{x - x_0}{R_p}\right)$ as the prox-function and run ARDD for N_0 iterations. Then*

$$\mathbb{E}[f(y_{N_0})] - f^* \leqslant \frac{aL_2R_p^2\Omega_p}{N_0^2} + \frac{b\sigma^2 N_0}{mL_2} + \Delta,$$

where $a = 384n^2\rho_n$, $b = \frac{4}{n}$, $\Delta = \frac{61N_0}{24L_2}\Delta_\zeta + \frac{122N_0}{3L_2}\Delta_\eta^2 + \frac{12\sqrt{2nR_p^2\Omega_p}}{N_0^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2$ and the expectation is taken with respect to all the randomness.

Proof. Note that $R_p^2\, d\left(\frac{x - x_0}{R_p}\right)$ is strongly convex with constant 1 w.r.t $\|\cdot\|_p$. Since $0 = \arg\min d(x)$, we have, for the prox-function $\bar{d}(x) = R_p^2\, d\left(\frac{x - x_0}{R_p}\right)$ and corresponding Bregman divergence $\bar{V}[x_0](x)$,

$$\Theta_p = \bar{V}[x_0](x_*) = \bar{d}(x_*) - \bar{d}(x_0) - \langle\nabla\bar{d}(x_0), x_* - x_0\rangle = \bar{d}(x_*) \leq \frac{R_p^2\Omega_p}{2}.$$

[5] Note that we use $s = x_k - x_*$ which does not depend on $\xi_1, \xi_2, \ldots, \xi_m$ from the $(k+1)$th iterate and it does not depend on e_{k+1}. Therefore we can use tower property of mathematical expectation and take firstly conditional expectation w.r.t. $\xi_1, \ldots, \xi_m$ and after that take full expectation.

Applying Theorem 1 an taking additional expectation w.r.t to x_0, we finish the proof of the lemma. □

Proof of Theorem 3. We prove by induction that

$$\mathbb{E}\|u_k - x^*\|_p^2 \le R_k^2 = R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right). \tag{60}$$

For $k = 0$, this inequality obviously holds. Let us assume that it holds for some $k \ge 0$ and prove the induction step. Applying Lemma 10 at the step k of Algorithm 3, we obtain that

$$\mathbb{E}f(u_{k+1}) - f^* = \mathbb{E}f(y_{N_0}) - f^* \leqslant \frac{aL_2R_k^2\Omega_p}{N_0^2} + \frac{b\sigma^2 N_0}{m_kL_2} + \Delta.$$

By definition of N_0, we have

$$\frac{aL_2R_k^2\Omega_p}{N_0^2} \leqslant \frac{aL_2R_k^2\Omega_p}{\frac{8aL_2\Omega_p}{\mu_p}} = \frac{\mu_pR_k^2}{8}.$$

By definition of m_k, we have

$$m_k \geqslant \frac{8b\sigma^2N_0}{L_2\mu_pR_p^2 2^{-k}} \geqslant \frac{8b\sigma^2N_0}{L_2\mu_p\left(R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right)\right)} = \frac{8b\sigma^2N_0}{L_2\mu_pR_k^2}$$

and

$$\frac{b\sigma^2N_0}{m_kL_2} \leqslant \frac{b\sigma^2N_0}{L_2\frac{8b\sigma^2N_0}{L_2\mu_pR_k^2}} = \frac{\mu_pR_k^2}{8}.$$

Hence,

$$\mathbb{E}f(u_{k+1}) - f^* \leqslant \frac{\mu_pR_k^2}{4} + \Delta = \frac{\mu_p}{4}\left(R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right)\right)$$
$$+\Delta = \frac{\mu_p}{2}\left(R_p^2 2^{-(k+1)} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-(k+1)}\right)\right) = \frac{\mu_pR_{k+1}^2}{2}.$$

Since f is strongly convex, we have

$$\mathbb{E}\|u_{k+1} - x^*\|_p^2 \leqslant \frac{2}{\mu_p}\left(\mathbb{E}f(u_{k+1}) - f^*\right) \leqslant R_{k+1}^2.$$

This finishes the induction step and, as a byproduct, we obtain inequality (18).

It remains to estimate the complexity. To make the right hand side of (18) smaller than ε it is sufficient to choose $K = \left\lceil \log_2 \frac{\mu_pR_p^2}{\varepsilon}\right\rceil$. To estimate the total number of oracle calls, we write

$$\begin{aligned}\text{Number of calls} &= \sum_{k=0}^{K-1} N_0m_k \leqslant \sum_{k=0}^{K-1} N_0\left(1 + \frac{8b\sigma^2N_0 2^k}{L_2\mu_pR_p^2}\right)\\ &\leqslant KN_0 + \frac{8b\sigma^2N_0^2 2^K}{L_2\mu_pR_p^2}\\ &\leqslant \sqrt{\frac{8aL_2\Omega_p}{\mu_p}}\log_2\frac{\mu_pR_p^2}{\varepsilon} + \frac{8b\sigma^2}{L_2\mu_pR_p^2}\cdot\frac{8aL_2\Omega_p}{\mu_p}\cdot\frac{\mu_pR_p^2}{\varepsilon}\\ &\leqslant \sqrt{\frac{8aL_2\Omega_p}{\mu_p}}\log_2\frac{\mu_pR_p^2}{\varepsilon} + \frac{64ab\sigma^2\Omega_p}{\mu_p\varepsilon}\\ &= \tilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2\Omega_p}{\mu_p}}\log_2\frac{\mu_pR_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right),\end{aligned}$$

where we used that $a = 384n^2\rho_n$, $b = \frac{4}{n}$ and ρ_n is given in Lemma 1. □

5.2. Non-accelerated algorithm

Lemma 11. *Assume that we start RDD Algorithm 2 from a random point x_0 such that $\mathbb{E}_{x_0}\|x^* - x_0\|_p^2 \leqslant R_p^2$, use the function $R_p^2\ d\left(\frac{x-x_0}{R_p}\right)$ as the prox-function and run RDD for N_0 iterations. Then*

$$\mathbb{E}[f(y_{N_0})] - f^* \leqslant \frac{aL_2R_p^2\Omega_p}{N_0} + \frac{b\sigma^2}{mL_2} + \Delta,$$

where $a = 192n\rho_n$, $b = 2$, $\Delta = \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 + \frac{8\sqrt{2nR_p^2\Omega_p}}{N_0}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2$ *and the expectation is taken with respect to all the randomness.*

Proof. Note that $R_p^2\ d\left(\frac{x-x_0}{R_p}\right)$ is strongly convex with constant 1 w.r.t $\|\cdot\|_p$. Since $0 = \arg\min d(x)$, we have, for the prox-function $\bar{d}(x) = R_p^2\ d\left(\frac{x-x_0}{R_p}\right)$ and corresponding Bregman divergence $\bar{V}[x_0](x)$,

$$\Theta_p = \bar{V}[x_0](x_*) = \bar{d}(x_*) - \bar{d}(x_0) - \langle\nabla\bar{d}(x_0), x_* - x_0\rangle = \bar{d}(x_*) \le \frac{R_p^2\Omega_p}{2}.$$

Applying Theorem 2 an taking additional expectation w.r.t to x_0, we finish the proof of the lemma. □

Proof of Theorem 4. We prove by induction that

$$\mathbb{E}\|u_k - x^*\|_p^2 \le R_k^2 = R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right). \tag{61}$$

For $k = 0$, this inequality obviously holds. Let us assume that it holds for some $k \ge 0$ and prove the induction step. Applying Lemma 11 at the step k of Algorithm 4, we obtain that

$$\mathbb{E}f(u_{k+1}) - f^* = \mathbb{E}f(y_{N_0}) - f^* \leqslant \frac{aL_2R_k^2\Omega_p}{N_0} + \frac{b\sigma^2}{m_kL_2} + \Delta.$$

By definition of N_0, we have

$$\frac{aL_2R_k^2\Omega_p}{N_0} \leqslant \frac{aL_2R_k^2\Omega_p}{\frac{8aL_2\Omega_p}{\mu_p}} = \frac{\mu_pR_k^2}{8}.$$

By definition of m_k, we have

$$m_k \geqslant \frac{8b\sigma^2}{L_2\mu_pR_p^2 2^{-k}} \geqslant \frac{8b\sigma^2}{L_2\mu_p\left(R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right)\right)} = \frac{8b\sigma^2}{L_2\mu_pR_k^2}$$

and

$$\frac{b\sigma^2}{m_kL_2} \leqslant \frac{b\sigma^2}{L_2\frac{8b\sigma^2}{L_2\mu_pR_k^2}} = \frac{\mu_pR_k^2}{8}.$$

Hence,

$$\begin{aligned}\mathbb{E}f(u_{k+1}) - f^* &\leqslant \frac{\mu_pR_k^2}{4} + \Delta = \frac{\mu_p}{4}\left(R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right)\right)\\ &\quad +\Delta = \frac{\mu_p}{2}\left(R_p^2 2^{-(k+1)} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-(k+1)}\right)\right)\\ &\quad = \frac{\mu_pR_{k+1}^2}{2}.\end{aligned}$$

Since f is strongly convex, we have

$$\mathbb{E}\|u_{k+1} - x^*\|_p^2 \leqslant \frac{2}{\mu_p}\left(\mathbb{E}f(u_{k+1}) - f^*\right) \leqslant R_{k+1}^2.$$

This finishes the induction step and, as a byproduct, we obtain inequality (21).

It remains to estimate the complexity. To make the right hand side of (21) smaller than ε it is sufficient to choose $K =$

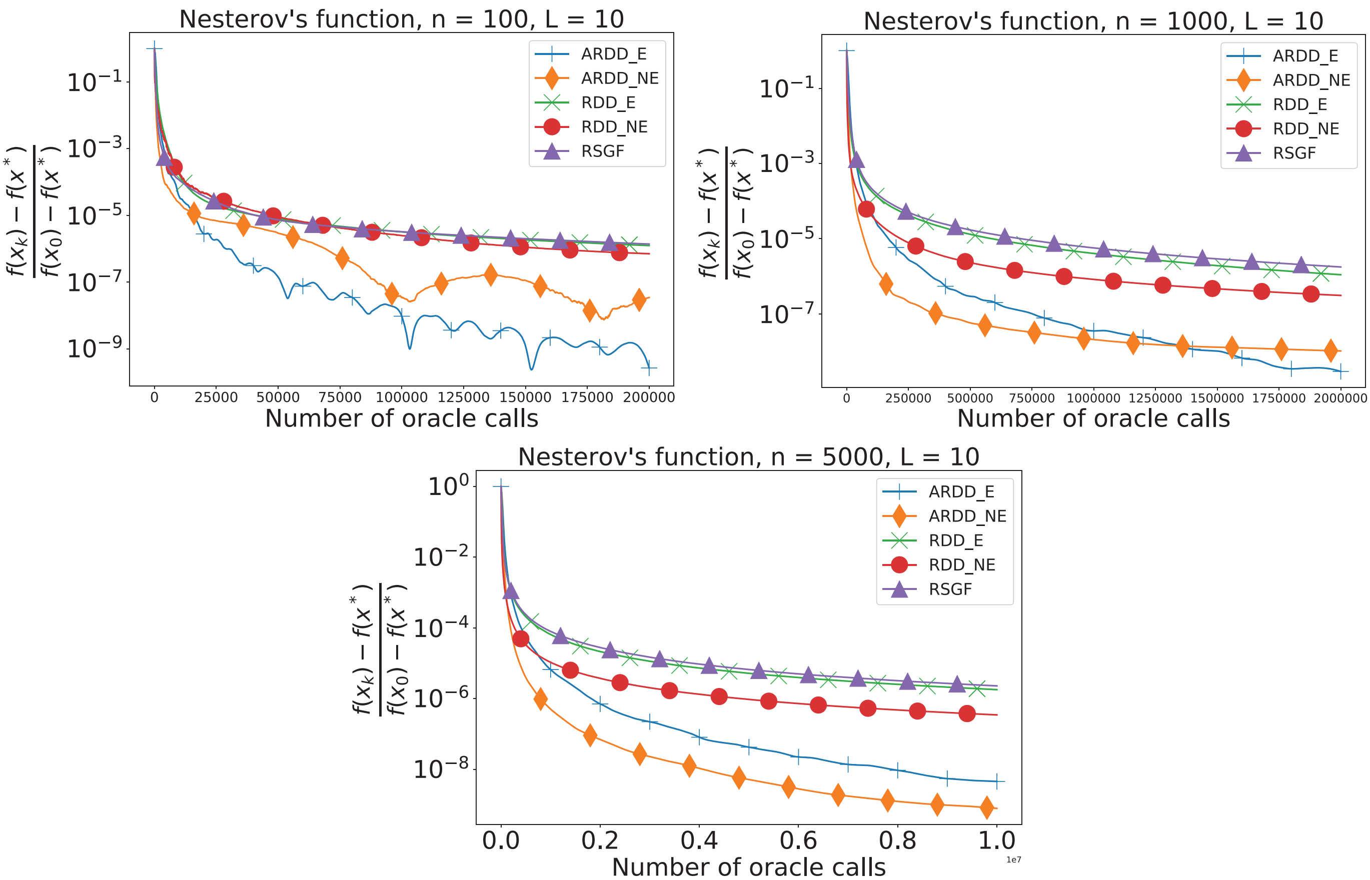

Fig. 1. ARDD, RDD and RSGF applied to minimize Nesterov's function (62). We use _E and _NE to define ℓ_2 and ℓ_1 proximal setups respectively (see (8) and (9) for the details). In the plot for $n = 5000$ number of oracle calls is divided by 10^7.

$\left\lceil \log_2 \frac{\mu_p R_p^2}{\varepsilon} \right\rceil$. To estimate the total number of oracle calls, we write

$$\begin{aligned}
\text{Number of calls} &= \sum_{k=0}^{K-1} N_0 m_k \leqslant \sum_{k=0}^{K-1} N_0 \left(1 + \frac{8b\sigma^2 2^k}{L_2 \mu_p R_p^2}\right) \\
&\leqslant K N_0 + \frac{8b\sigma^2 N_0 2^K}{L_2 \mu_p R_p^2} \\
&\leqslant \frac{8aL_2\Omega_p}{\mu_p} \log_2 \frac{\mu_p R_p^2}{\varepsilon} + \frac{8b\sigma^2}{L_2 \mu_p R_p^2} \cdot \frac{8aL_2\Omega_p}{\mu_p} \cdot \frac{\mu_p R_p^2}{\varepsilon} \\
&\leqslant \frac{8aL_2\Omega_p}{\mu_p} \log_2 \frac{\mu_p R_p^2}{\varepsilon} + \frac{64ab\sigma^2\Omega_p}{\mu_p \varepsilon} \\
&= \tilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}} L_2 \Omega_p}{\mu_p} \log_2 \frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}} \sigma^2 \Omega_p}{\mu_p \varepsilon}\right\}\right),
\end{aligned}$$

where we used that $a = 192 n \rho_n$, $b = 2$ and ρ_n is given in Lemma 1. □

6. Numerical experiments

In this section we numerically test our methods on the "worst in the world" function from Nesterov (2004) and least squares problem. In these problems there is no noise of type $\eta(x, \xi, e)$ from (3) since one can compute directional derivatives with machine precision. Moreover, for both examples one can compute exact functional values, therefore, using small enough smoothing parameter t (see (23)) it is possible to approximate directional derivatives via finite differences with high enough accuracy. That is, for the problems we consider in this section the difference between directional derivative oracle and derivative-free oracle is negligible to influence the behaviour of our methods. Taking it into account we consider only derivative-free oracle in the experiments and compare our methods with RSGF from Ghadimi and Lan (2013).

6.1. *Nesterov's function*

We start with numerical tests on Nesterov's function

$$f(x) = \frac{L}{8}\left(x_1^2 + \sum_{i=0}^{n-1} (x_i - x_{i+1})^2 + x_i^2\right) - \frac{L}{4} x_1 \tag{62}$$

which is convex, L-smooth and attains its minimal value $f^* = \frac{L}{8}\left(-1 + \frac{1}{n+1}\right)$ at such $x^* = (x_1^*, \ldots, x_n^*)^\top$ that $x_i^* = 1 - \frac{i}{n+1}$ for $i = 1, \ldots, n$ Nesterov (2004). We take the starting point x_0 such that all coordinates expect the first one coincides with corresponding coordinates of x^* and we take 10 as the first coordinate of x_0. We also choose $L = 10$, $t = 10^{-8}$ and consider $n = 100, 1000, 5000$. The results can be found in Fig. 1.

In these settings $\|x_0 - x^*\|_1 = \|x_0 - x^*\|_2$ and our theory establishes (see Tables 1 and 2) better complexity bounds for the case when $p = 1$ then for the Euclidean case especially for big n. The experiments confirm this claim: as one can see in Fig. 1, the choice of ℓ_1 proximal setup becomes more beneficial than standard Euclidean setup for $n = 1000$ and $n = 5000$ to reach good enough accuracy. Indeed, our choice of the starting point and L implies that $f(x_0) - f(x^*) \approx 200$ and for $n = 1000$ and $n = 5000$ ARDD with ℓ_1 proximal setup (ARDD_NE in Fig. 1) make $f(x_N) - f(x^*)$ of order $10^{-3} - 10^{-5}$ faster than ARDD with $p = 2$ (ARDD_E in

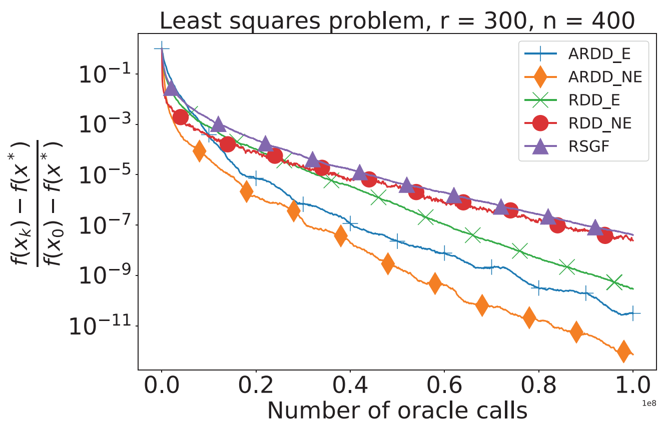

Fig. 2. ARDD, RDD and RSGF applied to solve least squares problem (63). We use _E and _NE to define ℓ_2 and ℓ_1 proximal setups respectively (see (8) and (9) for the details). For all methods batch size m equals 50. By oracle call we mean one computation of functional value of a summand. Number of oracle calls is divided by 10^8.

Fig. 1) and RDD with $p=1$ (RDD_NE in Fig. 1) finds such x^N that $f(x^N)-f(x^*)$ is of order 10^{-3} faster than its Euclidean counterpart (RDD_E in Fig. 1). Finally, all of our methods outperform RSGF on the considered problem.

To perform mirror descent step for $p=1$ we apply relations obtained in Appendix B from Gorbunov et al. (2018). See other details connected with parameters tuning in Appendix C of this work.

6.2. Least squares problem

In this section we consider least squares problem:

$$\min_{x\in\mathbb{R}^n}\left\{f(x)=\frac{1}{2r}\|Ax-b\|_2^2=\frac{1}{r}\sum_{i=1}^{r}\frac{1}{2}(A_ix-b_i)^2\right\}. \tag{63}$$

Here A is $r\times n$ real matrix, $b\in\mathbb{R}^r$ and A_i denotes the ith row of A. Clearly, $f(x)$ is convex and smooth function. Moreover, each summand $f_i(x)=\frac{1}{2}(A_ix-b_i)^2$ is also convex and $L_{2,i}$-smooth function with $L_{2,i}=\|A_i\|_2^2$. One can consider (63) as (1) with $F(x,\xi)=f_\xi(x)=\frac{1}{2}(A_\xi x-b_\xi)^2$ where ξ is uniformly distributed on $\{1,2,\ldots,r\}$. Then, by definition of L_2 we have

$$L_2=\sqrt{\mathbb{E}_\xi L_{2,\xi}^2}=\sqrt{\frac{1}{r}\sum_{i=1}^{r}\|A_i\|_2^2}=\frac{\|A\|_F}{\sqrt{r}} \tag{64}$$

where $\|A\|_F$ denotes Frobenius norm of matrix A.

In our preliminary experiments elements of A and b were sampled independently from the standard normal distribution and then matrix A was normalized by its ℓ_2-norm. In particular, we choose $r=300$ and $n=400$ which implies that $f(x)$ is just convex but not strongly convex and $f(x^*)=0$. Moreover, we compute the solution x^* as A^+b where A^+ denotes Moore-Penrose inverse of A and choose the starting point x_0 as x^* and 100 to the first component. In our tests the suboptimality of the starting point, i.e. $f(x_0)-f(x^*)$, was approximately 3. The results can be found in Fig. 2. We want to notice that in these preliminary experiments with stochasticity in functional values in experiments with ARDD it was needed to tune not only α_{k+1} that appears in the mirror descent step, but also the stepsize for the gradient step, see the details in Appendix C.

7. Conclusion

In this paper we propose four novel directional derivative methods for smooth stochastic convex and strongly convex optimization with corollaries for derivative-free optimization. These methods are able to work with Euclidean and non-Euclidean proximal setups. We prove complexity results showing that in non-Euclidean case complexities of our methods outperform state-of-the-art results for directional derivative and derivative-free methods in terms of the dependence on the dimension of the problem under assumption that ℓ_1 and ℓ_2 norms of x_0-x^* are close to each other, e.g. when $x_0=0$ and x^* is sparse. Moreover, we analyze our methods under general assumptions on the noisy oracle and provide bounds for the admissible noise levels. Since we use mini-batches, we are able to separate iteration complexity and sample complexity, the former being up to a dimension-dependent factor the same as for accelerated gradient method in the standard deterministic full-gradient setting. This makes our methods amenable to parallel computation setting Dvurechensky, Gasnikov, and Lagunovskaya (2018b) and leads to acceleration in this setting compared to standard stochastic gradient methods Duchi et al. (2015). Finally, we conduct several experiments providing numerical justifications of the obtained results.

Using an additional "light-tail" assumption that $\mathbb{E}_\xi[\exp(\|g(x,\xi)-\nabla f(x)\|_2^2/\sigma^2)]\leqslant\exp(1)$ and techniques of Gorbunov, Dvinskikh, and Gasnikov (2019) our algorithms and analysis can be extended to obtain results in terms of probability of large deviations. For example, in the case of controlled noise levels Δ_ζ, Δ_η this means that an algorithm outputs a point $\hat{x}$ which satisfies $\mathbb{P}\{f(\hat{x})-f(x^*)\leqslant\varepsilon\}\geqslant 1-\delta$, where $\delta\in(0,\ 1)$ is the confidence level, for the price of extra $\ln\frac{1}{\delta}$ factor in N and m. As directions of future research we would like to point a primal-dual extension for problems with linear constraints in the spirit of (Dvurechensky et al., 2016; Chernov, Dvurechensky, and Gasnikov, 2016; Anikin, Gasnikov, Dvurechensky, Tyurin, and Chernov, 2017; Bayandina, Dvurechensky, Gasnikov, Stonyakin, and Titov, 2018a; Dvurechensky, Dvinskikh, Gasnikov, Uribe, and Nedić, 2018a; Dvinskikh, Gorbunov, Gasnikov, Dvurechensky, and Uribe, 2019; Nesterov, Gasnikov, Guminov, and Dvurechensky, 2020), an extension with line-search to adapt to an unknown value of L_2 using the techniques in (Cartis and Scheinberg, 2018; Berahas, Cao, and Scheinberg, 2019c; Dvinskikh, Ogaltsov, Gasnikov, Dvurechensky, and Spokoiny, 2020), an extension for the case of intermediate smoothness (Kamzolov, Dvurechensky, and Gasnikov, 2020; Nesterov, 2015) or interpolation between accelerated and non-accelerated methods (Dvurechensky and Gasnikov, 2016; Gasnikov and Dvurechensky, 2016), as well as extension to a more general type of inexactness called inexact model of the objective Stonyakin et al. (2020, 2019).

Acknowledgments

The research is supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) 075-00337-20-03, project No. 0714-2020-0005.

Appendix A. Proof of Lemma 1

Here we prove that, for $e\in RS_2(1)$

$$\mathbb{E}[\|e\|_q^2]\leqslant\min\{q-1,\ 16\ln n-8\}n^{\frac{2}{q}-1}, \tag{A.1}$$

$$\mathbb{E}[\langle s,\ e\rangle^2\|e\|_q^2]\leqslant 6\|s\|_2^2\min\{q-1,16\ln n-8\}n^{\frac{2}{q}-2}. \tag{A.2}$$

We start with proving the following inequality which could be rough for big q:

$$\mathbb{E}[\|e\|_q^2]\leqslant(q-1)n^{\frac{2}{q}-1},\quad 2\leqslant q<\infty. \tag{A.3}$$

We have

$$\mathbb{E}[\|e\|_q^2] = \mathbb{E}\left[\left(\sum_{k=1}^n |e_k|^q\right)^{\frac{2}{q}}\right] \overset{①}{\leqslant} \left(\mathbb{E}\left[\sum_{k=1}^n |e_k|^q\right]\right)^{\frac{2}{q}} \overset{②}{=} (n\mathbb{E}[|e_2|^q])^{\frac{2}{q}}, \tag{A.4}$$

where ① is due to probabilistic version of Jensen's inequality (function $\varphi(x) = x^{\frac{2}{q}}$ is concave, because $q \geq 2$) and ② is because mathematical expectation is linear and components of vector e are identically distributed.

Moreover, due to Poincare lemma, we have

$$e =^d \frac{\xi}{\sqrt{\xi_1^2 + \dots + \xi_n^2}}, \tag{A.5}$$

where ξ is Gaussian random vector which mathematical expectation is zero vector and covariance matrix is identical. Then

$$\mathbb{E}[|e_2|^q] = \mathbb{E}\left[\frac{|\xi_2|^q}{\left(\xi_1^2 + \ldots + \xi_n^2\right)^{\frac{q}{2}}}\right]$$
$$= \int\cdots\int_{\mathbb{R}^n} |x_2|^q \left(\sum_{k=1}^n x_k^2\right)^{-\frac{q}{2}} \cdot \frac{1}{(2\pi)^{\frac{n}{2}}}$$
$$\times \exp\left(-\frac{1}{2}\sum_{k=1}^n x_k^2\right) dx_1 \ldots dx_n.$$

Consider spherical coordinates:

$$x_1 = r\cos\varphi\sin\theta_1 \ldots \sin\theta_{n-2},$$
$$x_2 = r\sin\varphi\sin\theta_1 \ldots \sin\theta_{n-2},$$
$$x_3 = r\cos\theta_1\sin\theta_2 \ldots \sin\theta_{n-2},$$
$$x_4 = r\cos\theta_2\sin\theta_3 \ldots \sin\theta_{n-2},$$
$$\ldots$$
$$x_n = r\cos\theta_{n-2},$$
$$r > 0,\ \varphi \in [0, 2\pi),\ \theta_i \in [0, \pi],\ i = \overline{1, n-2}.$$

The Jacobian of mapping is

$$\det\left(\frac{\partial(x_1, \ldots, x_n)}{\partial(r, \varphi, \theta_1, \theta_2, \ldots, \theta_{n-2})}\right) = r^{n-1}\sin\theta_1(\sin\theta_2)^2 \ldots (\sin\theta_{n-2})^{n-2}.$$

Then mathematical expectation $\mathbb{E}[|e_2|^q]$ could be rewritten in the following form:

$$\mathbb{E}[|e_2|^q]$$
$$= \int\cdots\int_{\substack{r>0,\ \varphi\in[0,2\pi),\\ \theta_i\in[0,\pi],\ i=\overline{1,n-2}}} r^{n-1}|\sin\varphi|^q|\sin\theta_1|^{q+1}|\sin\theta_2|^{q+2}\ldots|$$
$$\times \sin\theta_{n-2}|^{q+n-2} \cdot \frac{e^{-\frac{r^2}{2}}}{(2\pi)^{\frac{n}{2}}} dr \ldots d\theta_{n-2}$$
$$= \frac{1}{(2\pi)^{\frac{n}{2}}} I_r \cdot I_\varphi \cdot I_{\theta_1} \cdot I_{\theta_2} \cdot \ldots \cdot I_{\theta_{n-2}},$$

where

$$I_r = \int_0^{+\infty} r^{n-1} e^{-\frac{r^2}{2}} dr,$$
$$I_\varphi = \int_0^{2\pi} |\sin\varphi|^q d\varphi = 2\int_0^{\pi} |\sin\varphi|^q d\varphi,$$
$$I_{\theta_i} = \int_0^{\pi} |\sin\theta_i|^{q+i} d\theta_i,\ i = \overline{1, n-2}.$$

Now we are going to compute these integrals. Start with I_r:

$$I_r = \int_0^{+\infty} r^{n-1} e^{-\frac{r^2}{2}} dr = /r = \sqrt{2t}/ = \int_0^{+\infty} (2t)^{\frac{n}{2}-1} e^{-t} dt = 2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}\right).$$

To compute other integrals it is useful to consider the following integral ($\alpha > 0$):

$$\int_0^{\pi} |\sin\varphi|^\alpha d\varphi = 2\int_0^{\frac{\pi}{2}} |\sin\varphi|^\alpha d\varphi$$
$$= 2\int_0^{\frac{\pi}{2}} (\sin^2\varphi)^{\frac{\alpha}{2}} d\varphi = /t = \sin^2\varphi/$$
$$= \int_0^1 t^{\frac{\alpha-1}{2}}(1-t)^{-\frac{1}{2}} dt = B\left(\frac{\alpha+1}{2}, \frac{1}{2}\right)$$
$$= \frac{\Gamma\left(\frac{\alpha+1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{\alpha+2}{2}\right)} = \sqrt{\pi}\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha+2}{2}\right)}.$$

From this we obtain

$$\mathbb{E}[|e_2|^q] = \frac{1}{(2\pi)^{\frac{n}{2}}} I_r \cdot I_\varphi \cdot I_{\theta_1} \cdot I_{\theta_2} \cdot \ldots \cdot I_{\theta_{n-2}}$$
$$= \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot 2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}\right) \cdot 2\sqrt{\pi}\frac{\Gamma\left(\frac{q+1}{2}\right)}{\Gamma\left(\frac{q+2}{2}\right)}$$
$$\cdot \sqrt{\pi}\frac{\Gamma\left(\frac{q+2}{2}\right)}{\Gamma\left(\frac{q+3}{2}\right)} \cdot \sqrt{\pi}\frac{\Gamma\left(\frac{q+3}{2}\right)}{\Gamma\left(\frac{q+4}{2}\right)} \cdot \ldots \cdot \sqrt{\pi}\frac{\Gamma\left(\frac{q+n-1}{2}\right)}{\Gamma\left(\frac{q+n}{2}\right)}$$
$$= \frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{q+1}{2}\right)}{\Gamma\left(\frac{q+n}{2}\right)}. \tag{A.6}$$

Now, we want to show that $\forall q \geq 2$

$$\frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n}{2})\Gamma(\frac{q+1}{2})}{\Gamma(\frac{q+n}{2})} \leqslant \left(\frac{q-1}{n}\right)^{\frac{q}{2}}. \tag{A.7}$$

At the beginning show that (A.7) holds for $q = 2$ (and arbitrary n):

$$\frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n}{2})\Gamma(\frac{2+1}{2})}{\Gamma(\frac{2+n}{2})} - \frac{1}{n} = \frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n}{2}) \cdot \frac{1}{2}\Gamma(\frac{1}{2})}{\frac{n}{2}\Gamma(\frac{n}{2})}$$
$$-\frac{1}{n} = \frac{1}{n} - \frac{1}{n} = 0 \leqslant 0.$$

Consider the function

$$f_n(q) = \frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n}{2})\Gamma(\frac{q+1}{2})}{\Gamma(\frac{q+n}{2})} - \left(\frac{q-1}{n}\right)^{\frac{q}{2}}$$

where $q \geq 2$. Also consider $\psi(x) = \frac{d(\ln(\Gamma(x)))}{dx}$ with $x > 0$ which is called (*digamma function*). For gamma function it holds

$$\Gamma(x+1) = x\Gamma(x),\ x > 0.$$

Taking natural logarithm from it and taking derivative w.r.t. x:

$$\ln\Gamma(x+1) = \ln\Gamma(x) + \ln x,$$
$$\frac{d(\ln(\Gamma(x+1)))}{dx} = \frac{d(\ln(\Gamma(x)))}{dx} + \frac{1}{x},$$

which could be written in digamma-function-notation:

$$\psi(x+1) = \psi(x) + \frac{1}{x}. \tag{A.8}$$

One can show that digamma function is monotonically increases when $x > 0$. To prove this fact we are going to show that

$$\left(\Gamma'(x)\right)^2 < \Gamma(x)\Gamma''(x). \tag{A.9}$$

That is,

$$\left(\Gamma'(x)\right)^2 = \left(\int_0^{+\infty} e^{-t}\ln t \cdot t^{x-1} dt\right)^2$$
$$\overset{①}{<} \int_0^{+\infty}\left(e^{-\frac{t}{2}}t^{\frac{x-1}{2}}\right)^2 dt \cdot \int_0^{+\infty}\left(e^{-\frac{t}{2}}t^{\frac{x-1}{2}}\ln t\right)^2 dt$$

$$= \underbrace{\int_0^{+\infty} e^{-t}t^{x-1}dt}_{\Gamma(x)} \cdot \underbrace{\int_0^{+\infty} e^{t}t^{x-1}\ln^2 t dt}_{\Gamma''(x)},$$

where ① follows from Cauchy-Schwartz inequality (the equality cannot occur because functions $e^{-\frac{t}{2}}t^{\frac{x-1}{2}}$ and $e^{-\frac{t}{2}}t^{\frac{x-1}{2}}\ln t$ are linearly independent). From (A.9) follows that

$$\frac{d^2(\ln\Gamma(x))}{dx^2} = \left(\frac{\Gamma'(x)}{\Gamma(x)}\right)' = \frac{\Gamma''(x)}{\Gamma(x)} - \frac{\left(\Gamma'(x)\right)^2}{(\Gamma(x))^2} \overset{(A.9)}{>} 0,$$

which shows that digamma function increases.

Now we show that $f_n(q)$ decreases on the interval $[2,+\infty)$. To obtain it is sufficient to consider $\ln(f(q))$:

$$\ln(f_n(q)) = \ln\left(\frac{\Gamma(\frac{n}{2})}{\sqrt{\pi}}\right) + \ln\left(\Gamma\left(\frac{q+1}{2}\right)\right) - \ln\left(\Gamma\left(\frac{q+n}{2}\right)\right) - \frac{q}{2}(\ln(q-1) - \ln n), \quad \frac{d(\ln(f_n(q)))}{dq} = \frac{1}{2}\psi\left(\frac{q+1}{2}\right) - \frac{1}{2}\psi\left(\frac{q+n}{2}\right) - \frac{1}{2}\ln(q-1) - \frac{q}{2(q-1)} + \frac{1}{2}\ln n.$$

We are going to show that $\frac{d(\ln(f_n(q)))}{dq} < 0$ for $q \ge 2$. Let $k = \lfloor\frac{n}{2}\rfloor$ (the closest integer which is no greater than $\frac{n}{2}$). Then $\psi\left(\frac{q+n}{2}\right) > \psi\left(k-1+\frac{q+1}{2}\right)$ and $\ln n \leqslant \ln(2k+1)$, whence

$$\begin{aligned}
\frac{d(\ln(f_n(q)))}{dq} &< \frac{1}{2}\left(\psi\left(\frac{q+1}{2}\right) - \psi\left(k-1+\frac{q+1}{2}\right)\right) - \frac{1}{2}\ln(q-1) - \frac{q}{2(q-1)} + \frac{1}{2}\ln(2k+1) \\
&\overset{(A.8)}{=} \frac{1}{2}\left(\psi\left(\frac{q+1}{2}\right) - \sum_{i=1}^{k-1}\frac{1}{\frac{q+1}{2}+k-i-1} - \psi\left(\frac{q+1}{2}\right)\right) - \frac{q}{2(q-1)} + \frac{1}{2}\ln\left(\frac{2k+1}{q-1}\right) \\
&\overset{①}{\leqslant} -\frac{1}{2}\sum_{i=1}^{k-1}\frac{2}{q-1+2k-2i} - \frac{1}{q-1} + \frac{1}{2}\ln\left(\frac{2k+1}{q-1}\right) \\
&= -\frac{1}{2}\left(\frac{2}{q-1} + \frac{2}{q+1} + \frac{2}{q+3} + \ldots + \frac{2}{q+2k-3}\right) + \frac{1}{2}\ln\left(\frac{2k+1}{q-1}\right) \\
&\overset{②}{<} -\frac{1}{2}\ln\left(\frac{q+2k-1}{q-1}\right) + \frac{1}{2}\ln\left(\frac{2k+1}{q-1}\right) \\
&\overset{③}{\leqslant} -\frac{1}{2}\ln\left(\frac{2k+1}{q-1}\right) + \frac{1}{2}\ln\left(\frac{2k+1}{q-1}\right) = 0,
\end{aligned}$$

where ① and ③ is because $q \ge 2$, ② is due to estimation of integral of $\frac{1}{x}$ by integral of $g(x) = \frac{1}{q-1+2i}$, $x \in [q-1+2i, q-1+2i+2]$, $i = \overline{0, 2k-1}$ which is no less than $f(x)$:

$$\frac{2}{q-1} + \frac{2}{q+1} + \frac{2}{q+3} + \ldots + \frac{2}{q+2k-3} > \int_{q-1}^{q+2k-1}\frac{1}{x}dx = \ln\left(\frac{q+2k-1}{q-1}\right).$$

So, we shown that $\frac{d(\ln(f_n(q)))}{dq} < 0$ for $q \ge 2$ arbitrary natural number n. Therefore for any fixed number n the function $f_n(q)$ decreases as q increase, which means that $f_n(q) \leqslant f_n(2) = 0$, i.e., (A.7) holds. From this and (A.4),(A.6) we obtain that $\forall q \ge 2$

$$\mathbb{E}[||e||_q^2] \overset{(A.4)}{\leqslant} (n\mathbb{E}[|e_2|^q])^{\frac{2}{q}} \overset{(A.6)(A.7)}{\leqslant} (q-1)n^{\frac{2}{q}-1}. \quad (A.10)$$

However, inequality (A.10) is useless when q is big (with respect to n). Consider left hand side of (A.10) as function of q and find its minimum for $q \ge 2$. Consider $h_n(q) = \ln(q-1) + \left(\frac{2}{q}-1\right)\ln n$ (it is logarithm of the right hand side of (A.10)). Derivative of $h(q)$ is

$$\frac{dh(q)}{dq} = \frac{1}{q-1} - \frac{2\ln n}{q^2}, \frac{1}{q-1} - \frac{2\ln n}{q^2} = 0, q^2 - 2q\ln n + 2\ln n = 0.$$

If $n \ge 8$, then the point where the function obtains its minimum on the set $[2,+\infty)$ is $q_0 = \ln n\left(1+\sqrt{1-\frac{2}{\ln n}}\right)$ (for the case $n \le 7$ it turns out that $q_0 = 2$; further without loss of generality we assume $n \ge 8$). Therefore for all $q > q_0$ it is more useful to use the following estimation:

$$\begin{aligned}
\mathbb{E}[||e||_q^2] &\overset{①}{<} \mathbb{E}[||e||_{q_0}^2] \overset{(A.10)}{\leqslant} (q_0-1)n^{\frac{2}{q_0}-1} \overset{②}{\leqslant} (2\ln n - 1)n^{\frac{2}{\ln n}-1} \\
&= (2\ln n - 1)e^2\frac{1}{n} \leqslant (16\ln n - 8)\frac{1}{n} \leqslant (16\ln n - 8)n^{\frac{2}{q}-1}, \quad (A.11)
\end{aligned}$$

where ① is due to $\|e\|_q < \|e\|_{q_0}$ for $q > q_0$, ② follows from $q_0 \le 2\ln n, q_0 \ge \ln n$. Putting estimations (A.10) and (A.11) together we obtain (A.1).

Now we are going to prove (A.2). Firstly, we want to estimate $\sqrt{\mathbb{E}[||e||_q^4]}$. Due to probabilistic Jensen's inequality ($q \ge 2$)

$$\begin{aligned}
\mathbb{E}[||e||_q^4] &= \mathbb{E}\left[\left(\left(\sum_{k=1}^n |e_k|^q\right)^2\right)^{\frac{2}{q}}\right] \leqslant \left(\mathbb{E}\left[\left(\sum_{k=1}^n |e_k|^q\right)^2\right]\right)^{\frac{2}{q}} \\
&\overset{①}{\leqslant} \left(\mathbb{E}\left[\left(n\sum_{k=1}^n |e_k|^{2q}\right)\right]\right)^{\frac{2}{q}} \overset{②}{=} \left(n^2\mathbb{E}[|e_2|^{2q}]\right)^{\frac{2}{q}} \\
&\overset{(A.6)(A.7)}{\leqslant} n^{\frac{4}{q}}\left(\left(\frac{2q-1}{n}\right)^{\frac{2q}{2}}\right)^{\frac{2}{q}} = (2q-1)^2 n^{\frac{4}{q}-2},
\end{aligned}$$

where ① is because $\left(\sum_{k=1}^n x_k\right)^2 \leqslant n\sum_{k=1}^n x_k^2$ for $x_1, x_2, \ldots, x_n \in \mathbb{R}$ and ② follows from that mathematical expectation is linear and components of the random vector e are identically distributed. From this we obtain

$$\sqrt{\mathbb{E}[||e||_q^4]} \leqslant (2q-1)n^{\frac{2}{q}-1}. \quad (A.12)$$

Consider the right hand side of the inequality (A.12) as a function of q and find its minimum for $q \ge 2$. Consider $h_n(q) = \ln(2q-1) + \left(\frac{2}{q}-1\right)\ln n$ (logarithm of the right hand side (A.12)). Derivative of $h(q)$ is

$$\frac{dh(q)}{dq} = \frac{2}{2q-1} - \frac{2\ln n}{q^2}, \frac{2}{2q-1} - \frac{2\ln n}{q^2} = 0, q^2 - 2q\ln n + \ln n = 0.$$

If $n \ge 3$, the point where the function obtains its minimum on the set $[2,+\infty)$ is $q_0 = \ln n\left(1+\sqrt{1-\frac{1}{\ln n}}\right)$ (for the case $n \le 2$ it turns out that $q_0 = 2$; further without loss of generality we assume that $n \ge 3$). Therefore for all $q > q_0$:

$$\begin{aligned}
\sqrt{\mathbb{E}[||e||_q^4]} &\overset{①}{<} \sqrt{\mathbb{E}[||e||_{q_0}^4]} \leqslant (A.12)(2q_0-1)n^{\frac{2}{q_0}-1} \\
&\overset{②}{\leqslant} (4\ln n - 1)n^{\frac{2}{\ln n}-1} = (4\ln n - 1)e^2\frac{1}{n} \\
&\leqslant (32\ln n - 8)\frac{1}{n} \leqslant (32\ln n - 8)n^{\frac{2}{q}-1}, \quad (A.13)
\end{aligned}$$

where ① is due to $\|e\|_q < \|e\|_{q_0}$ for $q > q_0$, ② follows from $q_0 \le 2\ln n, q_0 \ge \ln n$. Putting estimations (A.12) and (A.13) together we get inequality

$$\sqrt{\mathbb{E}[||e||_q^4]} \leqslant \min\{2q-1, 32\ln n - 8\}n^{\frac{2}{q}-1}. \quad (A.14)$$

Please cite this article as: P. Dvurechensky, E. Gorbunov and A. Gasnikov, An accelerated directional derivative method for smooth stochastic convex optimization, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2020.08.027

Now we are going to find $\mathbb{E}[\langle s, e\rangle^4]$, where $s \in \mathbb{R}^n$ is some vector. Let $S_n(r)$ be a surface area of n-dimensional Euclidean sphere with radius r and $d\sigma(e)$ be unnormalized uniform measure on n-dimensional Euclidean sphere. From this it follows that $S_n(r) = S_n(1)r^{n-1}$, $\frac{S_{n-1}(1)}{S_n(1)} = \frac{n-1}{n\sqrt{\pi}}\frac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+1}{2})}$. Besides, let φ be the angle between s and e. Then

$$\begin{aligned}\mathbb{E}[\langle s, e\rangle^4] &= \frac{1}{S_n(1)}\int_S \langle s, e\rangle^4 d\sigma(\varphi)\\ &= \frac{1}{S_n(1)}\int_0^{\pi} ||s||_2^4\cos^3\varphi S_{n-1}(\sin\varphi)d\varphi\\ &= ||s||_2^4\frac{S_{n-1}(1)}{S_n(1)}\int_0^{\pi}\cos^4\varphi\sin^{n-2}\varphi d\varphi\\ &= ||s||_2^4\cdot\frac{n-1}{n\sqrt{\pi}}\frac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+1}{2})}\int_0^{\pi}\cos^4\varphi\sin^{n-2}\varphi d\varphi. \end{aligned} \tag{A.15}$$

Compute the integral:

$$\begin{aligned}&\int_0^{\pi}\cos^4\varphi\sin^{n-2}\varphi d\varphi = 2\int_0^{\frac{\pi}{2}}\cos^4\varphi\sin^{n-2}\varphi d\varphi\\ &= /t = \sin^2\varphi/ = \int_0^{\frac{\pi}{2}} t^{\frac{n-3}{2}}(1-t)^{\frac{3}{2}}dt = B\left(\frac{n-1}{2}, \frac{5}{2}\right)\\ &= \frac{\Gamma(\frac{5}{2})\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n+4}{2})} = \frac{\frac{3}{2}\cdot\frac{1}{2}\Gamma(\frac{1}{2})\Gamma(\frac{n-1}{2})}{\frac{n+2}{2}\cdot\Gamma(\frac{n+2}{2})}\\ &= \frac{3}{n+2}\cdot\frac{\sqrt{\pi}\Gamma(\frac{n-1}{2})}{2\Gamma(\frac{n+2}{2})}.\end{aligned}$$

From this and (A.15) we obtain

$$\begin{aligned}\mathbb{E}[\langle s, e\rangle^4] &= ||s||_2^4\cdot\frac{n-1}{n\sqrt{\pi}}\frac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+1}{2})}\cdot\frac{3}{n+2}\cdot\frac{\sqrt{\pi}\Gamma(\frac{n-1}{2})}{2\Gamma(\frac{n+2}{2})}\\ &= ||s||_2^4\cdot\frac{3(n-1)}{2n(n+2)}\cdot\frac{\Gamma(\frac{n-1}{2})}{\frac{n-1}{2}\Gamma(\frac{n-1}{2})} = \frac{3||s||_2^4}{n(n+2)} \overset{①}{\leqslant} \frac{3||s||_2^4}{n^2}.\end{aligned} \tag{A.16}$$

To prove (A.2), it remains to use (A.14), (A.16) and Cauchy-Schwartz inequality $((\mathbb{E}[XY])^2 \leqslant \mathbb{E}[X^2]\cdot\mathbb{E}[Y^2])$:

$$\begin{aligned}&\mathbb{E}[\langle s, e\rangle^2||e||_q^2] \overset{①}{\leqslant} \sqrt{\mathbb{E}[\langle s, e\rangle^4]\cdot\mathbb{E}[||e||_q^4]}\\ &\leqslant \sqrt{3}||s||_2^2\min\{2q-1, 32\ln n - 8\}n^{\frac{2}{q}-2}.\end{aligned}$$

Appendix B. Technical Results

Lemma 12. *Let $a_0, \ldots, a_{N-1}, b, R_1, \ldots, R_{N-1}$ be non-negative numbers such that*

$$R_l \leqslant \sqrt{2}\cdot\sqrt{\left(\sum_{k=0}^{l-1}a_k + b\sum_{k=1}^{l-1}\alpha_{k+1}R_k\right)} \quad l = 1, \ldots, N, \tag{B.1}$$

where $\alpha_{k+1} = \frac{k+2}{96n^2\rho_n L_2}$ for all $k \in \mathbb{N}$. Then for $l = 1, \ldots, N$

$$\sum_{k=0}^{l-1}a_k + b\sum_{k=1}^{l-1}\alpha_{k+1}R_k \leqslant \left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\cdot\frac{l^2}{96n^2\rho_n L_2}\right)^2. \tag{B.2}$$

Proof. For $l = 1$ it is trivial inequality. Assume that (B.2) holds for some $l < N$ and prove it for $l+1$. From the induction assumption and (B.1) we obtain

$$R_l \leqslant \sqrt{2}\left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\cdot\frac{l^2}{96n^2\rho_n L_2}\right), \tag{B.3}$$

whence

$$\begin{aligned}&\sum_{k=0}^{l}a_k + b\sum_{k=1}^{l}\alpha_{k+1}R_k = \sum_{k=0}^{l-1}a_k + b\sum_{k=1}^{l-1}\alpha_{k+1}R_k + a_l + b\alpha_{l+1}R_l\\ &\overset{①}{\leqslant}\left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\cdot\frac{l^2}{96n^2\rho_n L_2}\right)^2\\ &\quad + a_l + \sqrt{2}b\alpha_{l+1}\left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\cdot\frac{l^2}{96n^2\rho_n L_2}\right)\\ &= \sum_{k=0}^{l}a_k + 2\sqrt{\sum_{k=0}^{l-1}a_k}\cdot\sqrt{2}b\frac{l^2}{96n^2\rho_n L_2} + 2b^2\frac{l^4}{(96n^2\rho_n L_2)^2}\\ &\quad + \sqrt{2}b\alpha_{l+1}\left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\cdot\frac{l^2}{96n^2\rho_n L_2}\right)\\ &= \sum_{k=0}^{l}a_k + 2\sqrt{\sum_{k=0}^{l-1}a_k}\cdot\sqrt{2}b\left(\frac{l^2}{96n^2\rho_n L_2} + \frac{\alpha_{l+1}}{2}\right)\\ &\quad + 2b^2\left(\frac{l^4}{(96n^2\rho_n L_2)^2} + \alpha_{l+1}\cdot\frac{l^2}{96n^2\rho_n L_2}\right)\\ &\overset{②}{\leqslant}\sum_{k=0}^{l}a_k + 2\sqrt{\sum_{k=0}^{l}a_k}\cdot\sqrt{2}b\frac{(l+1)^2}{96n^2\rho_n L_2} + 2b^2\frac{(l+1)^4}{(96n^2\rho_n L_2)^2}\\ &= \left(\sqrt{\sum_{k=0}^{l}a_k} + \sqrt{2}b\cdot\frac{(l+1)^2}{96n^2\rho_n L_2}\right)^2,\end{aligned}$$

where ① follows from the induction assumption and (B.3), ② is because $\sum_{k=0}^{l-1}a_k \leqslant \sum_{k=0}^{l}a_k$ and

$$\begin{aligned}&\frac{l^2}{96n^2\rho_n L_2} + \frac{\alpha_{l+1}}{2} = \frac{2l^2+l+2}{192n^2\rho_n L_2} \leqslant \frac{(l+1)^2}{96n^2\rho_n L_2}, \frac{l^4}{(96n^2\rho_n L_2)^2}\\ &+ \alpha_{l+1}\cdot\frac{l^2}{96n^2\rho_n L_2} \leqslant \frac{l^4+(l+2)l^2}{(96n^2\rho_n L_2)^2} \leqslant \frac{(l+1)^4}{(96n^2\rho_n L_2)^2}.\end{aligned}$$

□

Lemma 13. *Let $a_0, \ldots, a_{N-1}, b, R_1, \ldots, R_{N-1}$ be non-negative numbers such that*

$$R_l \leqslant \sqrt{2}\cdot\sqrt{\left(\sum_{k=0}^{l-1}a_k + b\alpha\sum_{k=1}^{l-1}R_k\right)} \quad l = 1, \ldots, N. \tag{B.4}$$

Then for $l = 1, \ldots, N$

$$\sum_{k=0}^{l-1}a_k + b\alpha\sum_{k=1}^{l-1}R_k \leqslant \left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\alpha l\right)^2. \tag{B.5}$$

Proof. For $l = 1$ it is trivial inequality. Assume that (B.5) holds for some $l < N$ and prove it for $l+1$. From the induction assumption and (B.4) we obtain

$$R_l \leqslant \sqrt{2}\left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\alpha l\right), \tag{B.6}$$

whence

$$\begin{aligned}&\sum_{k=0}^{l}a_k + b\alpha\sum_{k=1}^{l}R_k = \sum_{k=0}^{l-1}a_k + b\alpha\sum_{k=1}^{l-1}R_k + a_l + b\alpha R_l\\ &\overset{①}{\leqslant}\left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\alpha l\right)^2 + a_l + \sqrt{2}b\alpha\left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\alpha l\right)\end{aligned}$$

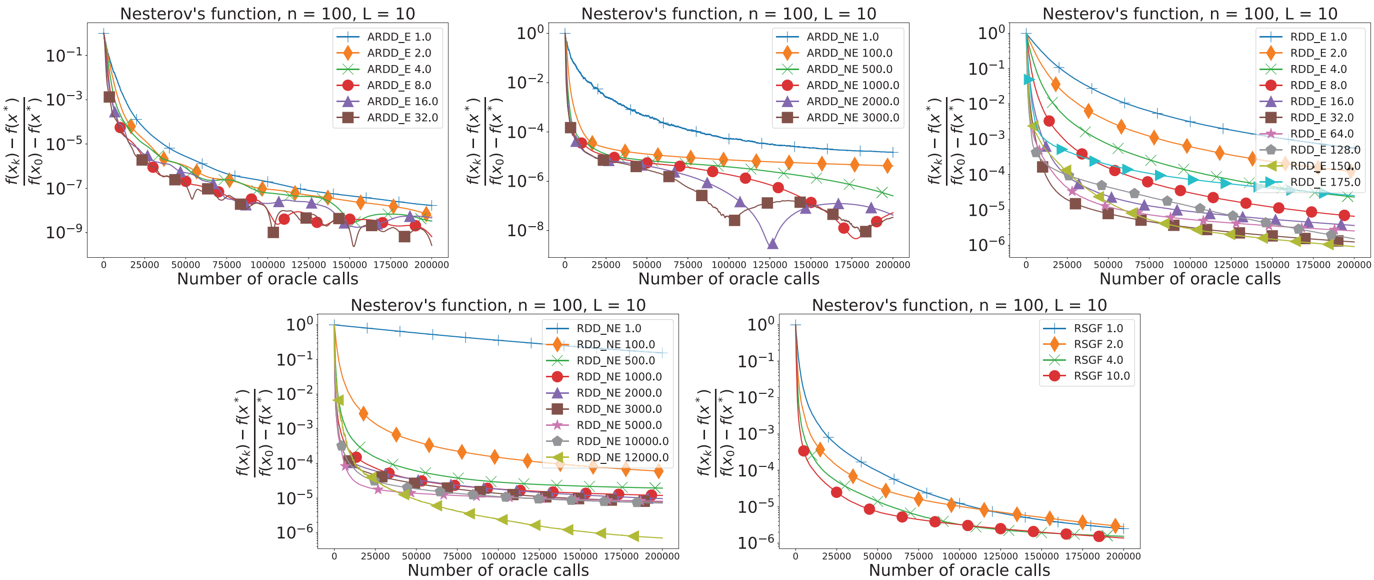

Fig. C.3. Stepsize tuning for ARDD, RDD and RSGF applied to minimize Nesterov's function (62). We use _E and _NE to define ℓ_2 and ℓ_1 proximal setups respectively (see (8) and (9) for the details). Numbers in labels in upper right corners denote different choices of γ that are used.

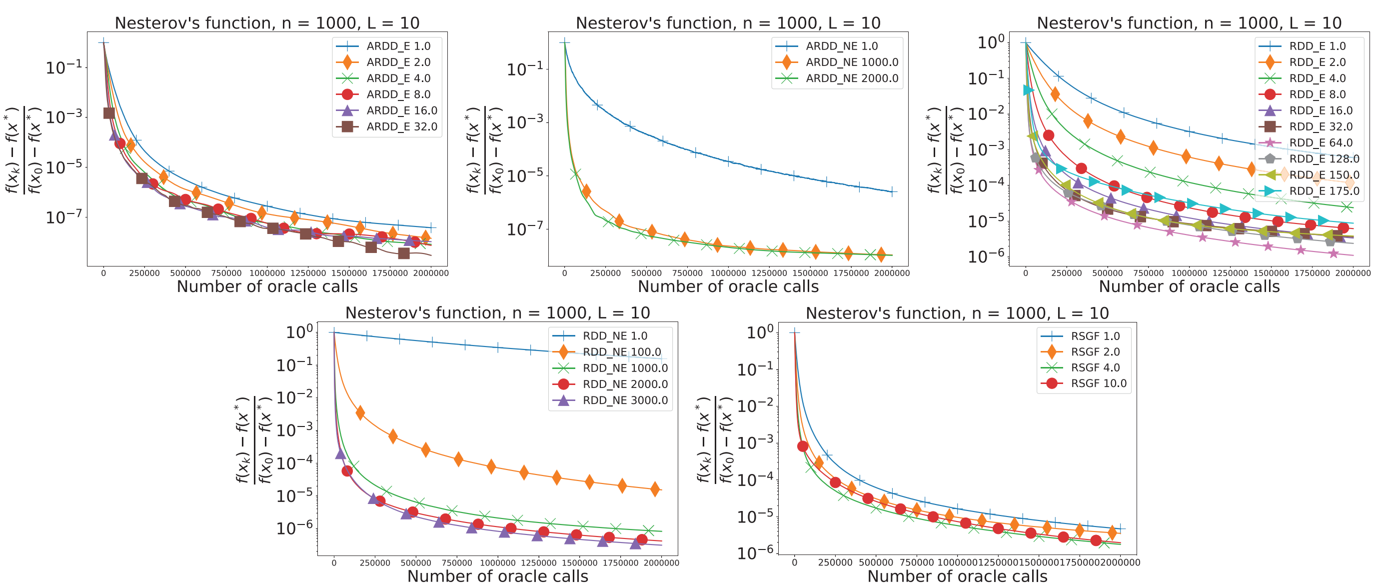

Fig. C.4. Stepsize tuning for ARDD, RDD and RSGF applied to minimize Nesterov's function (62). We use _E and _NE to define ℓ_2 and ℓ_1 proximal setups respectively (see (8) and (9) for the details). Number of oracle calls is divided by 10^7. Numbers in labels in upper right corners denote different choices of γ that are used.

$$
\begin{aligned}
&= \sum_{k=0}^{l} a_k + 2\sqrt{\sum_{k=0}^{l-1} a_k \cdot \sqrt{2}b\alpha l + 2b^2\alpha^2 l^2} + \sqrt{2}b\alpha\left(\sqrt{\sum_{k=0}^{l-1} a_k} + \sqrt{2}b\alpha l\right)\\
&= \sum_{k=0}^{l} a_k + 2\sqrt{\sum_{k=0}^{l-1} a_k \cdot \sqrt{2}b\alpha\left(l+\frac{1}{2}\right) + 2b^2\alpha^2\left(l^2+l\right)}\\
&\overset{②}{\leqslant} \sum_{k=0}^{l} a_k + 2\sqrt{\sum_{k=0}^{l} a_k \cdot \sqrt{2}b\alpha(l+1) + 2b^2\alpha^2(l+1)^2}\\
&= \left(\sqrt{\sum_{k=0}^{l} a_k} + \sqrt{2}b\alpha(l+1)\right)^2,
\end{aligned}
$$

where ① follows from the induction assumption and (B.6), ② is because $\sum_{k=0}^{l-1} a_k \leqslant \sum_{k=0}^{l} a_k$. □

Appendix C. Parameters tuning

In our analysis it is needed to choose $\alpha_{k+1} = \frac{k+2}{96n^2\rho_n L_2}$ for ARDD and $\alpha = \frac{1}{48n\rho_n L_2}$. However, one can tune these parameters in order to achieve better convergence rate in practice. In our experiments we choose $\alpha_{k+1} = \gamma \cdot \frac{k+2}{96n^2\rho_n L_2}$, $\alpha = \gamma \cdot \frac{1}{48n\rho_n L_2}$ and tune numerical factor γ. In Ghadimi and Lan (2013) authors prove convergence results for stepsize[6] $\alpha = \frac{1}{\sqrt{n+4}} \min\left\{\frac{1}{4L\sqrt{n+4}}, \frac{\tilde{D}}{\sigma\sqrt{N}}\right\}$ where $\tilde{D}$ is some numerical constant, therefore, in our experiments with RSGD we use stepsizes $\alpha = \gamma \cdot \frac{1}{\sqrt{n+4}} \min\left\{\frac{1}{4L\sqrt{n+4}}, \frac{1}{\sqrt{N}}\right\}$ where we also tune numerical factor γ.

[6] If $\sigma = 0$, then one should ignore the second term in the minimum.

Please cite this article as: P. Dvurechensky, E. Gorbunov and A. Gasnikov, An accelerated directional derivative method for smooth stochastic convex optimization, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2020.08.027

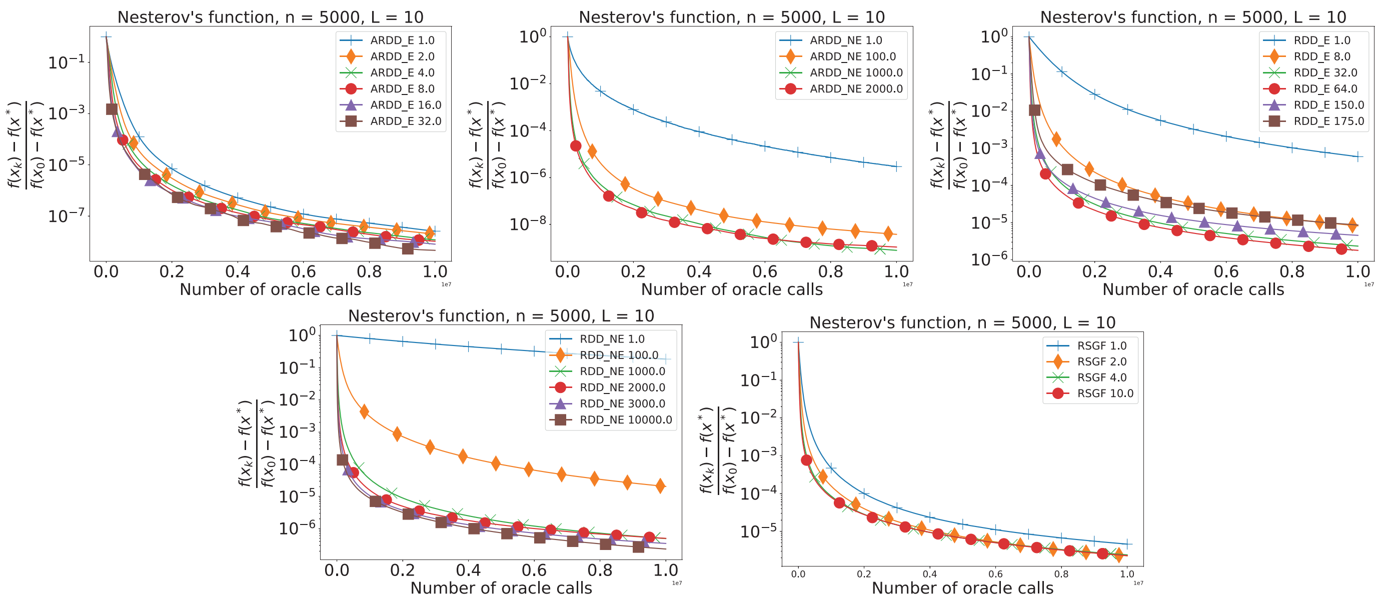

Fig. C.5. Stepsize tuning for ARDD, RDD and RSGF applied to solve least squares problem (63). We use _E and _NE to define ℓ_2 and ℓ_1 proximal setups respectively (see (8) and (9) for the details). For all methods batch size m equals 50. By oracle call we mean one computation of functional value of a summand. Number of oracle calls is divided by 10^8.

Table C.5
The optimal choices of γ for ARDD, RDD and RSGF applied to minimize Nesterov's function (62) for different dimension n.

	ARDD_E	ARDD_NE	RDD_E	RDD_NE	RSGF
$n = 100$	32	2000	32	12000	10
$n = 1000$	32	2000	64	3000	4
$n = 5000$	32	1000	64	3000	10

C.1. Nesterov's function

One can find our numerical results with tuning stepsizes for each method in Figs. C.3–C.4.

Our tests with Nesterov's function show that for this problem ARDD_E and RDD_ work better with $\gamma \in [32, 64]$ and RSGF shows the best performance with $\gamma \in [4, 10]$. Interestingly, ARDD and RDD with $p = 1$ require to choose γ significantly larger (of order $10^3 - 10^4$) than for Euclidean methods in order to get competitive or even better convergence rate. Moreover, ARDD_E, RDD_E and RSGF disconverge for $\gamma \geq 64$, 200, 20 respectively. So, our empirical observation is as follows: ARDD and RDD with non-Euclidean proximal setup are able to converge with significantly larger stepsizes than its Euclidean counterpart.

We summarize best options for γ that we use in the experiments presented in Section 6 in Table C.5.

C.2. Least squares problem

In addition to the tuning of γ in ARDD we also tried different options for L_2: instead of L_2 from (64) we tried $\beta \cdot \frac{\|A\|_F}{\sqrt{r}}$ with different β. We tried $\beta = 0.001, 0.01, 0.1, 1, 2, 5$ and 10, but the best results were obtained for $\beta = 0.01$. One can find our numerical results with tuning γ in Fig. C.5.

Besides $m = 50$ we tried different batch sizes. In general, the behaviour of the considered methods was similar after proper parameters tuning.

References

Agarwal, A., Dekel, O., & Xiao, L. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the Colt 2010 - the 23rd conference on learning theory*.

Allen-Zhu, Z. (2017). Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th annual acm sigact symposium on theory of computing*. In *STOC 2017* (pp. 1200–1205). New York, NY, USA: ACM. https://doi.org/10.1145/3055399.3055448. ArXiv:1603.05953

Allen-Zhu, Z., Qu, Z., Richtarik, P., & Yuan, Y. (2016). Even faster accelerated coordinate descent using non-uniform sampling. In M. F. Balcan, & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning*. In *Proceedings of Machine Learning Research: 48* (pp. 1110–1119). New York, New York, USA: PMLR. First appeared in arXiv:1512.09103

Anikin, A. S., Gasnikov, A. V., Dvurechensky, P. E., Tyurin, A. I., & Chernov, A. V. (2017). Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints. *Computational Mathematics and Mathematical Physics, 57*(8), 1262–1276.

Barrett, J. W., & Prigozhin, L. (2010). A quasi-variational inequality problem in superconductivity. *Mathematical Models and Methods in Applied Sciences, 20*(5), 679–706. https://doi.org/10.1142/S0218202510004404.

Barrett, J. W., & Prigozhin, L. (2014). Lakes and rivers in the landscape: A quasi-variational inequality approach. *Interfaces and Free Boundaries, 16*(2), 269–296. https://doi.org/10.4171/IFB/320.

Bayandina, A., Dvurechensky, P., Gasnikov, A., Stonyakin, F., & Titov, A. (2018a). Mirror descent and convex optimization problems with non-smooth inequality constraints. In P. Giselsson, & A. Rantzer (Eds.), *Large-scale and distributed optimization* (pp. 181–215). Springer International Publishing. ArXiv:1710.06612

Bayandina, A., Gasnikov, A., & Lagunovskaya, A. (2018b). Gradient-free two-points optimal method for non smooth stochastic convex optimization problem with additional small noise. *Automation and Remote Control, 79*(7). ArXiv:1701.03821

Ben-Tal, A., & Nemirovski, A. (2015). *Lectures on Modern Convex Optimization (Lecture Notes)*. Personal Web-pzage of A. Nemirovski.

Berahas, A. S., Byrd, R. H., & Nocedal, J. (2019a). Derivative-free optimization of noisy functions via quasi-Newton methods. *SIAM Journal on Optimization, 29*(2), 965–993. https://doi.org/10.1137/18M1177718.

Berahas, A. S., Cao, L., Choromanski, K., & Scheinberg, K. (2019b). A theoretical and empirical comparison of gradient approximations in derivative-free optimization. arXiv:1905.01332,.

Berahas, A. S., Cao, L., & Scheinberg, K. (2019c). Global convergence rate analysis of a generic line search algorithm with noise. arXiv:1910.04055,.

Beznosikov, A., Gorbunov, E., & Gasnikov, A. (2020a). *Derivative-free method for composite optimization with applications to decentralized distributed optimization*. IFAC-PapersOnLine. Accepted, arXiv:1911.10645

Beznosikov, A., Sadiev, A., & Gasnikov, A. (2020b). Gradient-free methods for saddle-point problem. In A. Kononov, & et al. (Eds.), *Mathematical optimization theory and operations research 2020*. Cham: Springer International Publishing. Accepted, arXiv:2005.05913

Bogolubsky, L., Dvurechensky, P., Gasnikov, A., Gusev, G., Nesterov, Y., Raigorodskii, A. M., … Zhukovskii, M. (2016). Learning supervised pagerank with gradient-based and gradient-free optimization methods. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29* (pp. 4914–4922). Curran Associates, Inc.. ArXiv:1603.00717

Bollapragada, R., & Wild, S. M. (2019). Adaptive sampling quasi-Newton methods for derivative-free stochastic optimization. arXiv:1910.13516,.

Brent, R. (1973). Algorithms for Minimization Without Derivatives. *Dover Books on Mathematics*. Dover Publications.

Cartis, C., & Scheinberg, K. (2018). Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming, 169*(2), 337–375. https://doi.org/10.1007/s10107-017-1137-4.

Cauchy, A. (1847). Méthode générale pour la résolution des systémes d'équations simultanées. *Comptes rendus hebdomadaires des séances de l'Académie des sciences, 55*, 536–538.

Cesa-bianchi, N., Conconi, A., & Gentile, C. (2002). On the generalization ability of on-line learning algorithms. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 359–366). MIT Press.

Chen, Y., Orvieto, A., & Lucchi, A. (2020). An accelerated DFO algorithm for finite-sum convex functions. In *Proceedings of the 37th international conference on machine learning*. In *Proceedings of Machine Learning Research*. PMLR. (accepted), arXiv:2007.03311

Chernov, A., Dvurechensky, P., & Gasnikov, A. (2016). *Fast primal-dual gradient method for strongly convex minimization problems with linear constraints*. In Y. Kochetov, M. Khachay, V. Beresnev, E. Nurminski, & P. Pardalos (Eds.) (pp. 391–403)). Cham: Springer International Publishing.

Conn, A., Scheinberg, K., & Vicente, L. (2009). *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718768.

Dang, C. D., & Lan, G. (2015). Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization, 25*(2), 856–881. https://doi.org/10.1137/130936361.

Devolder, O. (2011). Stochastic first order methods in smooth convex optimization. CORE Discussion Paper 2011/70,.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., & Wibisono, A. (2015). Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Trans. Information Theory, 61*(5), 2788–2806. https://doi.org/10.1109/TIT.2015.2409256. ArXiv:1312.2139

Dvinskikh, D., Gorbunov, E., Gasnikov, A., Dvurechensky, P., & Uribe, C. A. (2019). On primal and dual approaches for distributed stochastic convex optimization over networks. In *Proceedings of the IEEE 58th conference on decision and control (cdc)* (pp. 7435–7440). https://doi.org/10.1109/CDC40024.2019.9029798. ArXiv:1903.09844

Dvinskikh, D., Ogaltsov, A., Gasnikov, A., Dvurechensky, P., & Spokoiny, V. (2020). On the line-search gradient methods for stochastic optimization. IFAC-PapersOnLine,. Accepted, arXiv:1911.08380.

Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C. A., & Nedić, A. (2018a). Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31*. In *NeurIPS 2018* (pp. 10783–10793). Curran Associates, Inc.. ArXiv:1806.03915

Dvurechensky, P., & Gasnikov, A. (2016). Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications, 171*(1), 121–145. https://doi.org/10.1007/s10957-016-0999-6.

Dvurechensky, P., Gasnikov, A., Gasnikova, E., Matsievsky, S., Rodomanov, A., & Usik, I. (2016). Primal-dual method for searching equilibrium in hierarchical congestion population games. In *Proceedings of the 9th international conference on discrete optimization and operations research and scientific school (door 2016) Vladivostok, Russia, September 19 - 23, 2016* (pp. 584–595). ArXiv:1606.08988

Dvurechensky, P., Gasnikov, A., & Tiurin, A. (2017). Randomized similar triangles method: A unifying framework for accelerated randomized optimization methods (coordinate descent, directional search, derivative-free method). arXiv:1707.08486,.

Dvurechensky, P. E., Gasnikov, A. V., & Lagunovskaya, A. A. (2018b). Parallel algorithms and probability of large deviation for stochastic convex optimization problems. *Numerical Analysis and Applications, 11*(1), 33–37. ArXiv:1701.01830

Fabian, V. (1967). Stochastic approximation of minima with improved asymptotic speed. *Annals of Mathematical Statistics, 38*(1), 191–200. https://doi.org/10.1214/aoms/1177699070.

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

Gasnikov, A., Dvurechensky, P., & Usmanova, I. (2016a). On accelerated randomized methods. *Proceedings of Moscow Institute of Physics and Technology, 8*(2), 67–100. In Russian, first appeared in arXiv:1508.02182

Gasnikov, A. V., & Dvurechensky, P. E. (2016). Stochastic intermediate gradient method for convex optimization problems. *Doklady Mathematics, 93*(2), 148–151.

Gasnikov, A. V., Dvurechensky, P. E., Zhukovskii, M. E., Kim, S. V., Plaunov, S. S., Smirnov, D. A., & Noskov, F. A. (2018). About the power law of the pagerank vector component distribution. Part 2. The Buckley–Osthus model, verification of the power law for this model, and setup of real search engines. *Numerical Analysis and Applications, 11*(1), 16–32.

Gasnikov, A. V., Gasnikova, E. V., Dvurechensky, P. E., Mohammed, A. A. M., & Chernousova, E. O. (2017a). About the power law of the pagerank vector component distribution. Part 1. Numerical methods for finding the pagerank vector. *Numerical Analysis and Applications, 10*(4), 299–312.

Gasnikov, A. V., Krymova, E. A., Lagunovskaya, A. A., Usmanova, I. N., & Fedorenko, F. A. (2017b). 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. https://doi.org/10.1134/S0005117917020035. ArXiv:1509.01679

Gasnikov, A. V., Lagunovskaya, A. A., Usmanova, I. N., & Fedorenko, F. A. (2016b). Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control, 77*(11), 2018–2034. https://doi.org/10.1134/S0005117916110114. ArXiv:1412.3890

Ghadimi, S., & Lan, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization, 23*(4), 2341–2368. https://doi.org/10.1137/120880811. ArXiv:1309.5549

Ghadimi, S., Lan, G., & Zhang, H. (2016). Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming, 155*(1), 267–305. https://doi.org/10.1007/s10107-014-0846-1. ArXiv:1308.6594

Gorbunov, E., Dvinskikh, D., & Gasnikov, A. (2019). Optimal decentralized distributed algorithms for stochastic convex optimization. arXiv preprint arXiv:1911.07363,.

Gorbunov, E., Dvurechensky, P., & Gasnikov, A. (2018). An accelerated method for derivative-free smooth stochastic convex optimization. arXiv:1802.09022,.

Hu, X., L. A. , P., Gyrgy, A., & Szepesvari, C. (2016). (bandit) convex optimization with biased noisy gradient oracles. In A. Gretton, & C. C. Robert (Eds.), *Proceedings of the 19th international conference on artificial intelligence and statistics*. In *Proceedings of Machine Learning Research: 51* (pp. 819–828). Cadiz, Spain: PMLR.

Juditsky, A., & Nesterov, Y. (2014). Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems, 4*(1), 44–80. https://doi.org/10.1287/10-SSY010.

Kamzolov, D., Dvurechensky, P., & Gasnikov, A. V. (2020). Universal intermediate gradient method for convex problems with inexact oracle. *Optimization Methods and Software, 0*(0), 1–28. https://doi.org/10.1080/10556788.2019.1711079. ArXiv:1712.06036

Kim, K., Nesterov, Y., Skokov, V., & Cherkasskii, B. (1984). Effektivnii algoritm vychisleniya proisvodnyh i ekstremalnye zadachi (efficient algorithm for calculation of derivatives and extreme problems). *Ekonomika i matematicheskie metody, 20*(2), 309–318.

Lan, G. (2012). An optimal method for stochastic composite optimization. *Mathematical Programming, 133*(1), 365–397. Firs appeared in June 2008

Larson, J., Menickelly, M., & Wild, S. M. (2019). Derivative-free optimization methods. *Acta Numerica, 28*, 287–404. https://doi.org/10.1017/S0962492919000060.

Lee, Y. T., & Sidford, A. (2013). Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *Proceedings of the IEEE 54th annual symposium on foundations of computer science*. In *FOCS '13* (pp. 147–156). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/FOCS.2013.24. First appeared in arXiv:1305.1922

Lin, Q., Lu, Z., & Xiao, L. (2014). An accelerated proximal coordinate gradient method. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 3059–3067). Curran Associates, Inc.. First appeared in arXiv:1407.1296

Mordukhovich, B. S., & Outrata, J. V. (2007). Coderivative analysis of quasi variational inequalities with applications to stability and optimization. *SIAM Journal on Optimization, 18*(2), 389–412. https://doi.org/10.1137/060665609.

Nemirovsky, A., & Yudin, D. (1983). *Problem complexity and method efficiency in optimization*. J. Wiley & Sons, New York.

Nesterov, Y. (2004). *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, Massachusetts.

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

Nesterov, Y. (2015). Universal gradient methods for convex optimization problems. *Mathematical Programming, 152*(1), 381–404. https://doi.org/10.1007/s10107-014-0790-0.

Nesterov, Y., Gasnikov, A., Guminov, S., & Dvurechensky, P. (2020). Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, 1–28. https://doi.org/10.1080/10556788.2020.1731747. ArXiv:1809.05895

Nesterov, Y., & Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Found. Comput. Math., 17*(2), 527–566. https://doi.org/10.1007/s10208-015-9296-2. First appeared in 2011 as CORE discussion paper 2011/16

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

Powell, W. B. (2019). A unified framework for stochastic optimization. *European Journal of Operational Research, 275*(3), 795– 821. https://doi.org/10.1016/j.ejor.2018.07.014.

Prigozhin, L. (1996). Variational model of sandpile growth. *European Journal of Applied Mathematics, 7*(3), 225–235. https://doi.org/10.1017/S0956792500002321.

Rosenbrock, H. H. (1960). An automatic method for finding the greatest or least value of a function. *The Computer Journal, 3*(3), 175–184. https://doi.org/10.1093/comjnl/3.3.175.

Shalev-Shwartz, S., & Zhang, T. (2014). Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In E. P. Xing, & T. Jebara (Eds.), *Proceedings of the 31st international conference on machine learning*. In *Proceed-*

ings of Machine Learning Research: 32 (pp. 64–72). Bejing, China: PMLR. First appeared in arXiv:1309.2375

Shamir, O. (2017). An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research, 18*, 52:1–52:11. First appeared in arXiv:1507.08752

Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization* (1st). New York, NY, USA: John Wiley & Sons, Inc..

Stonyakin, F., Tyurin, A., Gasnikov, A., Dvurechensky, P., Agafonov, A., Dvinskikh, D., Pasechnyuk, D., Artamonov, S., & Piskunova, V. (2020). Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model. arXiv:2001.09013,.

Stonyakin, F. S., Dvinskikh, D., Dvurechensky, P., Kroshnin, A., Kuznetsova, O., Agafonov, A., … Artamonov, S. (2019). Gradient methods for problems with inexact model of the objective. In M. Khachay, Y. Kochetov, & P. Pardalos (Eds.), *Mathematical optimization theory and operations research* (pp. 97–114). Cham: Springer International Publishing. ArXiv:1902.09001

Vorontsova, E. A., Gasnikov, A. V., Gorbunov, E. A., & Dvurechenskii, P. E. (2019). Accelerated gradient-free optimization methods with a non-euclidean proximal operator. *Automation and Remote Control, 80*(8), 1487–1501.

Wengert, R. E. (1964). A simple automatic derivative evaluation program. *Communications of the ACM, 7*(8), 463–464. https://doi.org/10.1145/355586.364791.