**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2022 Т. 14 № 2 С. 321–334**
DOI: 10.20537/2076-7633-2022-14-2-321-334

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 519.853.62

Градиентный метод с неточным оракулом для задач композитной невыпуклой оптимизации

П. Е. Двуреченский[1,2]

[1]Институт прикладного анализа и стохастики им. Вейерштрасса,
Германия, 10117, г. Берлин, Моренштрассе, д. 39
[2]Московский физико-технический институт,
Россия, 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9

E-mail: pavel.dvurechensky@wias-berlin.de

Получено 11.02.2022.
Принято к публикации 13.02.2022.

В этой статье мы предлагаем новый метод первого порядка для композитных невыпуклых задач минимизации с простыми ограничениями и неточным оракулом. Целевая функция задается как сумма «сложной», возможно, невыпуклой части с неточным оракулом и «простой» выпуклой части. Мы обобщаем понятие неточного оракула для выпуклых функций на случай невыпуклых функций. Неформально говоря, неточность оракула означает, что для «сложной» части в любой точке можно приближенно вычислить значение функции и построить квадратичную функцию, которая приближенно ограничивает эту функцию сверху. Рассматривается два возможных типа ошибки: контролируемая, которая может быть сделана сколь угодно маленькой, например, за счет решения вспомогательной задачи, и неконтролируемая. Примерами такой неточности являются: гладкие невыпуклые функции с неточным и непрерывным по Гёльдеру градиентом, функции, заданные вспомогательной равномерно вогнутой задачей максимизации, которая может быть решена лишь приближенно. Для введенного класса задач мы предлагаем метод типа проекции градиента / зеркального спуска, который позволяет использовать различные прокс-функции для задания неевклидовой проекции на допустимое множество и более гибкой адаптации к геометрии допустимого множества; адаптивно выбирает контролируемую ошибку оракула и ошибку неевклидового проектирования; допускает неточное проксимальное отображение с двумя типами ошибки: контролируемой и неконтролируемой. Мы доказываем скорость сходимости нашего метода в терминах нормы обобщенного градиентного отображения и показываем, что в случае неточного непрерывного по Гёльдеру градиента наш метод является универсальным по отношению к параметру и константе Гёльдера. Это означает, что методу не нужно знание этих параметров для работы. При этом полученная оценка сложности является равномерно наилучшей при всех параметрах Гёльдера. Наконец, в частном случае показано, что малое значение нормы обобщенного градиентного отображения в точке означает, что в этой точке приближенно выполняется необходимое условие локального минимума.

Ключевые слова: невыпуклая оптимизация, композитная оптимизация, неточный оракул, непрерывный по Гёльдеру градиент, универсальный градиентный метод

Исследование выполнено за счет гранта Российского научного фонда (проект № 21-71-30005).

COMPUTER RESEARCH AND MODELING
2022 VOL. 14 NO. 2 P. 321–334
DOI: 10.20537/2076-7633-2022-14-2-321-334

UDC: 519.853.62

A gradient method with inexact oracle for composite nonconvex optimization

P. E. Dvurechensky[1,2]

[1]Weierstrass Institute for Applied Analysis and Stochastics,
39 Mohrenstraße, Berlin, 10117, Germany
[2]Moscow Institute of Physics and Technology,
9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia

E-mail: pavel.dvurechensky@wias-berlin.de

Received 11.02.2022.
Accepted for publication 13.02.2022.

In this paper, we develop a new first-order method for composite nonconvex minimization problems with simple constraints and inexact oracle. The objective function is given as a sum of «hard», possibly nonconvex part, and «simple» convex part. Informally speaking, oracle inexactness means that, for the «hard» part, at any point we can approximately calculate the value of the function and construct a quadratic function, which approximately bounds this function from above. We give several examples of such inexactness: smooth nonconvex functions with inexact Hölder-continuous gradient, functions given by the auxiliary uniformly concave maximization problem, which can be solved only approximately. For the introduced class of problems, we propose a gradient-type method, which allows one to use a different proximal setup to adapt to the geometry of the feasible set, adaptively chooses controlled oracle error, allows for inexact proximal mapping. We provide a convergence rate for our method in terms of the norm of generalized gradient mapping and show that, in the case of an inexact Hölder-continuous gradient, our method is universal with respect to Hölder parameters of the problem. Finally, in a particular case, we show that the small value of the norm of generalized gradient mapping at a point means that a necessary condition of local minimum approximately holds at that point.

Keywords: nonconvex optimization, composite optimization, inexact oracle, Hölder-continuous gradient, universal gradient methods

Citation: *Computer Research and Modeling*, 2022, vol. 14, no. 2, pp. 321–334.

This research was funded by the Russian Science Foundation (project 21-71-30005).

Introduction

In this paper, we introduce a new first-order method for nonconvex composite optimization problems with inexact oracle. Namely, our problem of interest is as follows:

$$\min_{x\in X\subseteq\mathcal{E}}\{\psi(x) := f(x) + h(x)\}, \tag{1}$$

where X is a closed convex set, $h(x)$ is a simple convex function, e. g. $\|x\|_1$. We assume that $f(x)$ is a general function endowed with an inexact first-order oracle, which is defined below (see Definition 1). Informally speaking, at any point we can approximately calculate the value of the function and construct a quadratic function, which approximately bounds our $f(x)$ from above. An example of a problem with this kind of inexactness is given in [Bogolubsky et al., 2016], where the authors study a learning problem for the parametric PageRank model.

First-order methods have been widely developed since the earliest years of optimization theory, see, e. g., [Polyak, 1963]. The recent renaissance in their development started more than ten years ago and was mostly motivated by fast growing problem sizes in applications such as Machine Learning, Data Analysis, Telecommunications. For many years, researchers have mostly considered convex optimization problems since they have good structure and allow one to estimate the rate of convergence for proposed algorithms. Recently, nonconvex problems started to attract fast growing attention, as they appear often in Machine Learning, especially in Deep Learning. Thus, high standards of research on algorithms for convex optimization started to influence nonconvex optimization. Namely, it has become very important for newly developed methods to obtain a rate of convergence with respect to some criterion. Usually, this criterion is the norm of gradient mapping, which is a generalization of gradient for constrained problems, see, e. g. [Nesterov, 2004].

Already in [Polyak, 1987], the author analyzed how different types of inexactness in gradient values influence the gradient method for unconstrained smooth convex problems. At the moment, the theory for convex optimization algorithms with inexact oracle is well-developed in a series of papers [d'Aspremont, 2008; Devolder, Glineur, Nesterov, 2014; Dvurechensky, Gasnikov, 2016]. In [d'Aspremont, 2008], it was proposed to calculate inexactly the gradient of the objective function and to extend the Fast Gradient Method of [Nesterov, 2005] to be able to use inexact oracle information. In [Devolder, Glineur, Nesterov, 2014], a general concept of inexact oracle is introduced for convex problems, and Primal, Dual and Fast gradient methods are analyzed. In [Dvurechensky, Gasnikov, 2016], the authors develop the Stochastic Intermediate Gradient Method for problems with stochastic inexact oracle, which provides good flexibility for solving convex and strongly convex problems with both deterministic and stochastic inexactness.

The theory for nonconvex smooth, nonsmooth and stochastic problems is well developed in [Ghadimi, Lan, 2016; Ghadimi, Lan, Zhang, 2016]. In [Ghadimi, Lan, 2016], problems of the form (1), where $X \equiv \mathbb{R}^n$ and $f(x)$ is a smooth nonconvex function, are considered in the case where the gradient of $f(x)$ is exactly available, as well as when it is available through stochastic approximation. Later, in [Ghadimi, Lan, Zhang, 2016] the authors generalized these methods for constrained problems of the form (1) in both deterministic and stochastic settings.

Nevertheless, it seems to us that gradient methods for nonconvex optimization problems with deterministic inexact oracle lack sufficient development. The goal of this paper is to fill this gap.

It turns out that smooth minimization with inexact oracle is closely connected with minimization of functions with a Hölder-continuous gradient. We say that a function $f(x)$ has a Hölder-continuous gradient on X iff there exist $\nu \in [0, 1]$ and $L_\nu \geqslant 0$ s.t.

$$\|\nabla f(x) - \nabla f(y)\|_{\mathcal{E},*} \leqslant L_\nu\|x - y\|_{\mathcal{E}}^{\nu}, \quad x, y \in X.$$

In [Devolder, Glineur, Nesterov, 2014] it was shown that a convex problem with a Hölder-continuous subgradient can be considered as a smooth problem with deterministic inexact oracle. Later, universal gradient methods for convex problems with a Hölder-continuous subgradient were proposed in [Nesterov, 2015]. These algorithms do not require knowledge of Hölder parameter ν and Hölder constant L_ν. Thus, they are universal with respect to these parameters. [Ghadimi, Lan, Zhang, 2015] proposed methods for nonconvex problems of the form (1), where $f(x)$ has a Hölder-continuous gradient. These methods rely on Euclidean norm and are good when the Euclidean projection onto the set X is simple.

Our contribution in this paper is as follows.

1. We generalize for the nonconvex case the definition of inexact oracle in [Devolder, Glineur, Nesterov, 2014] and provide several examples, where such inexactness can arise. We consider two types of errors — controlled errors, which can be made as small as desired, and uncontrolled errors, which can only be estimated.

2. We introduce a new gradient method for problem (1) and prove a theorem (see Theorem 1) on its rate of convergence in terms of the norm of generalized gradient mapping. Our method is adaptive to the controlled oracle error, is capable to work with inexact proximal mapping, and has flexibility of choice of proximal setup, based on the geometry of set X.

3. We show that, in the case of problems with an inexact Hölder-continuous gradient, our method is universal, that is, it does not require to know in advance a Hölder parameter ν and Hölder constant L_ν for the function $f(x)$, but provides the best-known convergence rate uniformly in Hölder parameter ν.

Thus, we provide a universal algorithm for nonconvex Hölder-smooth composite optimization problems with deterministic inexact oracle.

The rest of the paper is organized as follows. In Section 1, we define the deterministic inexact oracle for nonconvex problems and provide several examples. In Section 2, we describe our algorithm, and prove the convergence theorem. Also, we provide two corollaries for particular cases of smooth functions and Hölder-smooth functions. Note that the latter case includes the former one. Finally, we provide some explanations about how convergence of the norm of generalized gradient mapping to zero leads to a good approximation for a point, where a necessary optimality condition for Problem (1) holds. Note that we use different reasoning from what can be found in the literature.

Notation. Let $\mathcal{E}$ be a finite-dimensional real vector space and $\mathcal{E}^*$ be its dual. We denote the value of the linear function $g \in \mathcal{E}^*$ at $x \in \mathcal{E}$ by $\langle g, x \rangle$. Let $\|\cdot\|_{\mathcal{E}}$ be some norm on $\mathcal{E}$, $\|\cdot\|_{\mathcal{E},*}$ be its dual.

1. Inexact Oracle

In this section, we define the inexact oracle and describe several examples where it naturally arises.

Definition 1. We say that a function $f(x)$ is equipped with an *inexact first-order oracle* on a set X if there exists $\delta_u > 0$ and at any point $x \in X$ for any number $\delta_c > 0$ there exists a constant $L(\delta_c) \in (0, +\infty)$ and one can calculate $\widetilde{f}(x, \delta_c, \delta_u) \in \mathbb{R}$ and $\widetilde{g}(x, \delta_c, \delta_u) \in \mathcal{E}^*$ satisfying

$$|f(x) - \widetilde{f}(x, \delta_c, \delta_u)| \leqslant \delta_c + \delta_u, \tag{2}$$

$$f(y) - (\widetilde{f}(x, \delta_c, \delta_u) - \langle \widetilde{g}(x, \delta_c, \delta_u), y - x \rangle) \leqslant \frac{L(\delta_c)}{2}\|x - y\|_{\mathcal{E}}^2 + \delta_c + \delta_u \quad \forall y \in X. \tag{3}$$

In this definition, δ_c represents the error of the oracle, which we can control and make as small as we would like to. On the opposite, δ_u represents the error, which we cannot control. The idea behind the definition is that at any point we can approximately calculate the value of the function and construct an upper quadratic bound.

Let us now consider several examples.

1.1. Smooth function with inexact oracle values

Let us assume that:

1. Function $f(x)$ is L-smooth on X, i. e. it is differentiable and, for all $x, y \in X$, $\|\nabla f(x) - \nabla f(y)\|_{\mathcal{E},*} \leqslant$ $\leqslant L\|x - y\|_{\mathcal{E}}$.
2. Set X to be bounded with $\max\limits_{x, y \in X} \|x - y\|_{\mathcal{E}} \leqslant D$.
3. There exist $\overline{\delta}_u^1, \overline{\delta}_u^2 > 0$ and at any point $x \in Q$, for any $\overline{\delta}_c^1, \overline{\delta}_c^2 > 0$, we can calculate approximations $\overline{f}(x)$ and $\overline{g}(x)$ s.t. $|\overline{f}(x) - f(x)| \leqslant \overline{\delta}_c^1 + \overline{\delta}_u^1$, $\|\overline{g}(x) - \nabla f(x)\|_{\mathcal{E},*} \leqslant \overline{\delta}_c^2 + \overline{\delta}_u^2$.

Then, using L-smoothness of $f(x)$, we obtain, for any $y \in X$,

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_{\mathcal{E}}^2 \leqslant \tag{4}$$

$$\leqslant \overline{f}(x) + \overline{\delta}_c^1 + \overline{\delta}_u^1 + \langle \nabla \overline{g}(x), y - x \rangle + \langle \nabla f(x) - \overline{g}(x), y - x \rangle + \frac{L}{2}\|x - y\|_{\mathcal{E}}^2 \leqslant \tag{5}$$

$$\leqslant \overline{f}(x) + \langle \nabla \overline{g}(x), y - x \rangle + \frac{L}{2}\|x - y\|_{\mathcal{E}}^2 + \overline{\delta}_c^1 + \overline{\delta}_u^1 + (\overline{\delta}_c^2 + \overline{\delta}_u^2)D. \tag{6}$$

Thus, $(\overline{f}(x), \overline{g}(x))$ is an inexact first-order oracle with $\delta_u = \overline{\delta}_u^1 + \overline{\delta}_u^2 D$, $\delta_c = \overline{\delta}_c^1 + \overline{\delta}_c^2 D$, and $L(\delta_c) \equiv L$.

1.2. Smooth function with a Hölder-continuous gradient

Assume that $f(x)$ is differentiable and its gradient is Hölder-continuous, i. e. for some $\nu \in [0, 1]$ and $L_\nu \geqslant 0$,

$$\|\nabla f(x) - \nabla f(y)\|_* \leqslant L_\nu \|x - y\|_{\mathcal{E}}^\nu \quad \forall x, y \in X. \tag{7}$$

Then

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_\nu}{1 + \nu}\|x - y\|_{\mathcal{E}}^{1+\nu} \quad \forall x, y \in X. \tag{8}$$

It can be shown, see [Nesterov, 2015], Lemma 2, that, for all $x \in X$ and any $\delta > 0$,

$$f(y) - (f(x) - \langle \nabla f(x), y - x \rangle) \leqslant \frac{L(\delta)}{2}\|x - y\|_{\mathcal{E}}^2 + \delta \quad \forall y \in X, \tag{9}$$

where

$$L(\delta) = \left(\frac{1 - \nu}{1 + \nu} \cdot \frac{2}{\delta}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}. \tag{10}$$

Thus, $(f(x), \nabla f(x))$ is an inexact first-order oracle with $\delta_u = 0$, $\delta_c = \delta$, and $L(\delta)$ given by (10).

Note that, if $(f(x), \nabla f(x))$ can only be calculated inexactly as in Subsection 1.1, their approximations will again be an inexact first-order oracle.

1.3. The function given by maximization subproblem

Assume that the function $f(x)\colon \mathcal{E} \to \mathbb{R}$ is defined by an auxiliary optimization problem

$$f(x) = \max_{u \in U \subseteq \mathcal{H}} \{\Psi(x, u) := -G(u) + \langle Au, x\rangle\}, \tag{11}$$

where $A\colon \mathcal{H} \to \mathcal{E}^*$ is a linear operator, and $G\colon \mathcal{H} \to R$ is a continuously differentiable uniformly convex function of degree $\rho \geqslant 2$ with parameter $\sigma_\rho \geqslant 0$. The latter means that

$$\langle \nabla G(u_1) - \nabla G(u_2), u_1 - u_2\rangle \geqslant \sigma_\rho \|u_1 - u_2\|_{\mathcal{H}}^{\rho} \quad \forall u_1, u_2 \in U, \tag{12}$$

where $\|\cdot\|_{\mathcal{H}}$ is some norm on $\mathcal{H}$. Note that $f(x)$ is differentiable and $\nabla f(x) = Au^*(x)$, where $u^*(x)$ is the optimal solution in (11) for fixed x.

Extending the proof in [Nesterov, 2015], we can prove the following.

Lemma 1. *If G is uniformly convex on X, then the gradient of f is Hölder-continuous with*

$$\nu = \frac{1}{\rho - 1}, \quad L_\nu = \frac{\|A\|_{\mathcal{H}\to\mathcal{E}^*}^{\frac{\rho}{\rho-1}}}{\sigma_\rho^{\frac{1}{\rho-1}}}, \tag{13}$$

where $\|A\|_{\mathcal{H}\to\mathcal{E}^*} = \max\{\|Au\|_{\mathcal{E},*}\colon \|u\|_{\mathcal{H}} = 1\}$.

Proof. From the optimality conditions in (11), we obtain

$$\langle A^T x_1 - \nabla G(u(x_1)), u(x_2) - u(x_1)\rangle \leqslant 0, \tag{14}$$

$$\langle A^T x_2 - \nabla G(u(x_2)), u(x_1) - u(x_2)\rangle \leqslant 0. \tag{15}$$

Adding these inequalities, we obtain, by definition of uniformly convex function,

$$\langle A^T(x_1 - x_2), u(x_1) - u(x_2)\rangle \geqslant \langle \nabla G(u(x_1)) - \nabla G(u(x_2)), u(x_1) - u(x_2)\rangle \overset{(12)}{\geqslant} \sigma_\rho \|u(x_1) - u(x_2)\|_{\mathcal{H}}^{\rho}. \tag{16}$$

On the other hand,

$$\|A(u(x_1) - u(x_2))\|_{\mathcal{E},*}^2 \leqslant \|A\|_{\mathcal{H}\to\mathcal{E}^*}^2 \|u(x_1) - u(x_2)\|_{\mathcal{H}}^2 \leqslant \tag{17}$$

$$\leqslant \|A\|_{\mathcal{H}\to\mathcal{E}^*}^2 \left(\frac{1}{\sigma_\rho}\langle A^T(x_1 - x_2), u(x_1) - u(x_2)\rangle\right)^{2/\rho} \leqslant \tag{18}$$

$$\leqslant \frac{\|A\|_{\mathcal{H}\to\mathcal{E}^*}^2}{\sigma_\rho^{2/\rho}} \|A(u(x_1) - u(x_2))\|_{\mathcal{E},*}^{2/\rho} \|x_1 - x_2\|_{\mathcal{E}}^{2/\rho}. \tag{19}$$

Thus,

$$\|A(u(x_1) - u(x_2))\|_{\mathcal{E},*}^{2-2/\rho} \leqslant \frac{\|A\|_{\mathcal{H}\to\mathcal{E}^*}^2}{\sigma_\rho^{2/\rho}} \|x_1 - x_2\|_{\mathcal{E}}^{2/\rho}, \tag{20}$$

which proves the lemma. □

Let us now consider a situation where the maximization problem in (11) can be solved only inexactly by some auxiliary numerical method. It is natural to assume that, for any $x \in X$ and any $\delta > 0$, we can calculate a point $u_x \in U$ s.t.

$$0 \leqslant f(x) - \Psi(x, u_x) = \Psi(x, u^*(x)) - \Psi(x, u_x) \leqslant \delta. \tag{21}$$

Since $\ln(t)$ is a concave function, for any $\rho \geqslant 2$ and t, $\tau \geqslant 0$, we have

$$\ln\left(\frac{1}{\rho}t^{\rho} + \frac{\rho-1}{\rho}\tau^{\frac{\rho}{\rho-1}}\right) \geqslant \frac{1}{\rho}\ln\left(t^{\rho}\right) + \frac{\rho-1}{\rho}\ln\left(\tau^{\frac{\rho}{\rho-1}}\right) = \ln(t\tau). \tag{22}$$

Using this inequality with

$$t = \sigma_{\rho}^{1/\rho}\|u^*(x) - u_x\|_{\mathcal{H}}, \quad \tau = \frac{\|A\|_{\mathcal{H}\to\mathcal{E}^*}}{\sigma_{\rho}^{1/\rho}}\|y - x\|_{\mathcal{E}}, \tag{23}$$

we obtain, for any $y \in X$,

$$\langle A(u^*(x) - u_x),\, y - x\rangle \leqslant \|A\|_{\mathcal{H}\to\mathcal{E}^*}\|u^*(x) - u_x\|_{\mathcal{H}}\|y - x\|_{\mathcal{E}} \leqslant \tag{24}$$

$$\leqslant \frac{\sigma_{\rho}}{\rho}\|u^*(x) - u_x\|_{\mathcal{H}}^{\rho} + \frac{\|A\|_{\mathcal{H}\to\mathcal{E}^*}^{\frac{\rho}{\rho-1}}}{\frac{\rho}{\rho-1}\sigma_{\rho}^{\frac{1}{\rho-1}}}\|y - x\|_{\mathcal{E}}^{\frac{\rho}{\rho-1}} = \tag{25}$$

$$= \frac{\sigma_{\rho}}{\rho}\|u^*(x) - u_x\|_{\mathcal{H}}^{\rho} + \frac{L_{\nu}}{1+\nu}\|y - x\|_{\mathcal{E}}^{1+\nu}, \tag{26}$$

where ν and L_{ν} are defined in (13). At the same time, since $\Psi(x, u)$ (11) is uniformly concave in the second argument, we have

$$\frac{\sigma_{\rho}}{\rho}\|u^*(x) - u_x\|_{\mathcal{H}}^{\rho} \leqslant \Psi(x, u^*(x)) - \Psi(x, u_x) \overset{(21)}{\leqslant} \delta. \tag{27}$$

Combining this inequality with the previous one, we obtain

$$\langle A(u^*(x) - u_x),\, y - x\rangle \leqslant \frac{L_{\nu}}{1+\nu}\|x - y\|_{\mathcal{E}}^{1+\nu} + \delta. \tag{28}$$

Since f has a Hölder-continuous gradient with parameters (13), using (8), we obtain

$$f(y) \leqslant f(x) + \langle\nabla f(x),\, y - x\rangle + \frac{L_{\nu}}{1+\nu}\|x - y\|_{\mathcal{E}}^{1+\nu} \overset{(21)}{\leqslant} \tag{29}$$

$$\overset{(21)}{\leqslant} \Psi(x, u_x) + \delta + \langle Au_x, y - x\rangle + \langle A(u^*(x) - u_x),\, y - x\rangle + \frac{2L_{\nu}}{1+\nu}\|x - y\|_{\mathcal{E}}^{1+\nu} \overset{(28)}{\leqslant} \tag{30}$$

$$\overset{(28)}{\leqslant} \Psi(x, u_x) + \langle Au_x,\, y - x\rangle + \frac{2L_{\nu}}{1+\nu}\|x - y\|_{\mathcal{E}}^{1+\nu} + 2\delta \overset{(8),(9),(10)}{\leqslant} \tag{31}$$

$$\overset{(8),(9),(10)}{\leqslant} \Psi(x, u_x) + \langle Au_x,\, y - x\rangle + \frac{2L(\delta)}{2}\|x - y\|_{\mathcal{E}}^{2} + 4\delta. \tag{32}$$

Thus, we have found that $(\Psi(x, u_x), Au_x)$ is an inexact first-order oracle with $\delta_u = 0$, $\delta_c = 4\delta$, and $L(\delta_c)$ given by (10) with $\delta = \frac{\delta_c}{4}$.

2. Adaptive gradient method for problems with inexact oracle

To construct our algorithm for problem (1), we introduce, as it usually done, proximal setup [Ben-Tal, Nemirovski, 2015]. We choose a *prox-function* $d(x)$ which is continuous, convex on X and

1) admits a continuous in $x \in X^0$ selection of subgradients $d'(x)$, where $x \in X^0 \subseteq X$ is the set of all x, where $d'(x)$ exists;

2) $d(x)$ is 1-strongly convex on X with respect to $\|\cdot\|_{\mathcal{E}}$, i. e., for any $x \in X^0$, $y \in X$ $d(y) - d(x) - \langle d'(x), y - x\rangle \geqslant \frac{1}{2}\|y - x\|_{\mathcal{E}}^2$.

We define also the corresponding *Bregman divergence* $V[z](x) = d(x) - d(z) - \langle d'(z), x - z\rangle$, $x \in X$, $z \in X^0$. Standard proximal setups, i. e. Euclidean, entropy, $\frac{\ell_1}{\ell_2}$, simplex, nuclear norm, spectahedron can be found in [Ben-Tal, Nemirovski, 2015]. We will use Bregman divergence in the so-called *composite prox-mapping*

$$\min_{x \in X} \left\{ \langle g, x\rangle + \frac{1}{\gamma} V[\overline{x}](x) + h(x) \right\}, \tag{33}$$

where $\gamma > 0$, $\overline{x} \in X^0$, $g \in \mathcal{E}^*$ are given. We allow this problem to be solved inexactly in the following sense.

Definition 2. Assume that we are given $\delta_{pu} > 0$, $\gamma > 0$, $\overline{x} \in X^0$, $g \in \mathcal{E}^*$. We call a point $\widetilde{x} = \widetilde{x}(\overline{x}, g, \gamma, \delta_{pc}, \delta_{pu}) \in X^0$ an *inexact composite prox-mapping* iff for any $\delta_{pc} > 0$ we can calculate $\widetilde{x}$ and there exists $p \in \partial h(\widetilde{x})$ s.t. it holds that

$$\left\langle g + \frac{1}{\gamma}[d'(\widetilde{x}) - d'(\overline{x})] + p, u - \widetilde{x} \right\rangle \geqslant -\delta_{pc} - \delta_{pu} \quad \forall u \in X. \tag{34}$$

We write

$$\widetilde{x} = \operatorname*{argmin}_{x \in X}{}^{\delta_{pc}+\delta_{pu}} \left\{ \langle g, x\rangle + \frac{1}{\gamma} V[\overline{x}](x) + h(x) \right\} \tag{35}$$

and define

$$g_X(\overline{x}, g, \gamma, \delta_{pc}, \delta_{pu}) := \frac{1}{\gamma}(\overline{x} - \widetilde{x}). \tag{36}$$

This is a generalization of inexact composite prox-mapping in [Ben-Tal, Nemirovski, 2015]. Note that, if $\widetilde{x}$ is an exact solution of (33), inequality (34) holds with $\delta_{pc} = \delta_{pu} = 0$ due to the first-order optimality condition. Similarly to Definition 1, δ_{pc} represents an error, which can be controlled and made as small as it is desired, δ_{pu} represents an error which cannot be controlled.

Our main scheme is Algorithm 1 below.

We will need the following simple extension of Lemma 1 in [Ghadimi, Lan, Zhang, 2016] to perform a theoretical analysis of our algorithm.

Lemma 2. *Let* $\widetilde{x} = \widetilde{x}(\overline{x}, g, \gamma, \delta_{pc}, \delta_{pu})$ *be an inexact composite prox-mapping and* $g_X(\overline{x}, g, \gamma, \delta_{pc}, \delta_{pu})$ *be defined in* (36)*. Then, for any* $\overline{x} \in X^0$, $g \in \mathcal{E}^*$ *and* $\gamma, \delta_{pc}, \delta_{pu} > 0$, *it holds that*

$$\gamma\langle g, g_X(\overline{x}, g, \gamma, \delta_{pc}, \delta_{pu})\rangle \geqslant \gamma \|g_X(\overline{x}, g, \gamma, \delta_{pc}, \delta_{pu})\|_{\mathcal{E}}^2 + (h(\widetilde{x}(\overline{x}, g, \gamma, \delta_{pc}, \delta_{pu})) - h(x)) - \delta_{pc} - \delta_{pu}. \tag{39}$$

Proof. Taking $u = \overline{x}$ in (34) and rearranging terms, we obtain, by convexity of $h(x)$ and strong convexity of $d(x)$,

$$\langle g, \overline{x} - \widetilde{x}\rangle \geqslant \frac{1}{\gamma}\langle d'(\widetilde{x}) - d'(\overline{x}), \widetilde{x} - \overline{x}\rangle + \langle p, \widetilde{x} - \overline{x}\rangle - \delta_{pc} - \delta_{pu} \geqslant \frac{1}{\gamma}\|\widetilde{x} - \overline{x}\|_{\mathcal{E}}^2 + (h(\widetilde{x}) - h(\overline{x})) - \delta_{pc} - \delta_{pu}. \tag{40}$$

Applying the definition (36), we finish the proof. □

Theorem 1. *Assume that* $f(x)$ *is equipped with an inexact first-order oracle in the sense of Definition* 1 *and for any constants* $c_1, c_2 > 0$ *there exists an integer* $i \geqslant 0$ *s.t.* $2^i c_1 \geqslant L\left(\frac{c_2}{c_1 2^i}\right)$*. Assume also that there exists a number* $\psi^* > -\infty$ *such that* $\psi(x) \geqslant \psi^*$ *for all* $x \in X$*. Then, after* N *iterations of Algorithm* 1*, it holds that*

$$\left\| M_K(x_K - x_{K+1}) \right\|_{\mathcal{E}}^2 \leqslant \left(\sum_{k=0}^{N-1} \frac{1}{2M_k} \right)^{-1} (\psi(x_0) - \psi^* + N(4\delta_u + \delta_{pu})) + \frac{\varepsilon}{2}. \tag{41}$$

Algorithm 1. Adaptive gradient method for problems with inexact oracle

```
Input: accuracy $\varepsilon > 0$, uncontrolled oracle error $\delta_u > 0$, uncontrolled error of composite prox-mapping $\delta_{pu} > 0$, starting point $x_0 \in X^0$, initial guess $L_0 > 0$, prox-setup: $d(x)$ — 1-strongly convex w.r.t. $\|\cdot\|_{\mathcal{E}}$, $V[z](x) := d(x) - d(z) - \langle d'(z), x - z\rangle$.
Set $k = 0$.
repeat
  Set $M_k = \frac{L_k}{2}$.
  repeat
    Set $M_k = 2M_k$, $\delta_{c,k} = \delta_{pc,k} = \frac{\varepsilon}{20M_k}$.
    Calculate $\widetilde{f}(x_k, \delta_{c,k}, \delta_u)$ and $\widetilde{g}(x_k, \delta_{c,k}, \delta_u)$.
    Calculate
      $$w_k = \operatorname*{argmin}_{x\in X}{}^{\delta_{pc,k}+\delta_{pu}} \left\{\langle\widetilde{g}(x_k, \delta_{c,k}, \delta_u), x\rangle + M_k V[x_k](x) + h(x)\right\}. \tag{37}$$
    Calculate $\widetilde{f}(w_k, \delta_{c,k}, \delta_u)$.
  until
      $$\widetilde{f}(w_k, \delta_{c,k}, \delta_u) \leqslant \widetilde{f}(x_k, \delta_{c,k}, \delta_u) + \langle\widetilde{g}(x_k, \delta_{c,k}, \delta_u), w_k - x_k\rangle + \frac{M_k}{2}\|w_k - x_k\|^2_{\mathcal{E}} + \frac{\varepsilon}{10M_k} + 2\delta_u. \tag{38}$$
  Set $x_{k+1} = w_k$, $L_{k+1} = \frac{M_k}{2}$, $k = k + 1$.
until $\min_{i\in 1,\ldots,k} \left\|M_i(x_i - x_{i+1})\right\|_{\mathcal{E}} \leqslant \varepsilon$
Output: The point $x_{K+1}$ s.t. $K = \operatorname*{argmin}_{i\in 1,\ldots,k} \left\|M_i(x_i - x_{i+1})\right\|_{\mathcal{E}}$.
```

Moreover, the total number of checks of Inequality (38) *is no more than*

$$2N - 1 + \log_2 \frac{M_{N-1}}{L_0}. \tag{42}$$

Proof. First of all, let us show that the procedure of searching for point w_k satisfying (37), (38) is finite. Let $i_k \geqslant 0$ be the current number of performed checks of inequality (38) on step k. Then $M_k = 2^{i_k}L_k$. At the same time, by Definition 1, $L(\delta_{c,k}) = L\left(\frac{\varepsilon}{16M_k}\right) = L\left(\frac{\varepsilon}{16\cdot 2^{i_k}L_k}\right)$. Hence, by the assumptions of the theorem, there exists $i_k \geqslant 0$ s.t. $M_k = 2^{i_k}L_k \geqslant L(\delta_{c,k})$. At the same time, we have

$$\widetilde{f}(w_k, \delta_{c,k}, \delta_u) - \frac{\varepsilon}{20M_k} - \delta_u \overset{(2)}{\leqslant} f(w_k) \overset{(3)}{\leqslant} \tag{43}$$

$$\overset{(3)}{\leqslant} \widetilde{f}(x_k, \delta_{c,k}, \delta_u) + \langle\widetilde{g}(x_k, \delta_{c,k}, \delta_u), w_k - x_k\rangle + {} + \frac{L(\delta_{c,k})}{2}\|w_k - x_k\|^2_{\mathcal{E}} + \frac{\varepsilon}{20M_k} + \delta_u, \tag{44}$$

which leads to (38) when $M_k \geqslant L(\delta_{c,k})$.

Let us now obtain the rate of convergence. We denote, for simplicity, $\widetilde{f}_k = \widetilde{f}(x_k, \delta_{c,k}, \delta_u)$, $\widetilde{g}_k = \widetilde{g}(x_k, \delta_{c,k}, \delta_u)$, $\widetilde{g}_{X,k} = g_X\left(x_k, \widetilde{g}_k, \frac{1}{M_k}, \delta_{pc,k}, \delta_{pu}\right)$. Note that

$$\widetilde{g}_{X,k} \overset{(35),(36),(37)}{=} M_k(x_k - x_{k+1}). \tag{45}$$

Using the definition of x_{k+1}, we obtain, for any $k = 0, \ldots, N-1$,

$$f(x_{k+1}) - \frac{\varepsilon}{20M_k} - \delta_u = f(w_k) - \frac{\varepsilon}{20M_k} - \delta_u \overset{(2)}{\leqslant} \tag{46}$$

$$\overset{(2)}{\leqslant} \tilde{f}(w_k, \delta_{c,k}, \delta_u) \overset{(38)}{\leqslant} \tag{47}$$

$$\overset{(38)}{\leqslant} \tilde{f}_k + \langle \tilde{g}_k, x_{k+1} - x_k \rangle + \frac{M_k}{2}\|x_{k+1} - x_k\|_{\mathcal{E}}^2 + \frac{\varepsilon}{10M_k} + 2\delta_u \overset{(45)}{=} \tag{48}$$

$$\overset{(45)}{=} \tilde{f}_k - \frac{1}{M_k}\left\langle \tilde{g}_k, \tilde{g}_{X,k} \right\rangle + \frac{1}{2M_k}\left\|\tilde{g}_{X,k}\right\|_{\mathcal{E}}^2 + \frac{\varepsilon}{10M_k} + 2\delta_u \overset{(2),(39)}{\leqslant} \tag{49}$$

$$\overset{(2),(39)}{\leqslant} f(x_k) + \frac{\varepsilon}{20M_k} + \delta_u - \left[\frac{1}{M_k}\left\|\tilde{g}_{X,k}\right\|_{\mathcal{E}}^2 + h(x_{k+1}) - h(x_k) - \frac{\varepsilon}{20M_k} - \delta_{pu}\right] + \\ + \frac{1}{2M_k}\left\|\tilde{g}_{X,k}\right\|_{\mathcal{E}}^2 + \frac{\varepsilon}{10M_k} + 2\delta_u. \tag{50}$$

This leads to

$$\psi(x_{k+1}) \leqslant \psi(x_k) - \frac{1}{2M_k}\left\|\tilde{g}_{X,k}\right\|_{\mathcal{E}}^2 + \frac{\varepsilon}{4M_k} + 4\delta_u + \delta_{pu}, \quad k = 0, \ldots, N-1.$$

Summing up these inequalities, we get

$$\left\|\tilde{g}_{X,K}\right\|_{\mathcal{E}}^2 \sum_{k=0}^{N-1} \frac{1}{2M_k} \leqslant \sum_{k=0}^{N-1} \frac{1}{2M_k}\left\|\tilde{g}_{X,k}\right\|_{\mathcal{E}}^2 \leqslant \psi(x_0) - \psi(x_N) + \frac{\varepsilon}{4}\sum_{k=0}^{N-1}\frac{1}{M_k} + N(4\delta_u + \delta_{pu}).$$

Finally, since, for all $x \in X$ $\psi(x) \geqslant \psi^* > -\infty$ and $\tilde{g}_{X,K} \overset{(45)}{=} M_K(x_K - x_{K+1})$, we obtain

$$\left\|M_K(x_K - x_{K+1})\right\|_{\mathcal{E}}^2 \leqslant \left(\sum_{k=0}^{N-1}\frac{1}{2M_k}\right)^{-1}(\psi(x_0) - \psi^* + N(4\delta_u + \delta_{pu})) + \frac{\varepsilon}{2}, \tag{51}$$

which is (41). The estimate for the number of checks of inequality (38) is proved in the same way as in [Nesterov, Polyak, 2006], but we provide the proof for the reader's convenience. Let $i_k \geqslant 1$ be the total number of checks of inequality (38) on the step $k \geqslant 0$. Then $i_0 = 1 + \log_2 \frac{M_0}{L_0}$ and, for $k \geqslant 1$, $M_k = 2^{i_k - 1}L_k = 2^{i_k-1}\frac{M_{k-1}}{2}$. Thus, $i_k = 2 + \log_2 \frac{M_k}{M_{k-1}}$, $k \geqslant 1$. Then, the total number of checks of Inequality (38) is

$$\sum_{k=0}^{N-1} i_k = 1 + \log_2 \frac{M_0}{L_0} + \sum_{k=1}^{N-1}\left(2 + \log_2 \frac{M_k}{M_{k-1}}\right) = 2N - 1 + \log_2 \frac{M_{N-1}}{L_0}. \tag{52}$$

□

Let us consider two corollaries of the theorem above. The first is a simple case where in Definition 1 $L(\delta_c) \equiv L$. The second is the case where $L(\delta_c)$ is given by (10).

Corollar 1. *Assume that there exists a constant $L > 0$ s.t. for the dependence $L(\delta_c)$ in Definition* 1 *it holds that $L(\delta_c) \leqslant L$ for all $\delta_c > 0$. Assume also that there exists a number $\psi^* > -\infty$ such that $\psi(x) \geqslant \psi^*$ for all $x \in X$. Then, after N iterations of Algorithm* 1, *it holds that*

$$\left\|M_K(x_K - x_{K+1})\right\|_{\mathcal{E}}^2 \leqslant \frac{4L(\psi(x_0) - \psi^*)}{N} + 4L(4\delta_u + \delta_{pu}) + \frac{\varepsilon}{2}. \tag{53}$$

Moreover, the total number of checks of inequality (38) *is not more than*

$$2N + \log_2 \frac{L}{L_0}.$$

Proof. By our assumptions, for all iterations $k \geqslant 0$, there exists $i_k \geqslant 0$ s.t. $M_k = 2^{i_k} L_k \geqslant \geqslant L(\delta_{c,k}) \equiv L$. Hence, we can apply Theorem 1. Let $i_k \geqslant 1$ be the total number of checks of inequality (38) on a step $k \geqslant 0$. Then, for all $k \geqslant 0$, the inequality $M_k = 2^{i_k} L_k \leqslant 2L$ should hold. Otherwise the termination of the inner cycle would happen earlier. Using these inequalities, we obtain

$$\left(\sum_{k=0}^{N-1} \frac{1}{2M_k}\right)^{-1} \leqslant \left(\sum_{k=0}^{N-1} \frac{1}{4L}\right)^{-1} = \frac{4L}{N}.$$

Thus, (53) follows from Theorem 1. The same argument proves the second statement of the corollary.

□

Corollar 2. *Assume that the dependence* $L(\delta_c)$ *in Definition* 1 *is given by* (10) *for some* $\nu \in (0, 1]$, *i. e.*

$$L(\delta_c) = \left(\frac{1-\nu}{1+\nu} \cdot \frac{2}{\delta_c}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}, \quad \delta_c > 0. \tag{54}$$

Assume also that there exists a number $\psi^* > -\infty$ *such that* $\psi(x) \geqslant \psi^*$ *for all* $x \in X$. *Then, after N iterations of Algorithm* 1, *it holds that*

$$\left\|M_K(x_K - x_{K+1})\right\|_{\mathcal{E}}^2 \leqslant 2^{\frac{1+3\nu}{2\nu}} \left(\frac{1-\nu}{1+\nu} \cdot \frac{40}{\varepsilon}\right)^{\frac{1-\nu}{2\nu}} L_\nu^{\frac{1}{\nu}} \left(\frac{\psi(x_0) - \psi^*}{N} + (4\delta_u + \delta_{pu})\right) + \frac{\varepsilon}{2}. \tag{55}$$

Moreover, the total number of checks of inequality (38) *is no more than*

$$2N - 1 + \frac{1+\nu}{2\nu} + \frac{1-\nu}{2\nu} \log_2 \left(40 \cdot \frac{1-\nu}{1+\nu}\right) + \frac{1-\nu}{2\nu} \log_2 \frac{1}{\varepsilon} + \log_2 \frac{L_\nu^{\frac{1}{\nu}}}{L_0}.$$

Proof. First, let us check that, for any constants $c_1, c_2 > 0$, there exists an integer $i \geqslant 0$ s.t. $2^i c_1 \geqslant L\left(\frac{c_2}{c_1 2^i}\right)$. Substituting $\delta_c = \frac{c_2}{c_1 2^i}$ to (54) gives

$$L\left(\frac{c_2}{c_1 2^i}\right) = 2^{\frac{1-\nu}{1+\nu} i} c_3,$$

where $c_3 > 0$ is some constant. Since $1 - \frac{1-\nu}{1+\nu} = \frac{2\nu}{1+\nu} > 0$, we conclude that the required $i \geqslant 0$ exists. Thus, we can apply Theorem 1.

Let $i_k \geqslant 1$ be the total number of checks of inequality (38) on a step $k \geqslant 0$. Then, for all $k \geqslant 0$, the inequality $M_k = 2^{i_k} L_k \leqslant 2L(\delta_{c,k})$ should hold. Otherwise the termination of the inner cycle would happen earlier. From this inequality and (54) it follows that

$$M_k \leqslant 2\left(\frac{1-\nu}{1+\nu} \cdot \frac{40 M_k}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}. \tag{56}$$

Solving this inequality for M_k, we obtain

$$M_k \leqslant 2^{\frac{1+\nu}{2\nu}} \left(\frac{1-\nu}{1+\nu} \cdot \frac{40}{\varepsilon}\right)^{\frac{1-\nu}{2\nu}} L_\nu^{\frac{1}{\nu}}. \tag{57}$$

Hence,

$$\left(\sum_{k=0}^{N-1} \frac{1}{2M_k}\right)^{-1} \leqslant \left(\sum_{k=0}^{N-1} \frac{1}{4L}\right)^{-1} = 2^{\frac{1+3\nu}{2\nu}} \left(\frac{1-\nu}{1+\nu} \cdot \frac{40}{\varepsilon}\right)^{\frac{1-\nu}{2\nu}} \frac{L_\nu^{\frac{1}{\nu}}}{N}. \tag{58}$$

Now (55) follows from Theorem 1.

Using (42) and the bound (57), we obtain the estimate for the total number of checks of inequality (38). □

Let us make some remarks about the results obtained. First, if we set in Corollary 2 $\nu = 1$, we recover the result of Corollary 1. Second, in the situation of Corollary 2, to make the controlled part of the right-hand side smaller than ε, we need to choose

$$N \geqslant \text{const} \cdot \frac{L_\nu^{\frac{1}{\nu}} (\psi(x_0) - \psi^*)}{\varepsilon^{\frac{1+\nu}{2\nu}}}.$$

One can see that the less ν is, the worse is the bound. This is expected as for nonsmooth nonconvex problems the norm of gradient mapping $g_X(\cdot)$ at the stationary point could not be equal to zero. Third, we can see that the uncontrolled error $4\delta_u + \delta_{pu}$ can dramatically influence the error estimate, especially, when ν tends to zero.

Finally, let us explain why small $\left\| M_K(x_K - x_{K+1}) \right\|_{\mathcal{E}}$ means that x_{K+1} is a good approximation for the stationary point of the initial problem (1).

Lemma 3 (see [Nesterov, 2018, Theorem 3.1.23]). *Let in Problem* (1) $f(x)$ *be continuously differentiable,* $h(x)$ *be convex,* X *be a closed convex set. Assume that* x^* *is a local minimum in this problem. Then, for all* $x \in X$,

$$\langle \nabla f(x^*), x - x^* \rangle + h(x) - h(x^*) \geqslant 0. \tag{59}$$

Assume, for simplicity, that we are in the situation of Subsection 1.1. This means that $f(x)$ is $L(f)$-smooth, we can uniformly approximate its gradient

$$\|\overline{g}(x) - \nabla f(x)\|_{\mathcal{E},*} \leqslant \overline{\delta}_c^2 + \overline{\delta}_u^2, \tag{60}$$

and the set X is bounded with diameter D. Also, assume that the chosen prox-function $d(\cdot)$ is $L(d)$-smooth.

From (34), (35), (37), we find that there exists $\nabla h(x_{K+1}) \in \partial h(x_{K+1})$ s.t., for all $x \in X$,

$$\left\langle \widetilde{g}(x_K, \delta_{c,K}, \delta_u) + M_K \left[d'(x_{K+1}) - d'(x_K)\right] + \nabla h(x_{K+1}), x - x_{K+1} \right\rangle \geqslant -\delta_{pc,K} - \delta_{pu}.$$

Hence, by convexity of $h(x)$,

$$\langle \nabla f(x_{K+1}), x - x_{K+1} \rangle + h(x) - h(x_{K+1}) \geqslant \langle \nabla f(x_{K+1}) - \nabla f(x_K), x - x_{K+1} \rangle + \tag{61}$$

$$+ \langle \nabla f(x_K) - \widetilde{g}(x_k, \delta_{c,k}, \delta_u), x - x_{K+1} \rangle + \tag{62}$$

$$+ \langle M_k \left[d'(x_K) - d'(x_{K+1})\right], x - x_{K+1} \rangle - \delta_{pc,K} - \delta_{pu}, \quad x \in X. \tag{63}$$

By $L(f)$-smoothness of f and boundedness of X, we obtain

$$\langle \nabla f(x_{K+1}) - \nabla f(x_K), x - x_{K+1} \rangle \geqslant -\frac{L(f)}{M_K} \|M_K(x_K - x_{K+1})\|_{\mathcal{E}} D.$$

From (60), by boundedness of X, we get

$$\langle \nabla f(x_K) - \widetilde{g}(x_K, \delta_{c,K}, \delta_u), x - x_{K+1} \rangle \geqslant -(\overline{\delta}_{c,K}^2 + \overline{\delta}_u^2) D.$$

Using $L(d)$ smoothness of $d(x)$ and boundedness of X, we obtain

$$\left\langle M_k \left[d'(x_K) - d'(x_{K+1}) \right], \, x - x_{K+1} \right\rangle \geqslant -L(d) \| M_K (x_K - x_{K+1}) \|_{\mathcal{E}} D.$$

Substituting the last three inequalities into (63), we find that, if $\| M_K (x_K - x_{K+1}) \|_{\mathcal{E}} \leqslant \varepsilon$, then

$$\langle \nabla f(x_{K+1}), \, x - x_{K+1} \rangle + h(x) - h(x_{K+1}) \geqslant -\Theta(\varepsilon) - \overline{\delta}_u^2 D - \delta_{pu}.$$

Thus, at the point x_{K+1} the necessary condition in Lemma 3 approximately holds.

Conclusion

In this article, we propose a new adaptive gradient method for nonconvex composite optimization problems with inexact oracle and inexact proximal mapping. We show that, for problems with an inexact Hölder-continuous gradient, our method is universal in terms of the Hölder parameter and constant. For the proposed method, we prove the convergence theorem in terms of generalized gradient mapping and show that a point returned by our algorithm is a point where the necessary optimality condition approximately holds.

Acknowledgments. The author is very grateful to A. Nemirovski, Yu. Nesterov, and B. Polyak for fruitful discussions.

References

Ben-Tal A., Nemirovski A. Lectures on Modern Convex Optimization (Lecture Notes). — Personal web-page of A. Nemirovski, 2015.

Bogolubsky L., Dvurechensky P., Gasnikov A. et al. Learning supervised PageRank with gradient-based and gradient-free optimization methods // Advances in Neural Information Processing Systems 29 / D. D. Lee, M. Sugiyama, U. V. Luxburg et al. (eds.). — Curran Associates, Inc., 2016. — P. 4914–4922. — arXiv:1603.00717.

d'Aspremont A. Smooth optimization with approximate gradient // SIAM J. on Optimization. — 2008. — Vol. 19, No. 3. — P. 1171–1183.

Devolder O., Glineur F., Nesterov Y. First-order methods of smooth convex optimization with inexact oracle // Mathematical Programming. — 2014. — Vol. 146, No. 1. — P. 37–75.

Dvurechensky P., Gasnikov A. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle // Journal of Optimization Theory and Applications. — 2016. — Vol. 171, No. 1. — P. 121–145.

Ghadimi S., Lan G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming // Mathematical Programming. — 2016. — Vol. 156, No. 1. — P. 59–99.

Ghadimi S., Lan G., Zhang H. Generalized uniformly optimal methods for nonlinear programming // ArXiV preprint. — 2015.

Ghadimi S., Lan G., Zhang H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization // Mathematical Programming. — 2016. — Vol. 155, No. 1. — P. 267–305. — arXiv:1308.6594.

Nesterov Yu. Introductory lectures on convex optimization: a basic course. — Massachusetts: Kluwer Academic Publishers, 2004.

Nesterov Yu. Smooth minimization of non-smooth functions // Mathematical Programming. — 2005. — Vol. 103, No. 1. — P. 127–152.

Nesterov Yu. Universal gradient methods for convex optimization problems // Mathematical Programming. — 2015. — Vol. 152, No. 1. — P. 381–404.

Nesterov Yu. Lectures on convex optimization. — Springer International Publishing, 2018. — Vol. 137.

Nesterov Yu., Polyak B. Cubic regularization of Newton method and its global performance // Mathematical Programming. — 2006. — Vol. 108, No. 1. — P. 177–205.

Polyak B. Gradient methods for the minimisation of functionals // USSR Computational Mathematics and Mathematical Physics. — 1963. — Vol. 3, No. 4. — P. 864–878.

Polyak B. Introduction to optimization. — New York: Optimization Software, 1987.