Gradient Method With Inexact Oracle for Composite Non-Convex Optimization

Pavel Dvurechensky *

March 28, 2017

Abstract

In this paper, we develop new first-order method for composite non-convex minimization problems with simple constraints and inexact oracle. The objective function is given as a sum of "hard", possibly non-convex 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 non-convex functions with inexact Hölder-continuous gradient, functions given by 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 to use different proximal setup to adapt to geometry of the feasible set, adaptively chooses controlled oracle error, allows for inexact proximal mapping. We provide convergence rate for our method in terms of the norm of generalized gradient mapping and show that, in the case of inexact Hölder-continuous gradient, our method is universal with respect to Hölder parameters of the problem. Finally, in a particular case, we show that 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, complexity, gradient descent methods, first-order methods, parameter free methods, universal gradient methods.

AMS Classification: 90C30, 90C06, 90C26.

Introduction

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

\[
\min_{x \in X} \{\psi(x) := f(x) + h(x)\},
\]

\*Weierstrass Institute for Applied Analysis and Stochastics, Berlin; Institute for Information Transmission Problems RAS, Moscow, pavel.dvurechensky@wias-berlin.de
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 problem with this kind of inexactness is given in Bogolubsky et al. [2016], where the authors study a learning problem for parametric PageRank model.

First-order methods are widely developed since the earliest years of optimization theory, see, e.g., Polyak [1963]. 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 mostly considered convex optimization problems since they have good structure and allow to estimate rate of convergence for proposed algorithms. Recently, non-convex 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 non-convex optimization. Namely, it have 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 gradient method for unconstrained smooth convex problems. At the moment, theory for convex optimization algorithms with inexact oracle is well-developed in a series of papers d’Aspremont [2008], Devolder et al. [2014], Dvurechensky and Gasnikov [2016]. In d’Aspremont [2008], it was proposed to calculate inexacty the gradient of the objective function and extend Fast Gradient Method of Nesterov [2005] to be able to use inexact oracle information. In Devolder et al. [2014], a general concept of inexact oracle is introduced for convex problems. Primal, Dual and Fast gradient methods are analyzed. In Dvurechensky and Gasnikov [2016], the authors develop 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 non-convex smooth, non-smooth and stochastic problems is well developed in Ghadimi and Lan [2016], Ghadimi et al. [2016]. In Ghadimi and Lan [2016], problems of the form (1), where $X \equiv \mathbb{R}^n$ and $f(x)$ is a smooth non-convex function are considered in the case when the gradient of $f(x)$ is exactly available, as well as when it is available through stochastic approximation. Later, in Ghadimi et al. [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 non-convex 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 Hölder-continuous gradient. We say that a function $f(x)$ has
Hölder-continuous gradient on $X$ iff there exist $\nu \in [0, 1]$ and $L_\nu \geq 0$ s.t.

$$
\| \nabla f(x) - \nabla f(y) \|_{E, \ast} \leq L_\nu \| x - y \|_E, \quad x, y \in X.
$$

In [Devolder et al. 2014] it was shown that a convex problem with Hölder-continuous subgradient can be considered as a smooth problem with deterministic inexact oracle. Later, universal gradient methods for convex problems with Hölder-continuous subgradient were proposed in [Nesterov 2015]. These algorithms do not require to know Hölder parameter $\nu$ and Hölder constant $L_\nu$. Thus, they are universal with respect to these parameters. [Ghadimi et al. 2015] proposed methods for non-convex problems of the form (1), where $f(x)$ has 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 non-convex case the definition of inexact oracle in [Devolder et al. 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 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, has flexibility of choice of proximal setup, based on the geometry of set $X$.

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

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

The rest of the paper is organized as follows. In Section 1, we define deterministic inexact oracle for non-convex problems and provide several examples. In Section 2, we describe our algorithm, 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 literature.

**Notation** Let $E$ be a finite-dimensional real vector space and $E^*$ be its dual. We denote the value of linear function $g \in E^*$ at $x \in E$ by $\langle g, x \rangle$. Let $\| \cdot \|_E$ be some norm on $E$, $\| \cdot \|_{E, \ast}$ 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 \( \tilde{f}(x, \delta_c, \delta_u) \in \mathbb{R} \) and \( \tilde{g}(x, \delta_c, \delta_u) \in \mathcal{E}^* \) satisfying

\[
|f(x) - \tilde{f}(x, \delta_c, \delta_u)| \leq \delta_c + \delta_u, \tag{2}
\]

\[
f(y) - (\tilde{f}(x, \delta_c, \delta_u) - \langle \tilde{g}(x, \delta_c, \delta_u), y - x \rangle) \leq \frac{L(\delta_c)}{2} \| x - y \|^2 + \delta_c + \delta_u, \quad \forall y \in X. \tag{3}
\]

In this definition, \( \delta_c \) represents the error of the oracle, which we can control and make as small as we would like to. On the opposite, \( \delta_u \) represents the error, which we can not 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 \( x, y \in X \), \( \| \nabla f(x) - \nabla f(y) \|_{\mathcal{E},*} \leq L \| x - y \|_{\mathcal{E}} \).

2. Set \( X \) is bounded with \( \max_{x, y \in X} \| x - y \|_{\mathcal{E}} \leq D \).

3. There exist \( \bar{\delta}^1_u, \bar{\delta}^2_u > 0 \) and at any point \( x \in Q \), for any \( \bar{\delta}^1_c, \bar{\delta}^2_c > 0 \), we can calculate approximations \( \bar{f}(x) \) and \( \bar{g}(x) \) s.t. \( |f(x) - \bar{f}(x)| \leq \bar{\delta}^1_c + \bar{\delta}^1_u, \| \bar{g}(x) - \nabla f(x) \|_{\mathcal{E},*} \leq \bar{\delta}^2_c + \bar{\delta}^2_u \).

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

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

\[
\leq \tilde{f}(x) + \bar{\delta}^1_c + \bar{\delta}^1_u + \langle \nabla \bar{g}(x), y - x \rangle + \langle \nabla f(x) - \bar{g}(x), y - x \rangle + \frac{L}{2} \| x - y \|_{\mathcal{E}}^2 \tag{5}
\]

\[
\leq \tilde{f}(x) + \langle \nabla \bar{g}(x), y - x \rangle + \frac{L}{2} \| x - y \|_{\mathcal{E}}^2 + \bar{\delta}^1_c + \bar{\delta}^1_u + (\bar{\delta}^2_c + \bar{\delta}^2_u) D. \tag{6}
\]

Thus, \( (\tilde{f}(x), \bar{g}(x)) \) is an inexact first-order oracle with \( \delta_u = \bar{\delta}^1_u + \bar{\delta}^2_u D, \delta_c = \bar{\delta}^1_c + \bar{\delta}^2_c D, \) and \( L(\delta_c) \equiv L \).
1.2 Smooth Function with Hölder-Continuous Gradient

Assume that \( f(x) \) is differentiable and its gradient is Hölder-continuous, i.e. for some \( \nu \in [0, 1] \) and \( L_\nu \geq 0 \),
\[
\| \nabla f(x) - \nabla f(y) \|_* \leq L_\nu \| x - y \|_E^\nu, \forall x, y \in X.
\] (7)
Then
\[
f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_\nu}{1 + \nu} \| x - y \|_E^{1+\nu}, \forall x, y \in X.
\] (8)

It can be shown, see Nesterov [2015], Lemma 2, that, for all \( x \in X \) and any \( \delta > 0 \),
\[
f(y) - (f(x) - \langle \nabla f(x), y - x \rangle) \leq \frac{L(\delta)}{2} \| x - y \|_E^2 + \delta, \forall y \in X,
\] (9)
where
\[
L(\delta) = \left( \frac{1 - \nu \cdot 2}{1 + \nu \cdot \delta} \right)^{\frac{1+\nu}{\rho}} L_\nu^{\frac{\rho}{1+\nu}}.
\] (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 inexact as in Subsection 1.1, their approximations will again be an inexact first-order oracle.

1.3 Function Given by Maximization Subproblem

Assume that function \( f(x) : E \to \mathbb{R} \) is defined by an auxiliary optimization problem
\[
f(x) = \max_{u \in U \subseteq H} \{ \Psi(x, u) := -G(u) + \langle Au, x \rangle \},
\] (11)
where \( A : H \to E^* \) is a linear operator, \( G : H \to R \) is a continuously differentiable uniformly convex function of degree \( \rho \geq 2 \) with parameter \( \sigma_\rho \geq 0 \). The last means that
\[
\langle \nabla G(u_1) - \nabla G(u_2), u_1 - u_2 \rangle \geq \sigma_\rho \| u_1 - u_2 \|_H^\rho, \forall u_1, u_2 \in U,
\] (12)
where \( \| \cdot \|_H \) is some norm on \( H \). 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
\[
\nu = \frac{1}{\rho - 1}, \quad L_\nu = \frac{\| A \|_{H \to E^*}}{\sigma_\rho^{\rho-1}}.
\] (13)
where \( \| A \|_{H \to E^*} = \max\{ \| Au \|_{E^*} : \| u \|_H = 1 \} \).
Thus, since 
\[
\ln(t) \geq 0
\]

Using this inequality with 
\[
t = \sigma^{1/\rho} \|x^* - x\|_H, \quad \tau = \frac{\|A\|_{\mathcal{H}}}{\sigma^{1/\rho}} \|y - x\|_E,
\]

we obtain, for any \( y \in X \),
\[
\langle A(x^* - x), y - x \rangle \leq \|A\|_{\mathcal{H}} \|x^* - x\|_H \|y - x\|_E
\]

\[
\leq \sigma^{1/\rho} \|x^* - x\|_H \|y - x\|_E
\]

\[
= \frac{\sigma^{1/\rho}}{\rho} \|x^* - x\|_H + \frac{\rho - 1}{\rho - 1} \sigma^{1/\rho} \|y - x\|_E
\]

\[
= \frac{\sigma^{1/\rho} \|x^* - x\|_H + \frac{\rho - 1}{\rho - 1} \sigma^{1/\rho} \|y - x\|_E}{1 + \nu}
\]

\[
= \frac{\sigma^{1/\rho} \|x^* - x\|_H + \frac{\rho - 1}{\rho - 1} \sigma^{1/\rho} \|y - x\|_E}{1 + \nu}
\]

\[
\leq \frac{\sigma^{1/\rho} \|x^* - x\|_H + \frac{\rho - 1}{\rho - 1} \sigma^{1/\rho} \|y - x\|_E}{1 + \nu}
\]

\[
= \frac{\sigma^{1/\rho} \|x^* - x\|_H + \frac{\rho - 1}{\rho - 1} \sigma^{1/\rho} \|y - x\|_E}{1 + \nu}
\]
where \( \nu \) and \( L_\nu \) are defined in (13). At the same time, since \( \Psi(x,u) \) (11) is uniformly concave in second argument, we have

\[
\frac{\sigma}{\rho} \| u^*(x) - u_x \|_H^\rho \leq \Psi(x,u^*(x)) - \Psi(x,u_x) \leq \delta. \tag{27}
\]

Combining this inequality with the previous one, we obtain

\[
\langle A(u^*(x) - u_x), y - x \rangle \leq \frac{L_\nu}{1 + \nu} \| x - y \|_E^{1+\nu} + \delta. \tag{28}
\]

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

\[
f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_\nu}{1 + \nu} \| x - y \|_E^{1+\nu} \tag{29}\]

\[
\leq \Psi(x,u_x) + \delta + \langle Au_x, y - x \rangle + \langle A(u^*(x) - u_x), y - x \rangle + \frac{2L_\nu}{1 + \nu} \| x - y \|_E^{1+\nu} \tag{30}\]

\[
\leq \Psi(x,u_x) + \langle Au_x, y - x \rangle + \frac{2L_\nu}{1 + \nu} \| x - y \|_E^{1+\nu} + 2\delta \tag{31}\]

\[
\leq \Psi(x,u_x) + \langle Au_x, y - x \rangle + \frac{2L(\delta)}{2} \| x - y \|_E^2 + 4\delta. \tag{32}\]

Thus, we have obtained that \( \langle \Psi(x,u_x), Au_x \rangle \) is an inexact first-order oracle with \( \delta_u = 0 \), \( \delta_c = 4\delta \), and \( L(\delta_c) \) given by (10) with \( \delta = \delta_c/4 \).

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 and 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 \( \| \cdot \|_E \), i.e., for any \( x \in X^0, y \in X \)

\[
d(y) - d(x) - \langle d'(x), y - x \rangle \geq \frac{1}{2} \| y - x \|_E^2.
\]

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, \( \ell_1/\ell_2 \), simplex, nuclear norm, spectahedron can be found in [Ben-Tal and Nemirovski 2015]. We will use Bregman divergence in so called composite prox-mapping

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

where \( \gamma > 0, \bar{x} \in X^0, g \in E^* \) are given. We allow this problem to be solved inexactly in the following sense.
Definition 1. Assume that we are given $\delta_{pu} > 0$, $\gamma > 0$, $\bar{x} \in X^0$, $g \in \mathcal{E}^*$. We call a point $\bar{x} = \bar{x}(\bar{x}, g, \gamma, \delta_{pc}, \delta_{pu}) \in X^0$ an inexact composite prox-mapping iff for any $\delta_{pc} > 0$ we can calculate $\bar{x}$ and there exists $p \in \partial h(\bar{x})$ s.t. it holds that

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

We write

$$
\bar{x} = \arg \min_{x \in X} \delta_{pc} + \delta_{pu} \left\{ \left\langle g, x \right\rangle + \frac{1}{\gamma} V[x](x) + h(x) \right\}
$$

and define

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

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

Our main scheme is Algorithm 1.

We will need the following simple extension of Lemma 1 in [Ghadimi et al. 2016] to perform the theoretical analysis of our algorithm.

Lemma 1. Let $\bar{x} = \bar{x}(\bar{x}, g, \gamma, \delta_{pc}, \delta_{pu})$ be an inexact composite prox-mapping and $g_X(\bar{x}, g, \gamma, \delta_{pc}, \delta_{pu})$ be defined in (36). Then, for any $\bar{x} \in X^0$, $g \in \mathcal{E}^*$ and $\gamma, \delta_{pc}, \delta_{pu} > 0$, it holds

$$
\gamma \left\langle g, g_X(\bar{x}, g, \gamma, \delta_{pc}, \delta_{pu}) \right\rangle \geq \gamma \| g_X(\bar{x}, g, \gamma, \delta_{pc}, \delta_{pu}) \|_F^2 + (h(\bar{x}(\bar{x}, g, \gamma, \delta_{pc}, \delta_{pu}))) - h(x) - \delta_{pc} - \delta_{pu}.
$$

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

$$
\left\langle g, \bar{x} - \bar{x} \right\rangle \geq \frac{1}{\gamma} \left\langle d'(\bar{x}) - d'(\bar{x}), \bar{x} - \bar{x} \right\rangle + \left\langle p, \bar{x} - \bar{x} \right\rangle - \delta_{pc} - \delta_{pu}
$$

$$
\geq \frac{1}{\gamma} \left( \| \bar{x} - \bar{x} \|_2^2 + (h(\bar{x}) - h(\bar{x})) \right) - \delta_{pc} - \delta_{pu}.
$$

(40)

Applying the definition (36), we finish the proof. \qed

Now we state the main

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 $i \geq 0$ s.t. $2^i c_1 \geq L \left( \frac{c_2}{c_2^2} \right)$. Assume also that there exists a number $\psi^* > -\infty$ such that $\psi(x) \geq \psi^*$ for all $x \in X$. Then, after $N$ iterations of Algorithm 1, it holds that

$$
\| M_K(x_k - x_{k+1}) \|_F^2 \leq \left( \sum_{k=0}^{N-1} \frac{1}{2M_k} \right)^{-1} (\psi(x_0) - \psi^* + N(4\delta_u + \delta_{pu})) + \frac{\varepsilon}{2}.
$$

(41)
Algorithm 1 Adaptive Gradient Method for Problems with Inexact Oracle

\begin{itemize}
  \item 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. \( \| \cdot \|_E \), \( V[z](x) := d(x) - d(z) - \langle d'(z), x - z \rangle \).
  
  \begin{enumerate}
    \item Set \( k = 0 \).
    \item repeat
      \begin{enumerate}
        \item Set \( M_k = L_k / 2 \).
        \item repeat
          \begin{enumerate}
            \item Set \( M_k = 2M_k, \delta_{c,k} = \delta_{pc,k} = \frac{\varepsilon}{20M_k} \).
            \item Calculate \( \tilde{f}(x_k, \delta_{c,k}, \delta_u) \) and \( \tilde{g}(x_k, \delta_{c,k}, \delta_u) \).
            \item Calculate
              \begin{equation}
                w_k = \arg \min_{x \in X} \delta_{pc,k} + \delta_{pu} \{ \langle \tilde{g}(x_k, \delta_{c,k}, \delta_u), x \rangle + M_k V[x_k](x) + h(x) \}. \tag{37}
              \end{equation}
            \item Calculate \( \tilde{f}(w_k, \delta_{c,k}, \delta_u) \).
          \end{enumerate}
        \end{enumerate}
      \end{enumerate}
    \item until \( \tilde{f}(w_k, \delta_{c,k}, \delta_u) \leq \tilde{f}(x_k, \delta_{c,k}, \delta_u) + \langle \tilde{g}(x_k, \delta_{c,k}, \delta_u), w_k - x_k \rangle + \frac{M_k}{2} \| w_k - x_k \|_E^2 + \frac{\varepsilon}{10M_k} + 2\delta_u. \tag{38} \)
    \item Set \( x_{k+1} = w_k, L_{k+1} = M_k / 2, k = k + 1 \).
  \end{enumerate}
\end{itemize}

Output: The point \( x_{K+1} \) s.t. \( K = \arg \min_{i \in 1, \ldots, k} \| M_i(x_i - x_{i+1}) \|_E \leq \varepsilon \).

Moreover, the total number of checks of Inequality (38) is not more than
\begin{equation}
2N - 1 + \log_2 \frac{M_{N-1}}{L_0}. \tag{42}
\end{equation}

Proof First of all let us show that the procedure of search of point \( w_k \) satisfying (37), (38) is finite. Let \( i_k \geq 0 \) be the current number of performed checks of inequality (38) on the step \( k \). Then \( M_k = 2^{i_k} L_k \). At the same time, by Definition \( L(\delta_{c,k}) = L \left( \frac{\varepsilon}{16M_k} \right) = L \left( \frac{\varepsilon}{16 \cdot 2^k L_k} \right) \). Hence, by the Theorem assumptions, there exists \( i_k \geq 0 \) s.t. \( M_k = 2^{i_k} L_k \geq L(\delta_{c,k}) \). At the same time, we have
\begin{equation}
\tilde{f}(w_k, \delta_{c,k}, \delta_u) - \frac{\varepsilon}{20M_k} - \delta_u \leq f(w_k) \tag{43}
\end{equation}
\begin{equation}
\leq \tilde{f}(x_k, \delta_{c,k}, \delta_u) + \langle \tilde{g}(x_k, \delta_{c,k}, \delta_u), w_k - x_k \rangle + \frac{L(\delta_{c,k})}{2} \| w_k - x_k \|_E^2 + \frac{\varepsilon}{20M_k} + \delta_u. \tag{44}
\end{equation}
which leads to (38) when $M_k \geq L(\delta_{c,k})$.

Let us now obtain the rate of convergence. We denote, for simplicity, $\tilde{f}_k = \tilde{f}(x_k, \delta_{c,k}, \delta_u)$, $
abla f(x_k, \delta_{c,k}, \delta_u)$, $\tilde{g} = \tilde{g}(x_k, \delta_{c,k}, \delta_u)$, $\tilde{g}_X = g_X(x_k, \delta_{c,k}, \delta_u)$. Note that

$$\tilde{g}_X = M_k(x_k - x_{k+1}).$$

(45)

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

$$f(x_{k+1}) - \frac{\varepsilon}{20M_k} - \delta_u = f(w_k) - \frac{\varepsilon}{20M_k} - \delta_u$$

(46)

$$\leq \tilde{f}(w_k, \delta_{c,k}, \delta_u)$$

(47)

$$\leq \tilde{f}_k + \langle \tilde{g}_k, x_{k+1} - x_k \rangle + \frac{M_k}{2} \|x_{k+1} - x_k\|^2 + \frac{\varepsilon}{10M_k} + 2\delta_u$$

(48)

$$\leq f(x_k) + \frac{\varepsilon}{20M_k} + \delta_u - \left[ \frac{1}{M_k} \|\tilde{g}_X\|^2 + h(x_{k+1}) - h(x_k) - \frac{\varepsilon}{20M_k} - \delta_u \right]$$

(49)

$$+ \frac{1}{2M_k} \|\tilde{g}_X\|^2 + \frac{\varepsilon}{10M_k} + 2\delta_u.$$

This leads to

$$\psi(x_{k+1}) \leq \psi(x_k) - \frac{1}{2M_k} \|\tilde{g}_X\|^2 + \frac{\varepsilon}{4M_k} + 4\delta_u + \delta_{pu}, \quad k = 0, \ldots, N - 1.$$

Summing up these inequalities, we get

$$\|\tilde{g}_X\|^2 \sum_{k=0}^{N-1} \frac{1}{2M_k} \leq \sum_{k=0}^{N-1} \frac{1}{2M_k} \|\tilde{g}_X\|^2 \leq \psi(x_0) - \psi(x_N) + \frac{\varepsilon}{4} \sum_{k=0}^{N-1} \frac{1}{M_k} + N(4\delta_u + \delta_{pu}).$$

Finally, since, for all $x \in X$, $\psi(x) \geq \psi^* > -\infty$ and $\tilde{g}_X = M_k(x_k - x_{k+1})$, we obtain

$$\|M_k(x_k - x_{k+1})\|^2 \leq \left( \sum_{k=0}^{N-1} \frac{1}{2M_k} \right)^{-1} (\psi(x_0) - \psi^* + N(4\delta_u + \delta_{pu})) + \frac{\varepsilon}{2},$$

(51)

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

$$\sum_{k=0}^{N-1} i_k = 1 + \log_2 \frac{M_0}{L_0} + \sum_{k=1}^{N-1} \left( 2 + \log_2 \frac{M_k}{M_{k-1}} \right) = 2N - 1 + \log_2 \frac{M_{N-1}}{L_0}.\quad (52)$$
Let us consider two corollaries of the theorem above. First is a simple case, when in Definition 1 \( L(\delta_c) \equiv L \). Second is the case, when \( L(\delta_c) \) is given by (10).

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

\[
\|M_K(x_K - x_{K+1})\|_E^2 \leq \frac{4L(\psi(x_0) - \psi^*)}{N} + 4L(4\delta_u + \delta_{pu}) + \frac{\varepsilon}{2}. \tag{53}
\]

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

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

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

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

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

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

\[
\|M_K(x_K - x_{K+1})\|_E^2 \leq 2^{\frac{1+\nu}{2\nu}} \left( \frac{1 - \nu}{1 + \nu} \cdot \frac{40}{\varepsilon} \right)^{\frac{1+\nu}{2\nu}} L_{\nu}^{\frac{1}{2}} \left( \frac{\psi(x_0) - \psi^*}{N} + (4\delta_u + \delta_{pu}) \right) + \frac{\varepsilon}{2}. \tag{55}
\]

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

\[2N - 1 + \frac{1+\nu}{2\nu} + \frac{1 - \nu}{2\nu} \log_2 \left( 40 \cdot \frac{1 - \nu}{1 + \nu} \right) + \frac{1 - \nu}{2\nu} \log_2 \frac{\varepsilon}{2} + \log_2 \frac{L_{\nu}^{\frac{1}{2}}}{L_0}.
\]
Proof First, let us check that, for any constants $c_1, c_2 > 0$, there exists an integer $i \geq 0$ s.t. $2^i c_1 \geq L \left( \frac{c_2}{c_1^2} \right)$. Substituting $\delta = \frac{c_2}{c_1^2}$ to (54) gives

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

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 $i \geq 0$ exists. Thus, we can apply Theorem 1.

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

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

Solving this inequality for $M_k$, we obtain

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

Whence,

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

Now (55) follows from Theorem 1.

Using (12) and the bound (57), we obtain the estimate for the total number of checks of Inequality (38).

Let us make some remarks about the obtained results. 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 $\varepsilon$ we need to choose

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

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

Finally, let us explain, why small $\|M_K(x_K - x_{K+1})\|_{\mathcal{E}}$ means that $x_{K+1}$ is a good approximation for stationary point of the initial problem (1). Let us prove the following result, which was communicated to us by Prof. Yu. Nesterov without proof.

Lemma 2. 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 $x \in X$,

$$\langle \nabla f(x^*), x - x^* \rangle + h(x) + h(x^*) \geq 0. \quad (59)$$
\textbf{Proof} Let us fix an arbitrary point } x \in X. \text{ Denote } x_t = tx + (1 - t)x^* \in X, \text{ } t \in [0, 1]. \text{ Since } x^* \text{ is a local minimum in (1), } X \text{ is a convex set, } h(x) \text{ is a convex function, we obtain for all sufficiently small } t > 0

\begin{align*}
0 \leq \frac{f(x_t) + h(x_t) - f(x^*) - h(x^*)}{t} \leq \frac{f(x_t) - f(x^*)}{t} + h(x) - h(x^*).
\end{align*}

Taking the limit as } t \to +0, \text{ we prove the stated inequality.}

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

\begin{align*}
\|\tilde{g}(x) - \nabla f(x)\|_{\mathcal{E},s} \leq \tilde{\delta}_c^2 + \tilde{\delta}_u^2,
\end{align*}

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

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

\begin{align*}
\langle \tilde{g}(x_K, \delta_c, \delta_u) + M_K [d'(x_{K+1}) - d'(x_K)] + \nabla h(x_{K+1}), x - x_{K+1} \rangle \geq -\delta_{pc,K} - \delta_{pu}.
\end{align*}

Whence, by convexity of } h(x),

\begin{align*}
\langle \nabla f(x_{K+1}), x - x_{K+1} \rangle + h(x) - h(x_{K+1}) & \geq \langle \nabla f(x_{K+1}) - \nabla f(x_K), x - x_{K+1} \rangle \quad (61)
\end{align*}

\begin{align*}
+ \langle \nabla f(x_K) - \tilde{g}(x_k, \delta_c, \delta_u), x - x_{K+1} \rangle \quad (62)
\end{align*}

\begin{align*}
+ \langle M_k [d'(x_{K+1}) - d'(x_K)] , x - x_{K+1} \rangle - \delta_{pc,K} - \delta_{pu}, \quad x \in X. \quad (63)
\end{align*}

By } L(f)-\text{smoothness of } f, \text{ boundedness of } X, \text{ we obtain}

\begin{align*}
\langle \nabla f(x_{K+1}) - \nabla f(x_K), x - x_{K+1} \rangle & \geq -\frac{L(f)}{M_K} \| M_K (x_K - x_{K+1}) \|_{\mathcal{E}}D.
\end{align*}

From (60), by boundedness of } X, \text{ we get

\begin{align*}
\langle \nabla f(x_K) - \tilde{g}(x_k, \delta_c, \delta_u), x - x_{K+1} \rangle & \geq -(\tilde{\delta}_c^2 + \tilde{\delta}_u^2) D.
\end{align*}

Using } L(d) \text{ smoothness of } d(x) \text{ and boundedness of } X, \text{ we obtain

\begin{align*}
\langle M_k [d'(x_{K+1}) - d'(x_K)] , x - x_{K+1} \rangle & \geq -L(d) \| M_K (x_K - x_{K+1}) \|_{\mathcal{E}}D.
\end{align*}

Substituting last three inequalities to (63), we obtain that, if } \| M_K (x_K - x_{K+1}) \|_{\mathcal{E}} \leq \varepsilon, \text{ then

\begin{align*}
\langle \nabla f(x_{K+1}), x - x_{K+1} \rangle + h(x) - h(x_{K+1}) & \geq -\Theta(\varepsilon) - \tilde{\delta}_u^2 D - \delta_{pu}.
\end{align*}

Thus, at the point } x_{K+1} \text{ the necessary condition in Lemma 2 approximately holds.}
Conclusion

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

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

References

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

Lev Bogolubsky, Pavel Dvurechensky, Alexander Gasnikov, Gleb Gusev, Yurii Nesterov, Andrey Raigorodskii, Aleksey Tikhonov, and Maksim Zhukovskii. Learning supervised pagerank with gradient-based and gradient-free optimization methods. NIPS 2016, 2016. URL http://papers.nips.cc/paper/6565-learning-supervised-pagerank-with-gradient-based-and-

Alexandre d’Aspremont. Smooth optimization with approximate gradient. SIAM J. on Optimization, 19(3):1171–1183, October 2008. ISSN 1052-6234. doi: 10.1137/060676386. URL http://dx.doi.org/10.1137/060676386

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. Mathematical Programming, 146 (1):37–75, 2014. ISSN 1436-4646. doi: 10.1007/s10107-013-0677-5. URL http://dx.doi.org/10.1007/s10107-013-0677-5

Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. Journal of Optimization Theory and Applications, 171(1):121–145, 2016. ISSN 1573-2878. doi: 10.1007/s10957-016-0999-6. URL http://dx.doi.org/10.1007/s10957-016-0999-6

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Mathematical Programming, 156 (1):59–99, 2016. ISSN 1436-4646. doi: 10.1007/s10107-015-0871-8. URL http://dx.doi.org/10.1007/s10107-015-0871-8

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Generalized uniformly optimal methods for nonlinear programming. ArXiV preprint, 2015. URL https://arxiv.org/abs/1508.07384

14
Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016. ISSN 1436-4646. doi: 10.1007/s10107-014-0846-1. URL http://dx.doi.org/10.1007/s10107-014-0846-1

Yurii Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005. ISSN 1436-4646. doi: 10.1007/s10107-004-0552-5. URL http://dx.doi.org/10.1007/s10107-004-0552-5

Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015. ISSN 1436-4646. doi: 10.1007/s10107-014-0790-0. URL http://dx.doi.org/10.1007/s10107-014-0790-0

Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. ISSN 1436-4646. doi: 10.1007/s10107-006-0706-8. URL http://dx.doi.org/10.1007/s10107-006-0706-8

Boris Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864 – 878, 1963. ISSN 0041-5553. doi: http://dx.doi.org/10.1016/0041-5553(63)90382-3. URL http://www.sciencedirect.com/science/article/pii/0041555363903823

Boris Polyak. *Introduction to Optimization*. New York, Optimization Software, 1987.