Stopping Rules for Gradient Methods for Non-convex Problems with Additive Noise in Gradient

Fedor Stonyakin$^{1,2}$ · Ilya Kuruzov$^{1,3}$ · Boris Polyak$^{1,4}$

Received: 20 October 2022 / Accepted: 15 May 2023 / Published online: 1 June 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
We study the gradient method under the assumption that an additively inexact gradient is available for, generally speaking, non-convex problems. The non-convexity of the objective function, as well as the use of an inexactness specified gradient at iterations, can lead to various problems. For example, the trajectory of the gradient method may be far enough away from the starting point. On the other hand, the unbounded removal of the trajectory of the gradient method in the presence of noise can lead to the removal of the trajectory of the method from the desired global solution. The results of investigating the behavior of the trajectory of the gradient method are obtained under the assumption of the inexactness of the gradient and the condition of gradient dominance. It is well known that such a condition is valid for many important non-convex problems. Moreover, it leads to good complexity guarantees for the gradient method. A rule of early stopping of the gradient method is proposed. Firstly, it guarantees achieving an acceptable quality of the exit point of the method in terms of the function. Secondly, the stopping rule ensures a fairly moderate distance of this point from the chosen initial position. In addition to the gradient method with a constant step, its variant with adaptive step size is also investigated in detail, which makes it possible to apply the developed technique in the case of an unknown Lipschitz constant for the

Communicated by Paulo J.S. Silva.

Fedor Stonyakin
fedyor@mail.ru

Ilya Kuruzov
kuruzov.ia@phystech.edu

Boris Polyak
boris@ipu.ru

1 Moscow Institute of Physics and Technology, Moscow, Russia
2 V. I. Vernadsky Crimean Federal University, Simferopol, Russia
3 Institute for Information Transmission Problems, Moscow, Russia
4 Institute for Control Sciences, Moscow, Russia
gradient. Some computational experiments have been carried out which demonstrate effectiveness of the proposed stopping rule for the investigated gradient methods.

**Keywords** Non-convex optimization · Polyak–Łojasiewicz condition · Inexact gradient · Stopping rule · Adaptive method

**Mathematics Subject Classification** 49M37 · 90C25 · 65K05

1 Introduction

Gradient methods are relatively simple, and they require a low iteration cost as well as a small amount of memory, which explains their popularity [1–4, 8, 10, 11]. In this paper, we consider the problem of minimizing the function which satisfies the well-known Polyak–Łojasiewicz condition [1, 10, 15] and has Lipschitz continuous gradient. We suppose that the proposed method has access not to the exact, but to the approximate value of the gradient at any requested point.

It is worth noting that the issue of studying the influence of gradient errors on the estimates of the convergence rate of the first-order methods attracted many researchers (see, for example, [2–4, 16, 19]). However, we will focus on the distinguished class of non-convex problems. The non-convexity of the objective function of the problem, as well as the use of an inexactness of the specified gradient at iterations, can lead to various problems. In particular, in the absence of any early stopping rules, divergence of the gradient method trajectory from the starting point can be quite a large. It is problematic when the initial point of the method already has some appropriate properties. On the other hand, the unlimited divergence of the trajectory of the Gradient Descent method can lead to a larger distance from the desired global solution (see Examples 2.1 and 2.2).

The purpose of this paper is to study the estimate of the distance from initial point to points produced by the Gradient Descent method and to propose an early stopping rule that guarantees some compromise, such as a significant divergence of the trajectory from the chosen starting point of the method. Note that the early stopping rules in iterative procedures are actively studied for various types of problems. Apparently, for the first time, the ideology of early stopping of iterations was proposed in [5]. This paper is devoted to a technique for the approximate solution of ill-posed or ill-conditioned problems arising during regularization (in the mentioned work, the authors considered the problem of solving a linear equation). In this case, an early termination is aimed at overcoming the problem of the potential accumulation of errors in the regularization of the original problem. The topic of our paper is related to well-known approaches related to the early termination of first-order methods in the case of using inexact information about the gradient at iterations (see [16], Ch. 6, paragraph 1, and also, for example, the recent preprint [19]). However, the results known to us for convex (not strongly convex) problems differ from those obtained in this paper. The main difference is that usually either the achievement of the worst level in function is guaranteed (compared with the comment after theorem 2, section 1, chapter 6 of [16]) or estimates such as \(\|x_N - x_*\| \leq \|x_0 - x_*\|\) without examining \(\|x_N - x_0\|\).
(here \( \{ x_k \}_{k \in \mathbb{N}} \) is the sequence generated by the method; \( x_* \) is the global solution of the minimization problem closest to the starting point of the method \( x_0 \)).

In this paper, we obtained Theorem 2 devoted to the Gradient Descent method with a constant step size with a sufficiently small value of the inexact gradient. It indicates the level of accuracy with respect to the function that can be guaranteed after the proposed early stopping rule is fulfilled. It is important to note that this result can be applied to any \( L \)-smooth non-convex problem. Further, using Polyak–Łojasiewicz condition, this result is refined in Theorem 3, which describes the estimate of a sufficient number of iterations to achieve the desired quality of the output point by the function. Moreover, it contains an estimate (25) of the distance from to the starting point. The obtained results are compared with the well-known distance estimate [15] from the starting point to the nearest global solution (see Remark 3.3).

However, the method with a constant step-size imposes the need to efficiently estimate the Lipschitz constant of the gradient of the objective function, which can be problematic in practice. Moreover, many real problems lead to functions that have not an Lipschitz continuous gradient, and a condition such as (1) holds for such functions only locally on some subset. Therefore, we propose variations in Theorems 2 and 3 for the Gradient Descent method with an adaptively selected step-size. This makes it possible to apply an analog of Theorem 2 with the early stopping rule (30) to an arbitrary non-convex problem without additional conditions.

The last section of the paper is devoted to numerical experiments which explain the purpose of using the proposed stopping rules for some specific examples of object functions in problems: logistic regression, Rosenbrock and Nesterov–Skokov functions, quadratic function.

2 Preliminaries

In data analysis, non-convex problems often arise under the standard assumption that the gradient of the objective function \( f \) is Lipschitz-continuous with some constant \( L > 0 \) (or in other words, the function \( f \) is \( L \)-smooth):

\[
\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \| \quad \forall x, y \in \mathbb{R}^n,
\]

where \( \| \cdot \| \) (here and everywhere in the paper) denotes the Euclidean norm. For these problems, by applying the gradient-type methods, the generated sub-sequence of points, generated by applying the gradient-type methods, converges to the zero value of \( \| \nabla f(x) \| \). For this fact, we have the following known result (see, for example, [8, 16]).

**Theorem 1** Let \( f \) be an \( L \)-smooth function. Let us consider the gradient method

\[
x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)
\]
for the following optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x).$$  \hspace{1cm} (3)

Then, the following inequality holds:

$$\min_{k=0, \ldots, N-1} \| \nabla f(x_k) \| \leq \sqrt{\frac{2L(f(x_0) - f(x_*))}{N}},$$  \hspace{1cm} (4)

where $x_0$ is a starting point of the method and $x_*$ is a global minimum of the problem (3).

Let $f$ be an $L$-smooth function and its gradient satisfy the Polyak–Łojasiewicz condition (for brevity, we will write PL-condition) for some constant $\mu > 0$ [15] (see also the recent papers [1, 10], and the references therein):

$$f(x) - f^* \leq \frac{1}{2\mu} \| \nabla f(x) \|^2 \quad \forall x \in \mathbb{R}^n,$$  \hspace{1cm} (5)

where $f^* = f(x_*)$ is the value of the function $f$ at one of the global solutions $x_*$ of the optimization problem under consideration. Then, the Gradient Descent method converges at the rate of a geometric progression

$$f(x_N) - f^* \leq \left(1 - \frac{\mu}{L}\right)^N (f(x_0) - f^*) \leq \exp \left(-\frac{\mu}{L} N \right) (f(x_0) - f^*),$$  \hspace{1cm} (6)

$$\| x_* - x_0 \| \leq \frac{\sqrt{2L(f(x_0) - f^*)}}{\mu}.$$  \hspace{1cm} (7)

From [10], it is known that the PL-condition (5) implies the following so-called quadratic growth condition:

$$f(x) - f^* \geq \frac{\mu}{2} \inf_{x_*} \| x - x_* \|^2 \quad \forall x \in \mathbb{R}^n,$$

whence one can obtain that (6) means that the Gradient Descent method also converges in argument at the rate of a geometric progression

$$\inf_{x_*} \| x_N - x_* \|^2 \leq \frac{2}{\mu} \exp \left(-\frac{\mu}{L} N \right) (f(x_0) - f^*).$$

It is worth noting that the gradient dominance condition (5) certainly holds for a strongly convex objective function $f$. However, there are known examples, where PL-condition holds, but one cannot be sure even that $f$ is convex (see, for example, [14]). So from [8], we can consider the problem of finding some solution to a system of nonlinear equations $g(x) = 0$ (written in a vector form), where $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $m \leq n$. 

\copyright Springer
Let us introduce the Jacobian matrix $J(x) = \frac{\partial g(x)}{\partial x} = \begin{bmatrix} \frac{\partial g_i(x)}{\partial x_j} \end{bmatrix}_{i,j=1}^{m,n}$ of the mapping $g$ and assume that there exists $\mu > 0$ such that for all $x \in \mathbb{R}^n$ the Jacobian matrix is uniformly non-singular, i.e., $\lambda_{\min}(J(x)[J(x)]^\top) \geq \mu$. In this case, the function $f(x) = \|g(x)\|_2$ satisfies condition (5) for an arbitrary $x^*$ such that $f(x^*) = 0$, i.e., $g(x^*) = 0$ [12]. We would like to mention separately the review [1], which describes in detail a deep learning-motivated example of a nonlinear equation-related minimization problem with over-parametrization for a non-convex smooth function with the PL-condition.

### 2.1 Formulation of the Problem

In this paper, we consider the problem of minimizing the function $f$ which satisfies PL-condition (5) and has $L$-Lipschitz continuous gradient with some constant $L > 0$ (1). We suppose that the method has access not to the exact, but to the approximate value of the gradient $\tilde{\nabla} f(x)$ at any requested point $x$, which means the following

$$\nabla f(x) = \tilde{\nabla} f(x) + v(x), \quad \|v(x)\| \leq \Delta$$

for some fixed $\Delta > 0$. Then, (5) means that

$$f(x) - f^* \leq \frac{1}{\mu}(\|\tilde{\nabla} f(x)\|^2 + \Delta^2) \quad \forall x \in \mathbb{R}^n. \tag{9}$$

Therefore, $\|\tilde{\nabla} f(x)\|^2 + \Delta^2 \geq \mu(f(x) - f^*)$, where

$$\|\tilde{\nabla} f(x)\|^2 \geq \mu(f(x) - f^*) - \Delta^2 \quad \forall x \in \mathbb{R}^n. \tag{10}$$

Let’s remind that the unlimited divergence of the trajectory of the Gradient Descent method can lead to a larger distance from the desired global solution. Let us describe some situations of this type.

#### Example 2.1

As a simple example of a non-strongly convex function that satisfies the gradient dominance condition, we consider

$$f(x) = \langle Ax, x \rangle, \tag{11}$$

where $A = \text{diag}(L, \mu, 0)$ is a 3-order diagonal matrix with exactly two positive entries $L > \mu > 0$. If for the problem of minimizing the function (11) we assume that there is a gradient error $v(x) = (0, 0, \Delta)$ for $\Delta > 0$, then for $x_0 = (0, 0, 0)$, $h_k > 0$ and $x_{k+1} = x_k - h_k \tilde{\nabla} f(x_k)$, we have $\lim_{k \to \infty} \|x_{k+1}\|_2 = \infty$.

#### Example 2.2

Further, we can consider the Rosenbrock function of two variables $x = (x^{(1)}, x^{(2)})$:

$$f(x) = 100 \left( x^{(2)} - \left( x^{(1)} \right)^2 \right)^2 + \left( 1 - x^{(1)} \right)^2.$$
Let our method start from \( x_0 = (1, 1) = x^* \). Then, at each step of the gradient method, the error of the gradient \( v(x_k) \) is such that \( x_k^{(2)} = (x_k^{(1)})^2 \) and without stopping rule the trajectory can go very far from the global solution \( x^* \). Similarly, the trajectory of the gradient method can be unbounded for the objective function of two variables \( f(x) = (x^{(2)} - (x^{(1)})^2)^2 \).

### 3 Proposed Approach and Main Theoretical Results

#### 3.1 Variant of the Gradient Descent Method with a Constant Step-Size

We assume that the values of the parameters \( L > 0 \) and \( \Delta > 0 \) are known. Also, the Gradient Descent method of the following form can be applied to solve the minimization problem of the function \( f \):

\[
x_{k+1} = x_k - \frac{1}{L} \tilde{\nabla} f(x_k)
\]

In view of (1) for the method (12), we get

\[
f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \| x_{k+1} - x_k \|^2
\]

\[
= f(x_k) - \frac{1}{L} \langle \nabla f(x_k), \tilde{\nabla} f(x_k) \rangle + \frac{1}{2L} \| \tilde{\nabla} f(x_k) \|^2
\]

\[
= f(x_k) + \frac{1}{2L} \left( \| \nabla f(x_k) \|^2 - 2 \langle \nabla f(x_k), \tilde{\nabla} f(x_k) \rangle + \| \tilde{\nabla} f(x_k) \|^2 \right)
- \| \nabla f(x_k) \|^2
\]

\[
= f(x_k) + \frac{1}{2L} \| \nabla f(x_k) - \tilde{\nabla} f(x_k) \|^2 - \frac{\| \tilde{\nabla} f(x_k) \|^2}{2L}
\]

\[
\leq f(x_k) + \frac{\Delta^2}{2L} - \frac{1}{2L} \| \nabla f(x_k) \|^2,
\]

i.e.,

\[
f(x_{k+1}) - f(x_k) \leq \frac{\Delta^2}{2L} - \frac{1}{2L} \| \nabla f(x_k) \|^2. \tag{13}
\]

Summing up inequalities (13) over \( k = 0, N - 1 \) leads us to an estimate

\[
\min_{k=0,\ldots,N-1} \| \nabla f(x_k) \| \leq \sqrt{\Delta^2 + \frac{2L(f(x_0) - f(x^*))}{N}} \leq \Delta + \sqrt{\frac{2L(f(x_0) - f(x^*))}{N}}. \tag{14}
\]
Note that, in contrast with (4), the estimate (14) points to the potential divergence of the Gradient Descent method in the case of an additively inexact gradient. Specific examples of such situations were described above.

Taking into account (5) and (13), we get

\[ f(x_{k+1}) - f(x_k) \leq \frac{\Delta^2}{2L} - \frac{2\mu(f(x_k) - f^*)}{2L} = -\frac{\mu}{L}(f(x_k) - f^*) + \frac{\Delta^2}{2L}; \]

thus,

\[ f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)(f(x_k) - f^*) + \frac{\Delta^2}{2L} \]

\[ \leq \left(1 - \frac{\mu}{L}\right)^{k+1}(f(x_0) - f^*) + \frac{\Delta^2}{2L} \left(1 + 1 - \frac{\mu}{L} + \cdots + \left(1 - \frac{\mu}{L}\right)^k\right) \]

\[ \leq \left(1 - \frac{\mu}{L}\right)^{k+1}(f(x_0) - f^*) + \frac{\Delta^2}{2\mu}, \]

i.e.,

\[ f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{k+1}(f(x_0) - f^*) + \frac{\Delta^2}{2\mu}. \] (15)

\textbf{Remark 3.1} It is important to note that bounds (14) and (15) cannot be improved for the Gradient Descent method with an additively inexact gradient in the general case. For example, the lower estimates of accuracy with respect to the function \(O\left(\frac{\Delta^2}{2\mu}\right)\) are known even on the class of strongly convex functions (see, for example, section 2.11.1 of the manual [20], as well as references therein). In this regard, we consider the following example:

\[
\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{2} \sum_{i=1}^{n} \lambda_i \left(x^{(i)}\right)^2, \tag{16}
\]

where \(0 \leq \mu = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n = L\), such that \(L \geq 2\mu\). The global solution of problem (16) is \(x^*_n = 0 \in \mathbb{R}^n\). Suppose that an inexact gradient is available at the current point of the feasible area. Besides, the error is only in the calculation of the first component of the gradient. That is, instead of \(\partial f(x)/\partial x^{(1)} = \mu x^{(1)}\), we have only \(\tilde{\partial} f(x)/\partial x^{(1)} = \mu x^{(1)} - \Delta\), for some \(\Delta > 0\). Then, for the simplest Gradient Descent method (12) one can obtain that for \(x_0^{(1)} \geq 0\) and sufficiently large \(k \in \mathbb{N}\) \((k \gg L/\mu)\) the following inequality holds:

\[
x_k^{(1)} \geq \frac{\Delta}{L} 1 - \left(1 - \frac{\mu/L}{L}\right)^k \approx \frac{\Delta}{\mu}. \tag{17}
\]

Therefore, \(f(x_k) - f(x_*) \gtrsim \frac{\Delta^2}{2\mu}\).
Further, in view of

\[ \| \nabla f(x_k) \|^2 \geq \frac{\| \tilde{\nabla} f(x_k) \|^2}{2} - \Delta^2, \]

from (13), the inexact gradient satisfies the following inequality:

\[ f(x_{k+1}) - f(x_k) \leq \frac{\Delta^2}{2L} - \frac{1}{2L} \left( \frac{\| \tilde{\nabla} f(x_k) \|^2}{2} - \Delta^2 \right), \]

whence we have

\[ f(x_{k+1}) - f(x_k) \leq \frac{\Delta^2}{L} - \frac{1}{4L} \| \tilde{\nabla} f(x_k) \|^2. \quad (18) \]

Inequality (18) shows that if the value \( \| \tilde{\nabla} f(x_k) \| \) is sufficiently large, it can be guaranteed that \( f(x_{k+1}) < f(x_k) \). Thus, for any \( C > 2 \), an alternative arises: either the inequality \( \| \tilde{\nabla} f(x_k) \| \leq C \Delta \) holds, or

\[ f(x_{k+1}) - f(x_k) < - \frac{\Delta^2}{L} \left( \frac{C^2}{4} - 1 \right). \]

In the first case, the inequality \( \| \tilde{\nabla} f(x_k) \| \leq C \Delta \) guarantees the achievement of an acceptable quality of the output point \( x_k \) with respect to the function due to PL-condition. In the second case, we can guarantee the decreasing with respect to the function for \( C > 2 \).

So, it is possible to get \( x_k \) such that the value of \( f(x_k) \) is close enough to the minimum \( f^* \). For definiteness, let us choose \( C = \sqrt{6} \) (to get a “convenient” coefficient) and consider 2 scenarios:

1. \( \| \tilde{\nabla} f(x_k) \| > \Delta \sqrt{6} \), whence, taking (18) into account, we obtain the inequality

\[ f(x_{k+1}) - f(x_k) < - \frac{\Delta^2}{2L}. \quad (19) \]

2.

\[ \| \tilde{\nabla} f(x_k) \| \leq \Delta \sqrt{6}, \quad (20) \]

whence, in view of (9), we have

\[ f(x_k) - f^* \leq \frac{7\Delta^2}{\mu}. \quad (21) \]

Let us consider estimate (21) acceptable for the function level and agree to terminate process (12) if (20) is satisfied.
Let us investigate an alternative situation in which for any $k = 0, 1, \ldots, N - 1$, it is true that $\| \tilde{\nabla} f(x_k) \| > \Delta \sqrt{6}$ and (19) holds, where

$$f(x_0) - f(x_N) = \sum_{k=0}^{N-1} (f(x_k) - f(x_{k+1})) > \frac{N\Delta^2}{2L},$$

i.e., $N < \frac{2L}{\Delta^2} (f(x_0) - f^*)$, which indicates the end of the process. Thus, we have the following result.

**Theorem 2** Let stopping criterion (20) be satisfied for the first time at the $N$-th iteration of the Gradient Descent method (12). Then, the output point $\hat{x} = x_N$ is guaranteed to satisfy the inequality

$$f(\hat{x}) - f^* \leq \frac{7\Delta^2}{\mu}.$$

In this case, the following estimate for the number of iterations before stopping criterion is valid

$$N < \frac{2L}{\Delta^2} (f(x_0) - f^*).$$

(22)

It is clear that for a small value of the parameter $\Delta > 0$, the right-hand side of inequality (22) leads to a significantly overestimated number of iterations. At the same time, the conducted computational experiments (see Sect. 4) showed no increase in the number of iterations with a significant decrease in $\Delta > 0$ due to the proposed early stopping rule (20).

However, in the case of a known $\mu$, the estimate for the number of steps $N$ can be improved if the quality in (21) is assumed to be sufficient. Using inequality (18), we get $\frac{1}{4L^2} \| \tilde{\nabla} f(x_k) \|^2 \leq \frac{\Delta^2}{L^2} + f(x_k) - f(x_{k+1})$, and due to $\tilde{\nabla} f(x_k) = L(x_k - x_{k+1})$, we have the following estimation for every $k \geq 0$:

$$\|x_{k+1} - x_k\|^2 \leq \frac{4\Delta^2}{L^2} + \frac{4(f(x_k) - f(x_{k+1}))}{L}$$

$$\leq \frac{4\Delta^2}{L^2} + \frac{4(f(x_k) - f^*)}{L}$$

$$\leq \frac{4\Delta^2}{L^2} + \frac{4\Delta^2}{\mu L} + \frac{4}{L} \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

Whence one can obtain the final estimation:

$$\|x_{k+1} - x_k\| \leq 2\Delta \sqrt{\frac{1}{L^2} + \frac{1}{\mu L} + 2 \left(1 - \frac{\mu}{L}\right)^k \frac{f(x_0) - f^*}{L}}.$$
Next, summing the inequalities above for \( k = 0, \ldots, N - 1 \) and taking into account (15), we have

\[
\|x_N - x_0\| \leq \sum_{k=0}^{N-1} \|x_{k+1} - x_k\| \leq 2N \Delta \sqrt{\frac{1}{L^2} + \frac{1}{\mu L}} + 2 \sum_{k=0}^{N-1} \left(1 - \frac{\mu}{L}\right)^{\frac{k}{2}} \sqrt{\frac{f(x_0) - f^*}{L}}.
\]  

(23)

If at some step (20) is satisfied, then the required accuracy by function (21) will be achieved. Therefore, we estimate \( N \) in an alternative situation [(20) does not hold for all \( k = 0, 1, \ldots, N - 1 \)]. We use inequality (15) and impose the requirement that the level of approximation with respect to the function \( f(x_N) - f^* \leq \frac{7\Delta^2}{\mu} \). In view of (15), it suffices to require that

\[
\left(1 - \frac{\mu}{L}\right)^N (f(x_0) - f^*) \leq \frac{6\Delta^2}{\mu},
\]

or

\[
\left(1 - \frac{\mu}{L}\right)^N \leq e^{-\frac{\mu N}{L}} \leq \frac{6\Delta^2}{\mu (f(x_0) - f^*)},
\]

where \( N \leq \left\lceil \frac{L}{\mu} \ln \frac{\mu(f(x_0) - f^*)}{6\Delta^2} \right\rceil \). In this case, (23) takes the following form:

\[
\|x_N - x_0\| \leq 2\Delta \sqrt{\frac{1}{\mu} + \frac{L}{\mu} \left\lceil \ln \frac{\mu(f(x_0) - f^*)}{6\Delta^2} \right\rceil + \frac{4\sqrt{L(f(x_0) - f^*)}}{\mu \mu}}.
\]

**Theorem 3** Let one of the following alternatives hold:

1. The Gradient Descent method (12) stops after \( N_\ast \) steps, where \( N_\ast \) is such that

\[
N_\ast = \left\lceil \frac{L}{\mu} \ln \frac{\mu(f(x_0) - f^*)}{6\Delta^2} \right\rceil.
\]

(24)

2. For some \( N \leq N_\ast \), at the \( N \)-th iteration of the method (12), stopping criterion (20) is satisfied for the first time.

Then, for the output point \( \hat{x} \) (\( \hat{x} = x_N \) or \( \hat{x} = x_{N_\ast} \)) of the method (12), the following inequalities hold:

\[
f(\hat{x}) - f^* \leq \frac{7\Delta^2}{\mu},
\]
\[ \| \hat{x} - x_0 \| \leq \frac{2 \Delta}{\mu} \sqrt{1 + \frac{L}{\mu} \ln \left( \frac{\mu(f(x_0) - f^*)}{6 \Delta^2} \right)} + \frac{4 \sqrt{L(f(x_0) - f^*)}}{\mu}. \]  

(25)

**Remark 3.2** Since it is often difficult to estimate the value of the parameter \( \mu \) and usually \( f^* \) is not known, the estimate of the number of iterations (24) is difficult to use in practice. If the implementation works only according to the stopping rule (20), then we can only confirm an upper bound on the number of iterations of the form (22), but in this case, we cannot guarantee (25). However, the estimate (23) remains relevant. Moreover, the estimation of the value \( \| \hat{x} - x_0 \| \) can be refined if the value of the parameter \( \mu \) is not available. Indeed, in view of (18) for the Gradient Descent method (12) with a constant step-size it holds that

\[ \frac{1}{4L} \| \tilde{\nabla} f(x_k) \|^2 \leq \frac{\Delta^2}{L^2} + \frac{4f(x_k) - f(x_{k+1})}{L}, \]

whence we have

\[ \| x_{k+1} - x_k \|^2 \leq \frac{4 \Delta^2}{L^2} + \frac{4(f(x_k) - f(x_{k+1}))}{L}, \]

i.e., \( \| x_{k+1} - x_k \| \leq \frac{2 \Delta}{L} + 2 \sqrt{\frac{f(x_k) - f(x_{k+1})}{L}} \). Further, after summing the inequalities above over \( k = 0, N - 1 \), we have:

\[
\| x_0 - x_N \| \leq \sum_{k=0}^{N-1} \| x_k - x_{k+1} \| \leq \frac{2N \Delta}{L} + 2 \sum_{k=0}^{N-1} \sqrt{\frac{f(x_k) - f(x_{k+1})}{L}} \\
\leq \frac{2N \Delta}{L} + 2 \sqrt{N} \sqrt{\sum_{k=0}^{N-1} \frac{f(x_k) - f(x_{k+1})}{L}} \\
= \frac{2N \Delta}{L} + 2 \sqrt{N} \sqrt{\frac{f(x_0) - f(x_N)}{L}} \\
\leq \frac{2N \Delta}{L} + 2 \sqrt{N} \sqrt{\frac{f(x_0) - f^*}{L}}.
\]

It is clear that for small values of the error \( \Delta > 0 \) the following inequality

\[ \| x_0 - x_N \| \leq \frac{2N \Delta}{L} + 2 \sqrt{N} \sqrt{\frac{f(x_0) - f^*}{L}} \]

may turn out to be worse than (25). Taking into account (22), we get

\[ \| x_0 - x_N \| \leq \frac{2 \Delta}{L} \cdot \frac{2L(f(x_0) - f^*)}{\Delta^2} + 2 \sqrt{\frac{2L(f(x_0) - f^*)}{\Delta^2} \cdot \frac{f(x_0) - f^*}{L}} \]

\[ = \frac{4 + 2 \sqrt{2}}{\Delta} (f(x_0) - f^*). \]

**Remark 3.3** By (7) and (25), the quantity \( \| \hat{x} - x_0 \| \) can be comparable with \( \| x_\ast - x_0 \| \) for a sufficiently small \( \Delta > 0 \).
Remark 3.4 In view of (25), it suffices to require that conditions (1) and (5) are satisfied only in $R$—neighborhood of the $x_0$, where

$$R = \frac{2\Delta}{\mu} \sqrt{1 + \frac{L}{\mu} \left[ \ln \frac{\mu(f(x_0) - f^*)}{6\Delta^2} \right] + \frac{4\sqrt{L(f(x_0) - f^*)}}{\mu}}.$$  

Remark 3.5 Note that stopping criterion $\|\nabla f(x_k)\| \leq C\Delta$ can be applied for all non-convex problems. Partially, in the case of general Łojasiewicz condition (see [7], Remark 2.23):

$$f(x) - f^* \leq \frac{1}{q\mu} \|\nabla f(x)\|^q$$

for some $q < 1$.

In this case, we can guarantee only quality $f(\hat{x}) - f^* \leq \frac{(C+1)^q \Delta^q}{\mu}$ for the output point $\hat{x}$. But in general case we have sublinear rate [6] of convergence and estimation (25) for distance between $\hat{x}$ and $x_0$ does not hold.

3.2 Some Variant of the Gradient Descent Method with an Adaptive Step-Size Policy

In many applied optimization problems, it is difficult to estimate the Lipschitz constant of the gradient of the objective function. For example, the well-known Rosenbrock function and its multidimensional generalizations (for example, the Nesterov–Skokov function [13]) have only a locally Lipschitz gradient. Thus, it is impossible to estimate for them the Lipschitz constant of the gradient without additional restrictions on the domain in which the method operates. Therefore, we present a generalization of the universal gradient method from [11] for working with an inexact gradient of the functions satisfying PL-condition.

For $L$-smooth functions, we have the following well-known inequality:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n.$$  

For the inexact gradient (8), we can get a similar inequality:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + L\|x - y\|^2 + \frac{\Delta^2}{2L}, \quad \forall x, y \in \mathbb{R}^n.$$  

This inequality contains an exact calculation of the value of the function $f$ at an arbitrary point from dom $f$. For most important applications with an inexact gradient, we do not have an opportunity to make such a calculation. An important example of such problems is some optimization problems in a Hilbert space [18] and, in a particular case, inverse problems [9]. Therefore, further, we will discuss the possibility of using an inexact function value when checking the iteration exit criterion.
Let us assume that we can calculate the inexact value \( \tilde{f} \) of the function \( f \) at any point \( x \), so that

\[
|f(x) - \tilde{f}(x)| \leq \delta. \tag{26}
\]

Then, we have the following inequality:

\[
\tilde{f}(x) \leq \tilde{f}(y) + \langle \tilde{\nabla} f(y), x - y \rangle + L \|x - y\|^2 + \frac{\Delta^2}{2L} + 2\delta, \quad \forall x, y \in \mathbb{R}^n. \tag{27}
\]

Further, when \( \mu \) is known, we select the constant \( L \) in such a way that (27) is satisfied for the points from the neighboring iterations (see Algorithm 1).

**Algorithm 1** Adaptive Gradient Descent with Inexact Gradient.

**Require:** \( L_{\text{min}} \geq 0, L_0 \geq L_{\text{min}}, \delta \geq 0, \Delta \geq 0. \)

1: Set \( k := 0 \)
2: Calculate

\[
x_{k+1} = x_k - \frac{1}{2L_k} \tilde{\nabla} f(x_k) \tag{28}
\]

3: If the following inequality holds:

\[
\tilde{f}(x_{k+1}) \leq \tilde{f}(x_k) + \langle \tilde{\nabla} f(x_k), x_{k+1} - x_k \rangle + L_k \|x_{k+1} - x_k\|^2 + \frac{\Delta^2}{2L_k} + 2\delta. \tag{29}
\]

then \( k := k + 1, L_k := \max\left(\frac{L_{k-1}}{2}, L_{\text{min}}\right) \) and go to Step 2. Otherwise, \( L_k := 2L_k \) and go to Step 3.

4: **return** \( x_k \).

Similar to the approach of the method with a constant step-size proposed above, in the case of a sufficiently small inexact gradient

\[
\|\tilde{\nabla} f(x_k)\| \leq 2\Delta \tag{30}
\]

we agree to interrupt Algorithm 1. In this case, according to (9) we can guarantee that

\[
f(x_k) - f^* \leq \frac{5\Delta^2}{\mu}. \]

An alternative case, where condition (30) is not satisfied, can be investigated similarly to the constant step-size case in Sect. 3.1. A detailed proof is given in [17] (Appendix A).

The theoretical results about the operation of Algorithm 1 are presented in the following theorem.

**Theorem 4** Suppose \( f(x) \) satisfies PL-condition (5) and conditions (26), \( \Delta^2 \geq 16L\delta \) hold. Let the parameter \( L_{\text{min}} \) in Algorithm 1 be such that \( L_{\text{min}} \geq \frac{\mu}{4} \) and one of the following alternatives holds:

1. Algorithm 1 works \( N_* \) steps where \( N_* \) is such that

\[
N_* = \left\lceil \frac{8L}{\mu} \log \frac{\mu(f(x_0) - f^*)}{\Delta^2} \right\rceil. \tag{31}
\]
2. For some $N \leq N^*$, at the $N$-th iteration of Algorithm 1, stopping criterion (30) is satisfied for the first time.

Then, for the output point $\hat{x}$ ($\hat{x} = x_N$ or $\hat{x} = x_{N^*}$) of Algorithm 1, we have the following inequalities

$$f(\hat{x}) - f^* \leq \frac{5\Delta^2}{\mu},$$

$$\|\hat{x} - x_0\| \leq 8\frac{\Delta}{\mu} \sqrt{\frac{1}{2}\gamma^2 + 4\gamma \frac{L}{\mu} \log \frac{\mu(f(x_0) - f^*)}{\Delta^2}} + 16\frac{\sqrt{\gamma L(f(x_0) - f^*)}}{\mu},$$

(32)

where $\gamma = \frac{L}{L_{min}}$. Also, the total number of calls to the subroutine for calculating inexact values of the objective function and step (28) is not more than $2N + \log \frac{2L}{L_0}$.

As we can see, estimate (32) from Theorem 4 for the Gradient Descent with an adaptive step-size differs significantly from the estimate (25) from Theorem 3 for the method with a constant step-size, namely by the presence of the factor $\gamma$. In the worst case, the ratio of these two estimates can be $O\left(\frac{L}{\mu}\right)$. However, as it will be shown in experiments, the distances $\|\hat{x} - x_0\|$ for the methods differ insignificantly. In addition, note, that Algorithm 1 uses subroutines for finding the inexact value of the objective function more often than the gradient method with a constant step. But the number of calls to these subroutines in adaptive Algorithm 1 is not more than $2N + \log \frac{2L}{L_0}$. This means that the "cost" of an iteration of the adaptive algorithm is on average comparable to about two iterations of the non-adaptive method (12). At the same time, the accuracy achieved by the proposed methods is also approximately equal.

Remark 3.6 Note that condition (30) is satisfied for any $L_k \geq L$. By construction, we obtain that $L_k \leq 2L$. In the estimates above, the quantity $2L$ estimates the maximum value of the parameter $L_k$. The estimates above remain valid if $L$ is replaced by $\frac{1}{2} \max_{j \leq k} L_j$ and $\gamma$ by $\frac{\max_{j \leq k} L_j}{2 \min_{j \leq k} L_j}$. Similarly, we can replace the algorithm parameter $L_{\text{min}}$ with $\min_{j \leq k} L_j$.

Remark 3.7 Note that the estimate for the number of iterations (31) in Theorem 4 indicates the finiteness of the process, but it can be strongly overestimated. In practice, the following relation is a more interesting:

$$N^* = \left\lceil \frac{4\hat{L}}{\mu} \log \frac{\mu(f(x_0) - f^*)}{\Delta^2} \right\rceil,$$

where $\hat{L} = \frac{\mu}{4} \frac{1}{1 - \left( \prod_{j=0}^{N^* - 1} \left( 1 - \frac{\mu}{4L_j} \right) \right)^\frac{1}{N^*}} \leq 2L$ is a parameter depending on the fitted parameters $L_j$ in Algorithm 1.
Remark 3.8 Also note that we can relax the requirement $L_{\min} \geq \frac{\mu}{4}$ to $L_{\min} > 0$. In this case, the estimate for the distance from the starting point to the point $x_N$ at the $N$-th iteration (see the proof of the expression (41) in [17]) has the following form

$$
\|x_N - x_0\| \leq N\Delta \sqrt{\frac{1}{2L_{min}^2} + \frac{4}{\mu L_{min}}} + 16 \sqrt{\frac{L}{L_{\min}}} \sqrt{\frac{L(f(x_0) - f^*)}{\mu}}.
$$

But we can no longer use estimate (32) from Theorem 4. In this case, it is possible to evaluate the sufficient number of iterations of Algorithm 1, assuming that the stopping condition $\|\tilde{\nabla} f(x_k)\| \leq 2\Delta$ is not satisfied. Further, we obtain an estimate (see the proof of (37) in [17]) for $\Delta^2 > 16L\delta$:

$$
N < \frac{2L}{\Delta^2 - 16L\delta}(f(x_0) - f^*).
$$

For experimental comparison of the Gradient Descent methods with constant and adaptive steps, a single stopping criterion must be chosen. If we consider criterion (20) for the adaptive Algorithm 1 instead of (30), then the results of Theorem 4 about the number of iterations (31) and the estimate of the distance from $x_0$ to $\hat{x}$ will remain valid. Thus, criterion (20) makes it possible to achieve the same theoretical guarantees. Therefore, further in the experimental comparison of the variants of the gradient method (Algorithms 1 and (12)), we will use the stopping criterion (20).

4 Numerical Experiments

4.1 The Quadratic Form Minimization Problem

In this section, we compare the number of iterations of method (12), required to stop according to criterion (20), and the estimate for the number of iterations (24) to achieve estimate (21). To obtain the theoretical estimate for the number of iterations (24), we need the values of the constants $L$ and $\mu$. Therefore, as the first example, we consider a quadratic function for which these constants are easy to calculate.

As an inexact gradient, we use an exact gradient with random noise (8). In our experiments, we consider the following types of the additive inexactness $v(x)$ in (8):

- **Random** Randomly generated from a uniform distribution, i.e., $v(x) \sim U(S^n(0))$, where $S^n(0)$ is the $n$ dimensional sphere with radius 1 at the center 0.
- **Antigradient** $v(x) = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$.
- **Constant** $v(x) = v \in \mathbb{R}^n$, such that $\|v\| = 1$.

Let us explain these types of inexactness. Random unbiased noise is the standard assumption in many applications for stochastic gradient descent (see [1], [3]). Anti-gradient can be considered as inexactness that should significantly slow down the convergence of our method. Finally, in Examples 2.1 and 2.2 in Section 2.1 of this article, it was shown that the Gradient Descent method with constant inexactness can move far away from the initial point.
Let us start with a simple example that allows us to estimate the parameters $L$ and $\mu$. As shown in [15], the function $f(x) = \frac{1}{2} \langle x, Ax \rangle$ satisfies PL-condition if the operator $A$ is non-negative definite and its spectrum is separated from zero. In such a case, $\mu$ is the smallest nonzero eigenvalue of the matrix $A$. At the same time, the Lipschitz constant of the gradient is the largest eigenvalue of the matrix $A$. Thus, we consider the following problem of quadratic programming:

$$
\min_{x \in \mathbb{R}^n} \frac{1}{2} \sum_{j=k+1}^{n} d_j x_j^2,
$$

where $k$ is the number of zero eigenvalues of the matrix $A$, and $d_j$ are some positive constants. Thus, we have a quadratic form with a non-negative definite diagonal matrix. In this case, we can explicitly find the constants $\mu = \min_{j=k+1,n} (d_j), L = \max_{j=k+1,n} (d_j)$.

In the conducted experiments, we take $L = 1$ and change $\mu$ from 0 to 1. The parameters $d_j$ will be taken uniformly random from the interval $[\mu, L]$. We take the dimension $n = 100$ and $k = 10$ of zero eigenvalues. Let us compare the required number of iterations to achieve condition (20) and the estimate of $N_*$ from Theorem 3. As an inexactness, we will take Random noise $v(x)$. The results for problem (33) are presented in Table 1.

In Table 1, we can see that in all cases $N < N_*$. It means that stopping condition (20) is reached earlier than the theoretical estimate of the number of iterations $N_*$ justified using PL-condition (see Theorem 3). Also, we can note that the method converges much faster than the stated estimate for large values of $\mu$. At the same time, for small values of $\mu$, the value of $N_*$ exceeds $N$ by at most 2.5 times. For the other types of noise of the gradient, a similar picture is observed.

Note, $N_*$ is overestimated because parameters $L$ and $\mu$ are global parameters. Namely, for the considered quadratic problem (33) we have the following inequality:

$$
f(x) - f^* \leq \frac{1}{2} \| \nabla f(x) \|_D^2,
$$

where $\|x\|_D^2 = \sum_{j=1}^{n} d_j^{-1} x_j^2$. This inequality depends on the spectrum of quadratic problem (33) and gives more accurate estimation than (5).

Now, we compare the results of the Gradient Descent with a constant step-size (12) and the proposed Gradient Descent method with an adaptive step-size (Algorithm 1) when using stopping criterion (20). Tables 2 and 3 in [17] present the results of the experiments for the quadratic function in (33). The experiments were carried...
out for the uniformly distributed noise \( v(x) \) on the sphere. In these experiments, the inexactness \( \delta = 16 \Delta^2 \) in the function was taken. Note that in this case, the correlation of inexactness satisfies the condition of Theorem 4.

From Table 2, we can see that the adaptive method is inferior in real time to the Gradient Descent for all parameters \( \mu \) and \( \Delta \). However, it needs a smaller number of iterations for big values of \( \mu \).

### 4.2 The Problem of Minimizing the Logistic Regression Function

Now let us check the work of the proposed stopping criterion in the case when it is rather difficult to estimate the constant \( \mu \) of the function which satisfies the PL-condition. In this case, we will not be able to use estimate (24). This situation has been discussed in Remark 3.2. The detailed experiments are presented in [17] (Appendix B.2).

However, we note that, as shown by the previous experiment, condition (20) can be achieved in a significantly smaller number of steps compared to the theoretical estimate of the number of iterations \( N_\ast \) from Theorem 3. We will consider the following optimization problem associated with logistic regression

\[
\min \left\{ f(x) = \frac{1}{m} \sum_{i=1}^{m} \log \left( 1 + \exp \left( -y_i \langle w_i, x \rangle \right) \right) \right\}, \quad \text{where } y = (y_1, \ldots, y_m)^\top \in [-1, 1]^m \text{ is the feasible variable vector, } W = [w_1 \ldots w_m] \in \mathbb{R}^{n \times m} \text{ is the feature matrix, where each vector } w_i \in \mathbb{R}^n \text{ is from the same space as the optimized weight vector } w.
\]

Note that this problem may not have a finite solution in the general case. So we will create such an artificial data set that there is a finite vector \( x_\ast \) minimizing the given function. The details of data generation are presented in [17] (Appendix B.2).

In the conducted experiments, we chose \( n = 200, m = 700 \) and \( k = 10 < \min \left( n, \frac{m}{2} \right) \). We consider in this section the case of constant inexactness. From Fig. 1, it can be seen that the trajectories of the method are similar. Nevertheless, adding inexactness slows down the convergence to the corresponding output points. On the
Fig. 1 Rate of convergence of the gradient method in the gradient norm for different values of $\Delta$ for the problem of minimizing logistic regression using stopping criterion (20) for the constant $v$.

Fig. 2 Results of the gradient method with respect to the norm of the gradient without using the stopping criterion for $\Delta = 0.1$ for the problem of logistic regression minimization for the constant inaccuracy $\Delta v$. a The convergence rate with respect to the norm of the gradient; b the distance from the starting point to $x_k$.

Other hand, the trajectories have become more similar compared to the case of the inexactness directed along the minus of the gradient (see [17] (Appendix B.2)). However, in this case, in Fig. 2b it can be seen that without using the stopping criterion, the distance $\|x_k - x_0\|$ grows rather quickly. Thus, in the case of randomly generated gradient noise after $10^5$ iterations of Gradient Descent method (12) the distance was 1.25 times larger compared to the result without noise in the gradient. At the same time, in the case of a constant gradient specification error, these values differ by more than two orders of magnitude (see [17] (Appendix B.2)).

4.3 Some Experiments with the Rosenbrock-Type Function

In this subsection, we describe results of our investigation of the behavior of the proposed adaptive Algorithm 1 for some non-convex problems. The details are presented in [17] (Appendixes B.3 and B.4). Firstly, we considered the well-known two-dimensional Rosenbrock function $f(x_1, x_2) = 100(x_2 - (x_1)^2)^2 + (x_1 - 1)^2$. 

Springer
This function is not convex, and it satisfies the Lipschitz condition for the gradient only locally. Indeed, if we consider the line \( x_2 = 0 \), then we get \( f(x_1, 0) = 100x_1^4 + (x_1 - 1)^2 \). The gradient of this function does not satisfy the Lipschitz condition. On the other hand, the Rosenbrock function satisfies locally PL-condition.

In the conducted experiments, we will vary the value of the parameter \( \Delta \) and take \( \delta = \Delta^2 \). In Table 4 in [17] (Appendix B.3), we show the results for different types of noise. As previously, from the results presented in Table 4, we can see that the number of required iterations increases with decreasing \( \Delta \) (which also tightens the stopping condition). Moreover, it increases logarithmically, which coincides with the results of Theorem 4. We can also note that the resulting distance from the starting point \( x_0 \) to the last point does not exceed the distance from the starting point \( x_0 \) to the nearest optimal one \( x_* = (1, 1) \) everywhere. In addition, for all considered types of the gradient error (noise), a comparable convergence rate is observed according to the number of iterations until stopping criterion (30) is satisfied, and to the running time for the corresponding values of \( \Delta \).

Further, let us consider a system of nonlinear equations \( g(x) = 0 \), where \( g_1 = \frac{1}{4}(x_1 - 1), g_i = x_i - 2x_{i-1}^2 + 1, i = 2, n \). The problem of solving this system is equivalent to minimizing the following Nesterov–Skokov function (see [13])

\[
\begin{align*}
f(x) &= \frac{1}{4} (1 - x_1)^2 + \sum_{i=1}^{n-1} \left( x_{i+1} - 2x_i^2 + 1 \right)^2. 
\end{align*}
\]

This function is analogous to the Rosenbrock function. It is also non-convex and satisfies the Lipschitz gradient condition only locally. Also, function (34) has a global minimum at the point \((1, \ldots, 1)^\top\) and an optimal value \( f^* = 0 \). Moreover, this function locally satisfies PL-condition (see the proof in [17] (Appendix D)).

As it was seen from the results of the previous experiments, our proposed stopping criterion (30) of Algorithm 1 can work equally well for all considered types of noise in the gradient. In the current experiments for the Nesterov–Skokov function, we used the random noise of the gradient which is uniformly distributed on the sphere. For the experiments, the starting point is \((-1, 1, \ldots, 1)^\top\) and therefore \( \|x_0 - x_*\| = 2 \). We will vary the value of the inexactness \( \Delta \) and the dimension of the problem \( n \).

Table 5 in [17] (Appendix B.4) shows the results of the adaptive gradient method 1 for the Nesterov–Skokov function (34). Firstly, we see that as the dimension of \( n \) increases, the difference between the required time to solve the problem for different \( \Delta \) grows significantly. Secondly, for different \( n \) with the same \( \Delta \), the method converges to a solution with significantly different accuracy. We can also note that \( \|x_N - x_0\| \) exceeds \( \|x_0 - x_*\| \) by at most 2 times. Moreover, significant upward deviations are observed for the cases when numerous iterations are made (\( n = 5 \) and \( \Delta = 10^{-4}, 10^{-3} \)). It can also be noted that even for sufficiently small values of the norm of the gradient, the accuracy by the function turns out to be quite low (which is typical for the Nesterov–Skokov function).
5 Conclusion

This paper studies stopping criteria for the gradient method with an inexact gradient. The authors focus on the case of non-convex functions. The paper presents a stopping criterion that finds a compromise between the accuracy of the obtained point and the distance to the starting point. Moreover, it is shown that the method moves away from the starting point to a distance comparable to the distance to the nearest solution if the function satisfies PL-condition.

Besides, the paper considers the cases of constant and adaptive step size in the gradient methods. For both cases, we present theoretical analysis and the number of iterations required to approach the stopping criterion or to find the point with the required quality.

In addition, the paper contains numerical experiments demonstrating the work of the stopping criterion. In particular, there are experiments on a quadratic function (convex, but not strongly convex), demonstrating the stopping criterion to be approached faster than the theoretical estimation of the iteration number $N_\alpha$. Also, we present experiments on the problem of logistic regression where the objective function is convex and meets PL-condition only locally. The proposed stopping criterion on this function stops the growth of the distance $\|x_k - x_0\|$. Moreover, we present experiments on non-convex functions: the Rosenbrock function and its multidimensional generalization, which is the Nesterov–Skokov function. The first function demonstrates that our stopping criterion works with general types of inexactness. The second function demonstrates that even a small inexactness can lead to quite a high value of the function and this value cannot be improved. Also, we demonstrate that for some noises, the gradient method can move away quite far on the Nesterov–Skokov function without a stopping criterion.

Acknowledgements The research by F. Stonyakin in Sects. 2 and 3.1 was supported by the strategic academic leadership program "Priority 2030" (Agreement 075-02-2021-1316, 30.09.2021). The research by B. Polyak in Sect. 1 and by I. Kuruzov in Sect. 4 was supported by the Russian Science Foundation (project No. 21-71-30005), https://rscf.ru/en/project/21-71-30005/.

References

1. Belkin, M.: Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. Acta Numer. 30, 203–248 (2021)
2. d’Aspremont, A.: Smooth optimization with approximate gradient. SIAM J. Optim. 19(3), 1171–1183 (2008)
3. Devolder, O.: Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. Ph.D. thesis, ICTEAM and CORE, Université Catholique de Louvain (2013)
4. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. 146(1), 37–75 (2014)
5. Emelin, I.V., Krasnosel’ski, M.A.: The stoppage rule in iterative procedures of solving ill-posed problems. Autom. Remote Control 39, 1783–1787.; Translation from Avtom. Telemekh. 1978(12), 59–63 (1979). (in Russian)
6. Frei, S., Gu, Q.: Proxy convexity: a unified framework for the analysis of neural networks trained by gradient descent (2021). arXiv preprint arXiv:2106.13792
7. Garrigos, G., Gower, R.M.: Handbook of convergence theorems for (stochastic) gradient methods (2023). arXiv preprint arXiv:2301.11235
8. Gasnikov, A.V.: Modern numerical optimization methods: The universal gradient descent method (2021). arXiv preprint arXiv:1711.00394
9. Kabanikhin, S.I.: Inverse and Ill-posed Problems. In: Inverse and Ill-posed Problems. deGruyter (2011)
10. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 795–811. Springer (2016)
11. Nesterov, Y.: Universal gradient methods for convex optimization problems. Math. Program. 152(1), 381–404 (2015)
12. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton method and its global performance. Math. Program. 108(1), 177–205 (2006)
13. Nesterov, Y., Skokov, V.: On the issue of testing unconstrained optimization algorithms. In: Numerical methods of mathematical programming, pp. 77–91. Moscow (1980) (in Russian)
14. Polyak, B., Tremba, A.: New versions of Newton method: step-size choice, convergence domain and under-determined equations. Optim. Methods Softw. 35(6), 1272–1303 (2020)
15. Polyak, B.T.: Gradient methods for minimizing functionals. Comput. Math. Math. Phys. 3(4), 864–878 (1963). (in Russian)
16. Polyak, B.T.: Introduction to optimization. Optim. Softw. Inc. N. Y. I, 32 (1987)
17. Polyak, B.T., Kuruzov, I.A., Stonyakin, F.S.: Stopping rules for gradient methods for non-convex problems with additive noise in gradient (2022). arXiv preprint arXiv:2205.07544
18. Vasiliev, F.: Optimization Methods. FP, Moscow (2002). (in Russian)
19. Vasin, A., Gasnikov, A., Spokoiny, V.: Stopping rules for accelerated gradient methods with additive noise in gradient (2021). arXiv preprint arXiv:2102.02921
20. Vorontsova, E., Hildbrand, R., Gasnikov, A., Stonyakin, F.: Convex Optimization (2021). arXiv preprint arXiv:2106.01946

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.