ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 9, pp. 797–816.

SPECIAL ISSUE

A Gradient Method with Inexact Gradient Information: A Study on Some Classes of (L_0, L_1)-Smooth Nonconvex Problems

S. S. Ablaev[*,**,a], **F. S. Stonyakin**[**,*,b], **M. N. Fedotov**[***,c], **M. S. Alkousa**[****,d], **O. S. Savchuk**[*,*****,e], **and A. V. Gasnikov**[****,**,f]

[*] *Vernadsky Crimean Federal University, Simferopol, Russia*
[**] *Moscow Institute of Physics and Technology, Dolgoprudny, Russia*
[***] *National Research University Higher School of Economics, Moscow, Russia*
[****] *Innopolis University, Innopolis, Russia*
[*****] *Adyghe State University, Maykop, Russia*
e-mail: [a] *seydamet.ablaev@yandex.ru,* [b] *fedyor@mail.ru,* [c] *MaximFd-Nk@yandex.ru,*
[d] *m.alkousa@innopolis.ru,* [e] *oleg.savchuk19@mail.ru,* [f] *gasnikov@yandex.ru*

Received March 3, 2025
Revised May 20, 2025
Accepted June 27, 2025

Abstract—In this paper, a gradient method is studied on the class of (L_0, L_1)-smooth objective functions under the condition that only approximate values of the gradient are available at the iterations of the method. This setting arises when using noisy data. Two classes of problems are considered, namely, quasar-convex functions with respect to any solution that satisfy the Polyak–Łojasiewicz gradient dominance condition (the PL condition) and quasar-convex functions without the PL condition but with an additional constraint on the quasar-convexity parameter. For quasar-convex functions with the PL condition, a nearly linear convergence rate is established in the neighborhood of the exact solution. If inexact gradient values are sufficiently small (which is achieved in a finite number of iterations), the method will converge with a nearly linear rate on the class of problems with the PL condition without the additional assumption of quasar-convexity. For $(0, M)$-smooth quasar-convex functions, an adaptive gradient method is proposed, and its convergence rate is estimated. As shown, when using exact gradient values, the method converges with a linear rate.

Keywords: gradient method, Δ-inexact gradient, (L_0, L_1)-smooth function, ρ-quasar-convex function, Polyak–Łojasiewicz condition, logistic regression

DOI: 10.31857/S0005117925090012

1. INTRODUCTION

With the rapid advancement of machine learning, it is increasingly topical to develop and analyze numerical optimization methods with effective convergence guarantees. To achieve the optimal convergence rate of conventional methods, such as gradient descent and its modifications, a key assumption is the smoothness of the objective function (i.e., the Lipschitz continuity of its gradient). In the context of modern machine learning problems, this assumption may turn out to be too stringent [1]. As a rule, quite natural and simple-to-implement methods perform poorly for deep learning problems, where the global smoothness condition often fails. For example, variance reduction methods [2–8] are known to be theoretically faster (for minimizing finite sums of smooth functions), but in practice may be inferior to methods that do not theoretically reduce variance [9]. According to the experiments presented in [1], when training neural networks, the norm of the

Hessian may correlate with the norm of the gradient of the loss function. Thus, the above factors motivate researchers to find and propose alternative assumptions relaxing the standard smoothness requirement for the objective function.

One such assumption is the so-called generalized (L_0, L_1)-smoothness, pioneered several years ago [1] for twice differentiable functions. This assumption allows the norm of the Hessian of the objective function to increase linearly with the growth of the gradient norm. In particular, (L_0, L_1)-smoothness can hold even for functions with polynomially growing gradients, which is a typical situation for deep learning problems. Moreover, the concept of (L_0, L_1)-smoothness covers a wide class of problems and can be generalized to differentiable, but not necessarily twice differentiable, functions [10, 11].

In recent years, many research works have considered various methods for solving (L_0, L_1)-smooth optimization problems. However, despite the interest from the scientific community, the existing results on the convergence of methods remain nonoptimal in several important cases, and the theoretical analysis is often insufficiently complete. While the most general class of nonconvex problems, admitting only the convergence of metrics like the gradient norm to zero, has been the focus of [1] and the later publications [10–16], the class of (L_0, L_1)-smooth convex functions has been investigated much less until very recent time (the late 2024–early 2025). In particular, when applied to (L_0, L_1)-smooth convex problems, known results on the convergence of methods, such as gradient descent with clipping [17] and gradient descent with Polyak's step size [18], either additionally assume L-smoothness [19, 20] or restrict the step size to make the method's trajectory bounded in a domain where the gradient satisfies the Lipschitz condition due to the (L_0, L_1)-smoothness of the objective function [21].

Let us review several recent works devoted to the optimization of (L_0, L_1)-smooth functions.

In [22], the class of strongly convex (L_0, L_1)-smooth functions was studied, and new convergence rate guarantees for several existing methods were obtained. In particular, better convergence rate estimates were derived for gradient descent with (smoothed) clipping and for gradient descent with Polyak's step size. Unlike the existing results, these estimates neither involve the standard smoothness assumption nor have an exponential dependence on the distance from the initial point to the solution.

Parallel to [22], similar results for the gradient method with various step size strategies, including Polyak's step, were independently obtained in [23], along with new complexity estimates for the gradient method with normalized step on the class of (L_0, L_1)-smooth problems.

A more detailed analysis of the convergence of gradient descent and its modifications under generalized (L_0, L_1)-smoothness was presented in [24]. In the convex case, it was shown that the convergence rate of the gradient method changes depending on the gradient norm: for $\|\nabla f(x_k)\| \geqslant \frac{L_0}{L_1}$, a linear rate is achieved, and a standard sublinear convergence rate is observed if $\|\nabla f(x_k)\| < \frac{L_0}{L_1}$.

Unlike the above-mentioned works, here we analyze the gradient method for (L_0, L_1)-smooth problems under the condition that only approximate (additively inexact) gradient values are available at each iteration of the method. That is, this research further develops, in a new direction, the paper [25], where similar problems were investigated for smooth problems in the classical sense. Note that several directions are considered. First, we study the convergence of the method for ρ-quasar-convex, with respect to any exact solution [26–28], (L_0, L_1)-smooth functions satisfying the Polyak–Łojasiewicz (PL) condition

$$f(x) - f(x_*) \leqslant \frac{1}{2\mu}\|\nabla f(x)\|^2 \quad \forall x \in \mathbb{R}^n,\ \mu > 0, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm and x_* is a minimizer of the function f. As is shown below, under some assumptions, the method converges with a linear rate to some neighborhood of the exact

solution. It is also proven that starting from some iteration, the method converges on the class of problems with the PL condition (without requiring quasar-convexity). Second, we separately consider the class of ρ-quasar-convex $(0, M)$-smooth functions, representing interest in the context of machine learning problems (particularly when training logistic regression models). For this class of problems, an adaptive modification of the gradient method using inexact gradients at iterations is proposed, and convergence rate estimates are obtained under some constraint on the quasar-convexity parameter ρ. The efficiency of the proposed methods is confirmed by computational experiments for logistic regression problems (which satisfy the PL condition on any compact set [29]) and a certain nonconvex quasar-convex problem [27].

2. PROBLEM STATEMENT AND NECESSARY BACKGROUND

Consider minimization problems of the form

$$\min_{x\in\mathbb{R}^n} f(x), \tag{2}$$

where $f:\mathbb{R}^n \longrightarrow \mathbb{R}$ is a convex and (L_0, L_1)-smooth function. Throughout this paper, x_* and $f^* = f(x_*)$ denote a global minimizer and the global minimum value, respectively, of the function f; the norm $\|\cdot\|$ is Euclidean.

As already noted, the (L_0, L_1)-smoothness condition $(L_0, L_1 > 0)$ was initially introduced in [1] for twice differentiable functions $f:\mathbb{R}^n \longrightarrow \mathbb{R}$ as follows:

$$\|\nabla^2 f(x)\| \leqslant L_0 + L_1\|\nabla f(x)\| \quad \forall x \in \mathbb{R}^n, \tag{3}$$

where $\|\cdot\|$ indicates the Euclidean norm for vectors and the spectral norm for matrices.

In [10], this concept was extended to the class of differentiable, but not necessarily twice differentiable, functions.

Definition 1. A function $f:\mathbb{R}^n \longrightarrow \mathbb{R}$ is said to be (L_0, L_1)-smooth if, for any $x, y \in \mathbb{R}^n$ such that $\|y - x\| \leqslant \frac{1}{L_1}$,

$$\|\nabla f(x) - \nabla f(y)\| \leqslant (L_0 + L_1\|\nabla f(y)\|)\,\|x - y\|. \tag{4}$$

Consider several examples of functions satisfying condition (4).

Example 1. The function $f(x) = \|x\|^{2m}$ is $(2m, 2m-1)$-smooth. At the same time, for all $L \geqslant 0$, f is not L-smooth for $m \geqslant 2$; for details, see [22]. (Recall that an L-smooth function is one whose gradient satisfies the Lipschitz condition.)

Example 2. The function $f(x) = e^{ax}$, $a \in \mathbb{R}^n$, is $(0, \|a\|)$-smooth. At the same time, for all $L \geqslant 0$, $f(x)$ is not L-smooth for $a \neq 0$ [22].

Example 3. The logistic function $f(x) = \log(1 + \exp(-a^T x))$, where $a \in \mathbb{R}^n$, is (L_0, L_1)-smooth with $L_0 = 0$ and $L_1 = \|a\|$ [22].

The concept of (L_0, L_1)-smoothness was generalized in [11] as follows: a function $f:\mathbb{R}^n \longrightarrow \mathbb{R}$ is said to be α-symmetrically generalized smooth if, for some $\alpha \in [0, 1]$ and any $x, y \in \mathbb{R}^n$,

$$\|\nabla f(y) - \nabla f(x)\| \leqslant \left(L_0 + L_1 \max_{\theta\in[0,1]} \|\nabla f(\theta y + (1-\theta)x)\|^\alpha\right)\|y - x\|. \tag{5}$$

As demonstrated in [11], condition (5) is more general than (3) and (4): the validity of (3) or (4) implies the validity of (5) (with $\alpha = 1$). Moreover, for twice differentiable functions, conditions (3) and (5) are equivalent for $\alpha = 1$ [11].

Now we give the definition of a quasar-convex function, a key in this work.

Definition 2. Let $\rho \in (0, 1]$ and $\mu > 0$. A function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is said to be (ρ, μ)-quasar strongly convex if, for all $x \in \mathbb{R}^n$,

$$f(x_*) \geqslant f(x) + \frac{1}{\rho}\langle \nabla f(x), x_* - x\rangle + \frac{\mu}{2}\|x_* - x\|^2, \tag{6}$$

where x_* is a global minimizer of the function f. For $\mu = 0$ the function f is called a ρ-quasar convex function.

The PL condition can be treated as a relaxation of ordinary strong convexity and, moreover, of strong quasar-convexity [27]: if $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a (ρ, μ)-quasar strongly convex function, then it satisfies the PL condition (1) with the constant $\mu\rho^2$.

Example 4 [27]. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be defined as

$$f(x) = \|x\|^2 g\left(\frac{x}{\|x\|}\right), \quad \forall x \in \mathbb{R}^n,$$

where

$$g(x) = 1 + \sum_{i=1}^{n} a_i \sin(b_i x_i)^2, \quad \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Then the function f is $(1, 2)$-quasar strongly convex, i.e., satisfies the PL-condition (1) with constant 2.

The issues of convergence of the gradient method for convex problems with the generalized smoothness condition (4) have been actively investigated recently; for example, see [22–24]. The cited research works presented results on the convergence of the normalized gradient method, the gradient method with Polyak's step, and the accelerated gradient method based on the assumption of generalized smoothness. It was also shown that the gradient method converges with a linear rate as long as the gradient norms at the iterations remain sufficiently large. This paper is a continuation and further development of the above works.

Here, we focus on studying gradient descent for the minimization problems (2) under the condition that, at each iteration of the method, only a gradient with an additive error is available instead of exact gradient values. Recall that for $\Delta > 0$, a vector $\widetilde{\nabla} f(x)$ is called the inexact gradient of a function f at a point x if

$$\|\nabla f(x) - \widetilde{\nabla} f(x)\| \leqslant \Delta;$$

and $\Delta > 0$ is called the gradient inexactness parameter.

In the case of small gradient norms at the method's iterations, it becomes challenging to derive theoretical estimates of the convergence rate. The reason is that, unlike the case of using exact gradients, it becomes impossible to ensure a monotonic decrease of the gradient norm at each iteration. This problem can be circumvented by additionally requiring that the objective function f satisfies the PL condition. According to [25], for problems with inexact gradients, the PL condition implies the inequality

$$\|\widetilde{\nabla} f(x)\|^2 \geqslant \mu(f(x) - f^*) - \Delta^2 \quad \forall x \in \mathbb{R}^n,\ \mu > 0. \tag{7}$$

This section provides the problem statement and necessary background. In Section 3, we study the gradient method for (L_0, L_1)-smooth problems under an additive error in the gradient at each iteration. The method's convergence rate on the class of ρ-quasar-convex functions satisfying the PL condition is estimated. As established, the convergence rate is linear up to a certain level of gradient values. Section 4 deals with an adaptive modification of the gradient method for $(0, M)$-smooth ρ-quasar-convex problems. In Section 5, we present the results of computational experiments for the logistic regression problem (which satisfies the PL condition on any compact set [29]) and a certain nonconvex quasar-convex problem [27].

3. GRADIENT METHOD FOR (L_0, L_1)-SMOOTH ρ-QUASAR CONVEX FUNCTIONS

Let $f : \mathbb{R}^n \to \mathbb{R}$ be an (L_0, L_1)-smooth (4) and ρ-quasar convex function with respect to any solution x_* (see Definition 2, $\mu = 0$):

$$f(x_*) \geqslant f(x) + \frac{1}{\rho}\langle \nabla f(x), x_* - x\rangle, \quad \rho \in (0, 1]. \tag{8}$$

If f is strongly quasar-convex (see Definition 2, $\mu > 0$), then the solution x_* is unique and f is strongly quasar-convex with respect to any solution. Note that problems with nonconvex smooth (hence, (L_0, L_1)-smooth) ρ-quasar-convex functions are often encountered in applications (see Example 4) [27].

To solve problem (2), we investigate the gradient method of the form

$$x_{k+1} = x_k - \eta_k \widetilde{\nabla} f(x_k) \tag{9}$$

with the step size

$$\eta_k = \frac{\alpha}{L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)}, \tag{10}$$

where $\alpha \in (0, 1]$ and $\widetilde{\nabla} f(x_k)$ denotes the inexact gradient of the function f at a point x_k.

Note that, despite the use of approximate gradient values in the algorithm, the theoretical analysis assumes the existence of the exact gradient.

Let us prove several lemmas necessary for deriving convergence rate estimates of the method.

Lemma 1 [10, Lemma A3]. *Let a function $f(x)$ be (L_0, L_1)-smooth (4). Then, for all $x, y \in \mathbb{R}^n$ such that $\|x - y\| \leqslant \frac{1}{L_1}$,*

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_0 + L_1\|\nabla f(x)\|}{2}\|y - x\|^2. \tag{11}$$

Proof. Let $g(t) = f(x + t(y - x))$ for $t \in [0, 1]$. Then

$$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(t)dt = \int_0^1 \langle \nabla f(x + t(y - x)), y - x\rangle\, dt.$$

We add and subtract $\langle \nabla f(x), y - x\rangle$ to/from the integrand:

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x), y - x\rangle\, dt + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x\rangle\, dt.$$

Computing the first integral and applying the Cauchy–Bunyakovsky–Schwarz inequality to the expression under the second integral, we obtain

$$f(y) - f(x) \leqslant \langle \nabla f(x), y - x\rangle + \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|\|y - x\|dt.$$

In view of

$$\|\nabla f(y) - \nabla f(x)\| \leqslant (L_0 + L_1\|\nabla f(x)\|)\, \|y - x\|,$$

we have the inequality

$$f(y) - f(x) \leqslant \langle \nabla f(x), y - x\rangle + \|y - x\| \int\limits_0^1 (L_0 + L_1\|\nabla f(x)\|)\, \|y - x\| t dt$$

and, consequently,

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_0 + L_1\|\nabla f(x)\|}{2}\|y - x\|^2.$$

Remark 1. Under an additive error in the gradient, inequality (11) takes the form

$$f(y) \leqslant f(x) + \langle \widetilde{\nabla} f(x), y - x\rangle + \frac{L_0 + L_1(\|\widetilde{\nabla} f(x)\| + \Delta)}{2}\|y - x\|^2 + \Delta\|y - x\|. \tag{12}$$

Lemma 2. *Let $f(x)$ be an (L_0, L_1)-smooth function* (4)*. Assume that $\|\nabla f(x_k)\| > \Delta$. Then at the iterations of algorithm* (9) *with the step size* (10)*, the values of this function decrease monotonically, $f(x_{k+1}) \leqslant f(x_k)$, and*

$$f(x_k) - f(x_{k+1}) \geqslant \frac{\alpha(2-\alpha)}{2}\frac{(\|\nabla f(x_k)\| - \Delta)^2}{L_0 + L_1\|\nabla f(x_k)\| + 2L_1\Delta}. \tag{13}$$

Proof. Consider inequality (11) with $y = x_{k+1}$ and $x = x_k$. Note that $\|x_{k+1} - x_k\| \leqslant \frac{1}{L_1}$. Then we have

$$\begin{aligned}
&f(x_{k+1}) - f(x_k)\\
&\leqslant \frac{-\alpha\langle \nabla f(x_k), \widetilde{\nabla} f(x_k)\rangle}{L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} + \frac{L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)}{2\left(L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)\right)^2}\alpha^2\|\widetilde{\nabla} f(x_k)\|^2\\
&= \frac{\alpha^2\|\widetilde{\nabla} f(x_k)\|^2 - 2\alpha\langle \nabla f(x_k), \widetilde{\nabla} f(x_k)\rangle}{2\left(L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)\right)}\\
&= \frac{\alpha^2\|\widetilde{\nabla} f(x_k)\|^2 - 2\alpha\langle \nabla f(x_k), \widetilde{\nabla} f(x_k)\rangle + \|\nabla f(x_k)\|^2 - \|\nabla f(x_k)\|^2}{2\left(L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)\right)}\\
&= \frac{\|\alpha\widetilde{\nabla} f(x_k) - \nabla f(x_k)\|^2 - \|\nabla f(x_k)\|^2}{2\left(L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)\right)}.
\end{aligned}$$

Therefore,

$$f(x_k) - f(x_{k+1}) \geqslant \frac{\|\nabla f(x_k)\|^2 - \|\alpha\widetilde{\nabla} f(x_k) - \nabla f(x_k)\|^2}{2\left(L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)\right)}.$$

We utilize the following relation for $\alpha \in (0, 1]$ and $a, b \in \mathbb{R}^n$:

$$\|\alpha a - b\| = \|\alpha(a - b) + (\alpha - 1)b\| \leqslant \alpha\|a - b\| + (1 - \alpha)\|b\|.$$

As a result,

$$
\begin{aligned}
&f(x_k) - f(x_{k+1}) \\
&\geqslant \frac{\|\nabla f(x_k)\|^2 - \left(\alpha\|\widetilde{\nabla} f(x_k) - \nabla f(x_k)\| + (1-\alpha)\|\nabla f(x_k)\|\right)^2}{2\left(L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)\right)} \\
&\geqslant \frac{\|\nabla f(x_k)\|^2 - (\alpha\Delta + (1-\alpha)\|\nabla f(x_k)\|)^2}{2\left(L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)\right)} \\
&= \frac{\alpha\left(\|\nabla f(x_k)\| - \Delta\right)\left((2-\alpha)(\|\nabla f(x_k)\| - \Delta) + 2\Delta\right)}{2\left(L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)\right)} \\
&> \frac{\alpha(2-\alpha)}{2}\frac{(\|\nabla f(x_k)\| - \Delta)^2}{L_0 + L_1\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} \\
&\geqslant \frac{\alpha(2-\alpha)}{2}\frac{(\|\nabla f(x_k)\| - \Delta)^2}{L_0 + L_1\|\nabla f(x_k)\| + 2L_1\Delta}.
\end{aligned}
$$

Lemma 3. *Let f be an (L_0, L_1)-smooth function* (4) *satisfying the PL condition* (1)*, and let x_* be the exact solution of problem* (2) *closest to x_{k+1}. Then the following inequality holds for the gradient method* (9) *with the step size* (10)*:*

$$
\|x_{k+1} - x_*\| \leqslant \sqrt{\frac{2}{\mu}\left(f(x_0) - f^*\right)}. \tag{14}
$$

Proof. According to [29, Appendix A], for the closest exact solution x_*, the PL condition implies the quadratic growth of the function:

$$
\frac{\mu}{2}\|x_{k+1} - x_*\|^2 \leqslant f(x_{k+1}) - f^*.
$$

Due to the monotonicity of the function f, the inequality becomes

$$
\frac{\mu}{2}\|x_{k+1} - x_*\|^2 \leqslant f(x_0) - f^*,
$$

and the desired inequality is immediate:

$$
\|x_{k+1} - x_*\| \leqslant \sqrt{\frac{2}{\mu}\left(f(x_0) - f^*\right)}.
$$

Let us denote $g_k = \|\nabla f(x_k)\| - \Delta$. From Lemma 2 it follows that

$$
f(x_k) - f(x_{k+1}) > \frac{\alpha(2-\alpha)}{2}\frac{g_k^2}{L_0 + L_1 g_k + 3L_1\Delta}. \tag{15}
$$

Assume that $g_k \geqslant 3\Delta$. We will investigate the convergence of the method depending on the value of g_k. Two cases are possible: $g_k > \frac{L_0}{L_1}$ and $g_k \leqslant \frac{L_0}{L_1}$.

For $g_k > \frac{L_0}{L_1}$, from (15) we have

$$
f(x_k) - f(x_{k+1}) > \frac{\alpha(2-\alpha)}{2}\frac{g_k^2}{L_1 g_k + L_1 g_k + L_1 g_k} > \frac{\alpha(2-\alpha)}{2}\frac{g_k}{3L_1} = \frac{\alpha(2-\alpha)}{6L_1}g_k. \tag{16}
$$

If $g_k \leqslant \frac{L_0}{L_1}$, then inequality (15) gives

$$f(x_k) - f(x_{k+1}) > \frac{\alpha(2-\alpha)}{2}\frac{g_k^2}{L_0 + L_0 + L_0} > \frac{\alpha(2-\alpha)}{2}\frac{g_k^2}{3L_0} = \frac{\alpha(2-\alpha)}{6L_0}g_k^2. \tag{17}$$

The following result is true for this class of problems.

Theorem 1. *Let f be a ρ-quasar-convex function with respect to each exact solution and, simultaneously, an (L_0, L_1)-smooth function satisfying the PL condition* (1). *Also, let $\|\widetilde{\nabla} f(x_k)\| \geqslant 5\Delta$ and $\mu < \min\left\{\frac{6L_0}{\alpha(2-\alpha)}; \frac{72L_1^2(f(x_0)-f^*)}{\alpha^2(2-\alpha)^2\rho^2}\right\}$. Then the following inequality holds for algorithm* (9) *with the step size* (10):

$$\begin{aligned} f(x_{k+1}) - f^* \leqslant \max\left\{1 - \frac{\alpha(2-\alpha)\rho}{6L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}}; 1 - \frac{\alpha(2-\alpha)\mu}{6L_0}\right\}^{k+1}(f(x_0)-f^*) \\ + \Delta \max\left\{\frac{1}{L_1}; \frac{\Delta}{L_0}\right\}\max\left\{\frac{L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}}{\rho}; \frac{L_0}{\mu}\right\}. \end{aligned} \tag{18}$$

Proof. Consider first the case $g_k > \frac{L_0}{L_1}$. By the quasar-convexity of the function $f(x)$ at $x = x_k$, we have the inequality

$$\begin{aligned} f(x_k) - f^* &\leqslant \frac{1}{\rho}\langle \nabla f(x_k), x_k - x_*\rangle \leqslant \frac{1}{\rho}\|\nabla f(x_k)\|\|x_k - x_*\| \\ &= \frac{1}{\rho}(g_k + \Delta)\|x_k - x_*\|. \end{aligned} \tag{19}$$

Applying Lemma 3 to (19) yields

$$f(x_k) - f^* \leqslant \frac{1}{\rho}\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}(g_k + \Delta)$$

and, consequently,

$$g_k \geqslant \frac{\rho(f(x_k)-f^*)}{\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}} - \Delta. \tag{20}$$

Inequalities (16) and (20) imply the estimate

$$\begin{aligned} f(x_{k+1}) - f^* &\leqslant f(x_k) - f^* - \frac{\alpha(2-\alpha)}{6L_1}\left(\frac{\rho(f(x_k)-f^*)}{\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}} - \Delta\right) \\ &= \left(1 - \frac{\alpha(2-\alpha)\rho}{6L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}}\right)(f(x_k)-f^*) + \frac{\alpha(2-\alpha)\Delta}{6L_1}. \end{aligned} \tag{21}$$

For $g_k \leqslant \frac{L_0}{L_1}$, from (17) it follows that

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leqslant -\frac{\alpha(2-\alpha)}{6L_0}g_k^2 = -\frac{\alpha(2-\alpha)}{6L_0}(\|\nabla f(x_k)\| - \Delta)^2 \\ &\leqslant -\frac{\alpha(2-\alpha)}{6L_0}\left(\frac{1}{2}\|\nabla f(x_k)\|^2 - \Delta^2\right) \\ &= -\frac{\alpha(2-\alpha)}{12L_0}\|\nabla f(x_k)\|^2 + \frac{\alpha(2-\alpha)\Delta^2}{6L_0} \\ &= -\frac{\alpha(2-\alpha)\mu}{6L_0}(f(x_k)-f^*) + \frac{\alpha(2-\alpha)\Delta^2}{6L_0}. \end{aligned}$$

Thus,

$$f(x_{k+1})-f^*\leqslant\left(1-\frac{\alpha(2-\alpha)\mu}{6L_0}\right)(f(x_k)-f^*)+\frac{\alpha(2-\alpha)\Delta^2}{6L_0}. \tag{22}$$

Due to inequalities (21) and (22), we obtain

$$f(x_{k+1})-f^*\leqslant\max\left\{1-\frac{\alpha(2-\alpha)\rho}{6L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}};1-\frac{\alpha(2-\alpha)\mu}{6L_0}\right\}(f(x_k)-f^*)$$
$$+\max\left\{\frac{\alpha(2-\alpha)\Delta}{6L_1};\frac{\alpha(2-\alpha)\Delta^2}{6L_0}\right\},$$

or

$$f(x_{k+1})-f^*\leqslant\max\left\{1-\frac{\alpha(2-\alpha)\rho}{6L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}};1-\frac{\alpha(2-\alpha)\mu}{6L_0}\right\}(f(x_k)-f^*)$$
$$+\frac{\alpha(2-\alpha)\Delta}{6}\max\left\{\frac{1}{L_1};\frac{\Delta}{L_0}\right\}$$
$$\leqslant\max\left\{1-\frac{\alpha(2-\alpha)\rho}{6L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}};1-\frac{\alpha(2-\alpha)\mu}{6L_0}\right\}^{k+1}(f(x_0)-f^*)$$
$$+\frac{\alpha(2-\alpha)\Delta}{6}\max\left\{\frac{1}{L_1};\frac{\Delta}{L_0}\right\}\sum_{i=0}^{k}\max\left\{1-\frac{\alpha(2-\alpha)\rho}{6L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}};1-\frac{\alpha(2-\alpha)\mu}{6L_0}\right\}^{i}$$
$$\leqslant\max\left\{1-\frac{\alpha(2-\alpha)\rho}{6L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}};1-\frac{\alpha(2-\alpha)\mu}{6L_0}\right\}^{k+1}(f(x_0)-f^*)$$
$$+\frac{\alpha(2-\alpha)\Delta}{6}\max\left\{\frac{1}{L_1};\frac{\Delta}{L_0}\right\}\frac{1}{1-\max\left\{1-\frac{\alpha(2-\alpha)\rho}{6L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}};1-\frac{\alpha(2-\alpha)\mu}{6L_0}\right\}}.$$

The following chain of inequalities is true for positive numbers a and b:

$$\frac{1}{1-\max\{1-a;1-b\}}=\frac{1}{1-(1-\min\{a;b\})}=\frac{1}{\min\{a;b\}}=\max\left\{\frac{1}{a};\frac{1}{b}\right\}.$$

Hence,

$$f(x_{k+1})-f^*\leqslant\max\left\{1-\frac{\alpha(2-\alpha)\rho}{6L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}};1-\frac{\alpha(2-\alpha)\mu}{6L_0}\right\}^{k+1}(f(x_0)-f^*)$$
$$+\frac{\alpha(2-\alpha)\Delta}{6}\max\left\{\frac{1}{L_1};\frac{\Delta}{L_0}\right\}\frac{6}{\alpha(2-\alpha)}\max\left\{\frac{L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}}{\rho};\frac{L_0}{\mu}\right\}$$
$$\leqslant\max\left\{1-\frac{\alpha(2-\alpha)\rho}{6L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}};1-\frac{\alpha(2-\alpha)\mu}{6L_0}\right\}^{k+1}(f(x_0)-f^*)$$
$$+\Delta\max\left\{\frac{1}{L_1};\frac{\Delta}{L_0}\right\}\max\left\{\frac{L_1\sqrt{\frac{2}{\mu}(f(x_0)-f^*)}}{\rho};\frac{L_0}{\mu}\right\}.$$

Remark 2. For algorithm (9), the number of iterations with $\|\widetilde{\nabla} f(x_k)\| \geqslant 5\Delta$ can be estimated as

$$N \leqslant \left\lceil \frac{6L_1^2(f(x_0)-f^*)}{\alpha(2-\alpha)L_0} \right\rceil + \left\lceil \frac{2L_0(f(x_0)-f^*)}{3\alpha(2-\alpha)\Delta^2} \right\rceil.$$

Proof. Let us estimate the number of iterations where $g_k > \frac{L_0}{L_1}$. From (16) it follows that

$$f(x_k) - f(x_{k+1}) > \frac{\alpha(2-\alpha)L_0}{6L_1^2}.$$

Let $i_1, \ldots, i_m$, $p = \overline{1,m}$, be the iterations for which $g_{i_p} > \frac{L_0}{L_1}$. Then summation yields the inequality

$$f(x_0) - f^* \geqslant f(x_{i_1}) - f(x_{i_m}) \geqslant \frac{\alpha(2-\alpha)L_0}{6L_1^2} m$$

and, consequently,

$$m \leqslant \left\lceil \frac{6L_1^2(f(x_0)-f^*)}{\alpha(2-\alpha)L_0} \right\rceil, \quad L_0 \neq 0.$$

Now we estimate the number of iterations for which $g_k \leqslant \frac{L_0}{L_1}$. In view of $g_k^2 \geqslant 9\Delta^2$, inequality (17) leads to the estimate $f(x_k) - f(x_{k+1}) > \frac{9\alpha(2-\alpha)\Delta^2}{6L_0}$. Let $j_1, \ldots, j_n$, $l = \overline{1,n}$, be the iterations for which $g_{j_l} \leqslant \frac{L_0}{L_1}$. Then summation yields the inequality

$$f(x_0) - f^* \geqslant f(x_{j_1}) - f(x_{j_n}) \geqslant \frac{9\alpha(2-\alpha)\Delta^2}{6L_0} n,$$

and the number of such iterations satisfies the upper bound

$$n \leqslant \left\lceil \frac{2L_0(f(x_0)-f^*)}{3\alpha(2-\alpha)\Delta^2} \right\rceil.$$

Remark 3. According to inequality (21), the method (9) with the step size (10) converges with a rate close to that of a geometric progression for sufficiently large gradient norms (i.e., $g_k > \frac{L_0}{L_1}$).

Remark 4. Suppose that an acceptable solution quality is achieved under the condition $\|\widetilde{\nabla} f(x_k)\| < 5\Delta$. Indeed, if $\|\widetilde{\nabla} f(x_k)\| < 5\Delta$, then $\|\nabla f(x_k)\| < 6\Delta$, and the PL condition (1) gives

$$f(x_k) - f^* \leqslant \frac{1}{2\mu}\|\nabla f(x)\|^2 < \frac{18\Delta^2}{\mu}.$$

Remark 5. Since the number of iterations with $g_k > \frac{L_0}{L_1}$ is limited, starting from some iteration N the gradient norm will be less than $\frac{L_0}{L_1}$. Then it can be stated that the method (9) with the step size (10) converges on the class of problems with the PL condition (without requiring ρ-quasar-convexity) starting from this number. Thus, starting from some iteration, we obtain:

1) either

$$f(x_{k+1}) - f^* \leqslant \left(1 - \frac{\alpha(2-\alpha)\mu}{6L_0}\right)^k (f(x_N) - f^*) + \frac{\Delta^2}{\mu}$$

from inequality (22) by recursion,

2) or an acceptable accuracy has already been achieved.

4. ADAPTIVE GRADIENT METHOD FOR QUASAR-CONVEX $(0, M)$-SMOOTH FUNCTIONS

Consider the case $L_0 = 0$, corresponding, e.g., to the logistic regression problem. Denoting $M = L_1$ for convenience, we write condition (4) as

$$\|\nabla f(y)-\nabla f(x)\| \leqslant M\|\nabla f(x)\|\|y-x\| \ \forall x, y \in \mathbb{R}^n \text{ such that } \|y-x\| \leqslant \frac{1}{M}. \tag{23}$$

In this section, an adaptive modification of the gradient method using inexact gradient information will be investigated on the class of ρ-quasar convex functions.

Let f be a $(0, M)$-smooth and ρ-quasar-convex function with $\rho \in \left(\frac{1}{2}, 1\right]$. (Unlike the previous section, this condition holds not necessarily with respect to each exact solution.) Consider the following adaptive modification of the gradient method using inexact gradient values at the iterations.

Algorithm 1 (the adaptive gradient method for $(0, M)$-smooth problems).

1. Input: $x_0 \in \mathbb{R}^n$, $\Delta > 0$, $0 < M_0 < 2M$, and $\alpha > 0$.
2. $k = 0$.
3. Repeat:
4. $\quad M_{k+1} = \max\{M_0/2, M_k/2\}$,
5. $\quad \eta_k = \frac{\alpha}{M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\|+\Delta\right)}$,
6. $\quad x_{k+1} = x_k - \eta_k \widetilde{\nabla} f(x_k)$,
7. $\quad$ If $f(x_k) - f(x_{k+1}) \geqslant \frac{\alpha(2-\alpha)}{4M_{k+1}}\left(\|\widetilde{\nabla} f(x_k)\| - 2\Delta\right)$ and
$$f(x_k) - f(x_{k+1}) \geqslant \alpha\left(1 - \tfrac{\alpha}{2}\right)\frac{\|\widetilde{\nabla} f(x_k)\|^2}{M_{k+1}(\|\widetilde{\nabla} f(x_k)\|+\Delta)} - \frac{\Delta\alpha}{M_{k+1}},$$
8. $\quad$ then $\quad k := k + 1$;
9. $\quad$ otherwise,
10. $\quad\quad M_{k+1} = 2M_{k+1}$ and get back to line 5.
11. Output: $\widehat{x} = x_{k+1}$.

Now we prove several auxiliary statements for $(0, M)$-smooth functions. In particular, they will justify the correctness of Algorithm 1, i.e., the validity of the iteration exit criterion (see line 7 of Algorithm 1) for $(0, M)$-smooth functions for sufficiently large M_{k+1}.

Lemma 4. *Let $f(x)$ be a $(0, M)$-smooth function. Assume that $\|\widetilde{\nabla} f(x_k)\| > 5\Delta$ and $\{M_i\}_{i=0}^k$ is a sequence such that $M_{i+1} \geqslant M$. Then at the iterations of the algorithm*

$$x_{k+1} = x_k - \frac{\alpha}{M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)}\widetilde{\nabla} f(x_k), \quad \alpha \in (0, 1],$$

the values of this function are monotonically nonincreasing, $f(x_{k+1}) \leqslant f(x_k)$, and

$$f(x_k) - f(x_{k+1}) \geqslant \frac{\alpha(2-\alpha)}{2}\frac{(\|\widetilde{\nabla} f(x_k)\| - 2\Delta)^2}{M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)}. \tag{24}$$

Proof. For $L_0 = 0$ and $L_1 = M$, inequality (11) implies

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

With $y = x_{k+1}$ and $x = x_k$, we obtain

$$
\begin{aligned}
&f(x_{k+1}) - f(x_k) \\
&\leqslant \frac{-\alpha\langle\nabla f(x_k), \widetilde{\nabla} f(x_k)\rangle}{M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} + \frac{M\|\nabla f(x_k)\|}{2M_{k+1}^2\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)^2}\alpha^2\|\widetilde{\nabla} f(x_k)\|^2 \\
&\leqslant \frac{-\alpha\langle\nabla f(x_k), \widetilde{\nabla} f(x_k)\rangle}{M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} + \frac{M\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)}{2M\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)}\frac{\alpha^2\|\widetilde{\nabla} f(x_k)\|^2}{M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} \\
&= \frac{\alpha^2\|\widetilde{\nabla} f(x_k)\|^2 - 2\alpha\langle\nabla f(x_k), \widetilde{\nabla} f(x_k)\rangle}{2M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} \\
&= \frac{\alpha^2\|\widetilde{\nabla} f(x_k)\|^2 - 2\alpha\langle\nabla f(x_k), \widetilde{\nabla} f(x_k)\rangle + \|\nabla f(x_k)\|^2 - \|\nabla f(x_k)\|^2}{2M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} \\
&= \frac{\|\alpha\widetilde{\nabla} f(x_k) - \nabla f(x_k)\|^2 - \|\nabla f(x_k)\|^2}{2M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)}.
\end{aligned}
$$

Therefore,

$$
f(x_k) - f(x_{k+1}) \geqslant \frac{\|\nabla f(x_k)\|^2 - \|\alpha\widetilde{\nabla} f(x_k) - \nabla f(x_k)\|^2}{2M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)}.
$$

We take advantage of the following relation for $\alpha \in (0, 1]$ and $a, b \in \mathbb{R}^n$:

$$
\|\alpha a - b\| = \|\alpha(a - b) + (\alpha - 1)b\| \leqslant \alpha\|a - b\| + (1 - \alpha)\|b\|.
$$

As a result,

$$
\begin{aligned}
&f(x_k) - f(x_{k+1}) \\
&\geqslant \frac{\|\nabla f(x_k)\|^2 - \left(\alpha\|\widetilde{\nabla} f(x_k) - \nabla f(x_k)\| + (1 - \alpha)\|\nabla f(x_k)\|\right)^2}{2M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} \\
&\geqslant \frac{\|\nabla f(x_k)\|^2 - \left(\alpha\Delta + (1 - \alpha)\|\nabla f(x_k)\|\right)^2}{2M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} \\
&= \frac{\alpha\left(\|\nabla f(x_k)\| - \Delta\right)\left((2 - \alpha)(\|\nabla f(x_k)\| - \Delta) + 2\Delta\right)}{2M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} \\
&> \frac{\alpha(2 - \alpha)}{2}\frac{\left(\|\nabla f(x_k)\| - \Delta\right)^2}{M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} \\
&\geqslant \frac{\alpha(2 - \alpha)}{2}\frac{\left(\|\widetilde{\nabla} f(x_k)\| - 2\Delta\right)^2}{M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)}.
\end{aligned}
$$

Let $\{M_i\}_{i=0}^k$ be a sequence of positive numbers satisfying the assumptions of the previous theorem. Given $\|\widetilde{\nabla} f(x_k)\| \geqslant 5\Delta$, from (24) it follows that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geqslant \frac{\alpha(2-\alpha)(\|\widetilde{\nabla} f(x_k)\| - 2\Delta)^2}{2\left(M_{k+1}(\|\widetilde{\nabla} f(x_k)\| - 2\Delta) + 3M_{k+1}\Delta\right)} \\ &\geqslant \frac{\alpha(2-\alpha)}{4M_{k+1}}\left(\|\widetilde{\nabla} f(x_k)\| - 2\Delta\right). \end{aligned} \tag{25}$$

Lemma 5. *Let $f(x)$ be a $(0, M)$-smooth function, and let $\{M_i\}_{i=0}^k$ be some sequence such that $M_{i+1} \geqslant M$. Then for the algorithm*

$$x_{k+1} = x_k - \frac{\alpha}{M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)}\widetilde{\nabla} f(x_k), \quad \alpha \in (0, 1],$$

the following inequality holds:

$$\frac{\|\widetilde{\nabla} f(x_k)\|^2}{M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)} \leqslant \frac{2}{\alpha(2-\alpha)}(f(x_k) - f(x_{k+1})) + \frac{2\Delta}{M_{k+1}(2-\alpha)}. \tag{26}$$

Proof. For $L_0 = 0$ and $L_1 = M$, inequality (12) gives

$$f(y) \leqslant f(x) + \langle \widetilde{\nabla} f(x), y - x\rangle + \frac{M(\|\widetilde{\nabla} f(x)\| + \Delta)}{2}\|y - x\|^2 + \Delta\|y - x\|.$$

With $y = x_{k+1}$ and $x = x_k$, we have

$$\begin{aligned} f(x_{k+1}) &\leqslant f(x_k) + \langle \widetilde{\nabla} f(x_k), x_{k+1} - x_k\rangle \\ &\quad + \frac{M(\|\widetilde{\nabla} f(x_k)\| + \Delta)}{2}\|x_{k+1} - x_k\|^2 + \Delta\|x_{k+1} - x_k\| \\ &\leqslant f(x_k) - \frac{\alpha\|\widetilde{\nabla} f(x_k)\|^2}{M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)} + \frac{\alpha^2\|\widetilde{\nabla} f(x_k)\|^2}{2M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)} \\ &\quad + \frac{\Delta\alpha\|\widetilde{\nabla} f(x_k)\|}{M_{k+1}(\|\widetilde{\nabla} f(x)\| + \Delta)} \\ &= f(x_k) - \alpha\left(1 - \frac{\alpha}{2}\right)\frac{\|\widetilde{\nabla} f(x_k)\|^2}{M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)} + \frac{\Delta\alpha\|\widetilde{\nabla} f(x_k)\|}{M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)}. \end{aligned}$$

Let us estimate the term containing Δ :

$$\frac{\Delta\alpha\|\widetilde{\nabla} f(x_k)\|}{M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)} \leqslant \frac{\frac{\Delta\alpha}{M_{k+1}}M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)}{M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)} = \frac{\Delta\alpha}{M_{k+1}}.$$

Thus,

$$f(x_k) - f(x_{k+1}) \geqslant \alpha\left(1 - \frac{\alpha}{2}\right)\frac{\|\widetilde{\nabla} f(x_k)\|^2}{M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)} - \frac{\Delta\alpha}{M_{k+1}},$$

which can be equivalently written as

$$\frac{\|\widetilde{\nabla} f(x_k)\|^2}{M_{k+1}(\|\widetilde{\nabla} f(x_k)\| + \Delta)} \leqslant \frac{2}{\alpha(2-\alpha)}(f(x_k) - f(x_{k+1})) + \frac{2\Delta}{M_{k+1}(2-\alpha)}.$$

Inequalities (25) and (26) mean that for $M_{k+1} \geqslant M$, the iteration exit criterion of Algorithm 1 (see line 7) is automatically satisfied.

Lemma 6. *Let $f(x)$ be a ρ-quasar-convex function with $\rho \in \left(\frac{1}{2}, 1\right]$, and let $\{M_i\}_{i=0}^k$ be some sequence of positive numbers from Algorithm 1. If $\alpha \leqslant \frac{2\rho-1}{\rho}$, then*

$$\|x_{k+1} - x_*\| \leqslant \|x_k - x_*\| + \Delta\eta_k + \frac{1}{M_{k+1}}. \tag{27}$$

Proof. Since $x_{k+1} = x_k - \eta_k \widetilde{\nabla} f(x_k)$, the following equalities are valid:

$$\begin{aligned}\|x_{k+1} - x_*\|^2 &= \|x_k - x_* - \eta_k \widetilde{\nabla} f(x_k)\|^2 \\ &= \|x_k - x_*\|^2 - 2\eta_k \langle \widetilde{\nabla} f(x_k), x_k - x_* \rangle + \eta_k^2 \|\widetilde{\nabla} f(x_k)\|^2.\end{aligned}$$

By the quasar-convexity of the function f, we obtain

$$\begin{aligned}&\|x_{k+1} - x_*\|^2 \\ &\leqslant \|x_k - x_*\|^2 - 2\eta_k(\rho(f(x_k) - f^*) - \Delta\|x_k - x_*\|) + \eta_k^2 \|\widetilde{\nabla} f(x_k)\|^2 \\ &\overset{(26)}{\leqslant} \|x_k - x_*\|^2 - 2\eta_k(\rho(f(x_k) - f^*) - \Delta\|x_k - x_*\|) \\ &+ \alpha\eta_k \left(\frac{2}{\alpha(2-\alpha)}(f(x_k) - f^*) + \frac{2\Delta}{M_{k+1}(2-\alpha)}\right) \\ &= \|x_k - x_*\|^2 + \eta_k \left(\frac{2}{2-\alpha} - 2\rho\right)(f(x_k) - f^*) \\ &+ 2\Delta\eta_k \left(\|x_k - x_*\| + \frac{\alpha}{M_{k+1}(2-\alpha)}\right).\end{aligned}$$

Let the parameter α be chosen so that $\frac{2}{2-\alpha} - 2\rho \leqslant 0$. As $\alpha \in (0, 1]$, this is possible if and only if $\rho \in \left(\frac{1}{2}, 1\right]$. In this case, we have

$$\begin{aligned}\|x_{k+1} - x_*\|^2 &\leqslant \|x_k - x_*\|^2 + 2\Delta\eta_k\|x_k - x_*\| + \Delta^2\eta_k^2 + \frac{2\alpha\Delta\eta_k}{M_{k+1}(2-\alpha)} - \Delta^2\eta_k^2 \\ &= (\|x_k - x_*\| + \Delta\eta_k)^2 + \Delta\eta_k \left(\frac{2\alpha}{M_{k+1}(2-\alpha)} - \Delta\eta_k\right).\end{aligned}$$

Consider the function $y(t) = -t^2 + \frac{2\alpha}{M_{k+1}(2-\alpha)}t$. At the point $t = \frac{\alpha}{M_{k+1}(2-\alpha)}$ it reaches the maximum value $\frac{\alpha^2}{M_{k+1}^2(2-\alpha)^2}$, and $\frac{\alpha^2}{M_{k+1}^2(2-\alpha)^2} \leqslant \frac{1}{M_{k+1}^2}$ for $\alpha \leqslant 1$. Thus,

$$\|x_{k+1} - x_*\|^2 \leqslant (\|x_k - x_*\| + \Delta\eta_k)^2 + \frac{1}{M_{k+1}^2}.$$

Note that for $a, b, c > 0$, the inequality $a^2 \leqslant b^2 + c^2$ implies $a^2 \leqslant (b + c)^2$. Hence, $a \leqslant b + c$, and, consequently,

$$\|x_{k+1} - x_*\| \leqslant \|x_k - x_*\| + \Delta\eta_k + \frac{1}{M_{k+1}}.$$

Note that inequalities (25) and (26) surely hold under $M_{k+1} \geqslant M$. If $\|\widetilde{\nabla} f(x_k)\| \geqslant 5\Delta$ and $M_i \geqslant \frac{M_0}{2}$ $\forall i = \overline{1, k+1}$ with some $M_0 > 0$, then

$$\eta_k = \frac{\alpha}{M_{k+1}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)} \leqslant \frac{\alpha}{6M_{k+1}\Delta} \leqslant \frac{\alpha}{3M_0\Delta}.$$

Theorem 2. *Let $f(x)$ be a ρ-quasar-convex and $(0,M)$-smooth function, where $\rho \in \left(\frac{1}{2},1\right]$, $\alpha \in (0,1]$, and $\alpha \leqslant \frac{2\rho-1}{\rho}$. Then after $\left\lceil \frac{8M(f(x_0)-f^*)}{3\alpha(2-\alpha)\Delta} \right\rceil$ iterations of Algorithm 1, there exists an iteration k for which either $\|\widetilde{\nabla} f(x_k)\| < 5\Delta$, or the exact solution will be reached.*

Proof. In view of $\|\widetilde{\nabla} f(x_k)\| \geqslant 5\Delta$, summing over k from 0 to $N-1$ in inequality (25) yields the relations

$$f(x_0) - f^* \geqslant f(x_0) - f(x_N) \geqslant \frac{3\alpha(2-\alpha)\Delta}{4} \sum_{k=0}^{N-1} \frac{1}{M_{k+1}},$$

which can be written as

$$\sum_{k=0}^{N-1} \frac{1}{M_{k+1}} \leqslant \frac{4(f(x_0)-f^*)}{3\alpha(2-\alpha)\Delta}.$$

Considering that $M_i \leqslant 2M$, we estimate the number of iterations for which $\|\widetilde{\nabla} f(x_k)\| \geqslant 5\Delta$. We have $\frac{N}{2M} \leqslant \frac{4(f(x_0)-f^*)}{3\alpha(2-\alpha)\Delta}$, or, in the final form,

$$N \leqslant \left\lceil \frac{8M(f(x_0)-f^*)}{3\alpha(2-\alpha)\Delta} \right\rceil.$$

Remark 6. Suppose that an acceptable solution quality is achieved under the condition $\|\widetilde{\nabla} f(x_k)\| < 5\Delta$. Indeed, the ρ-quasar-convexity of the function f and inequality (27) imply

$$f(x_k) - f^* \leqslant \frac{1}{\rho}\left(\|\widetilde{\nabla} f(x_k)\| + \Delta\right)\|x_k - x_*\| \leqslant \frac{6\Delta}{\rho}\left(\|x_0 - x_*\| + \frac{(\alpha+6)N}{3M_0}\right).$$

Hence, with a small number of iterations, one can guarantee a solution quality within $O(\Delta)$. In the case of large values of N, this cannot be stated for sure. Note that if the objective function satisfies the PL condition, an acceptable solution quality will be achieved under a small norm of the inexact gradient (see Remark 4).

In the special case $\Delta = 0$, we have the following result.

Theorem 3. *Let $f(x)$ be a ρ-quasar-convex and $(0,M)$-smooth function, where $\rho \in \left(\frac{1}{2},1\right]$, $\alpha \in (0,1]$, and $\alpha \leqslant \frac{2\rho-1}{\rho}$. Then Algorithm 1 with $\Delta = 0$ converges at the rate of a geometric progression.*

Proof. Consider the case $\Delta = 0$, i.e., $\widetilde{\nabla} f(x) = \nabla f(x)$ for each x. Then from line 7 of Algorithm 1 it follows that

$$f(x_k) - f(x_{k+1}) \geqslant \frac{\alpha(2-\alpha)}{2M_{k+1}} \frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_k)\|} = \frac{(2-\alpha)\eta_k}{2}\|\nabla f(x_k)\|^2. \tag{28}$$

In this case, we also have the relations

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - 2\eta_k\langle \nabla f(x_k), x_k - x_*\rangle + \eta_k^2\|\nabla f(x_k)\|^2 \\
&\leqslant \|x_k - x_*\|^2 - 2\eta_k\rho(f(x_k) - f^*) + \eta_k^2\|\nabla f(x_k)\|^2 \\
&\leqslant \|x_k - x_*\|^2 - 2\eta_k\rho(f(x_k) - f^*) + \frac{2\eta_k}{(2-\alpha)}(f(x_k) - f^*) \\
&= \|x_k - x_*\|^2 + \eta_k(f(x_k) - f^*)\left(\frac{2}{2-\alpha} - 2\rho\right) \\
&\leqslant \|x_k - x_*\|^2 \leqslant \|x_0 - x_*\|^2 = R^2,
\end{aligned}$$

where $\alpha \in (0, 1]$ and $\rho \in \left(\frac{1}{2}, 1\right]$. Further, by the quasar-convexity of the function f,

$$f(x_k) - f^* \leqslant \frac{1}{\rho}\|\nabla f(x_k)\|\|x_k - x_*\| \leqslant \frac{1}{\rho}\|\nabla f(x_k)\|\|x_0 - x_*\| = \frac{R}{\rho}\|\nabla f(x_k)\|.$$

Recall that, due to Lemmas 4 and 5, for $M_{k+1} \geqslant M$ the iteration exit criterion of Algorithm 1 (see line 7) is surely satisfied. Then for each $k \geqslant 0$ we have $M_{k+1} \leqslant 2M$, and from (28) it follows that

$$\begin{aligned} f(x_{k+1}) - f^* &\leqslant f(x_k) - f^* - \frac{\alpha(2-\alpha)}{2M_{k+1}}\|\nabla f(x_k)\| \\ &\leqslant f(x_k) - f^* - \frac{\alpha(2-\alpha)}{2M_{k+1}}\frac{\rho}{R}(f(x_k) - f^*) \\ &\leqslant \left(1 - \frac{\alpha(2-\alpha)\rho}{2M_{k+1}R}\right)(f(x_k) - f^*) \\ &\leqslant \left(1 - \frac{\alpha(2-\alpha)\rho}{4MR}\right)(f(x_k) - f^*) \\ &\leqslant \left(1 - \frac{\alpha(2-\alpha)\rho}{4MR}\right)^{k+1}(f(x_0) - f^*). \end{aligned}$$

The proof of Theorem 3 is complete.

5. COMPUTATIONAL EXPERIMENTS

This section presents the results of computational experiments conducted to assess the efficiency of the proposed algorithms when minimizing the logistic function (29) and the quasar-convex function (30).

5.1. Testing Results for Adaptive Algorithm 1

Let us compare Algorithm 1 with the adaptive universal gradient method [30] for the problem of minimizing the logistic function from Example 3:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m}\log(1 + \exp(-a_i^\top x)), \quad a_i \in \mathbb{R}^n. \tag{29}$$

As is well known, the logistic function (29) is convex and L-smooth with the constant $L = \frac{1}{m}\sum\limits_{i=1}^{m}\|a_i\|^2$, as well as (L_0, L_1)-smooth with the parameters $L_0 = 0$ and $L_1 = \max\limits_{1\leqslant i\leqslant m}\|a_i\|$.

The algorithms were run with the parameters $n = 1000$ and $m = 100$ and the initial point $x_0 = (1, \ldots, 1) \in \mathbb{R}^n$. The gradient inexactness parameter Δ took values from the set $\{0.5; 0.1; 0.05; 0.01\}$. The vectors $\{a_i\}_{i=1}^m$ were generated randomly from the standard Gaussian distribution. Figure 1 shows the iteration-to-iteration dynamics of the objective function values for the algorithms compared. According to the numerical results, Algorithm 1, which uses (L_0, L_1)-smoothness, demonstrates higher efficiency compared to the adaptive universal gradient method [30].

5.2. A Nonconvex Problem Example from the Class of Quasar-Convex Functions

This subsection presents a comparison of the proposed method (9) with the step size (10) and the gradient method with a constant step size of 0.001 on the example of a quasar-convex function from [27]:

$$f(x) = h(\|x\|_2)g\left(\frac{x}{\|x\|_2}\right), \tag{30}$$

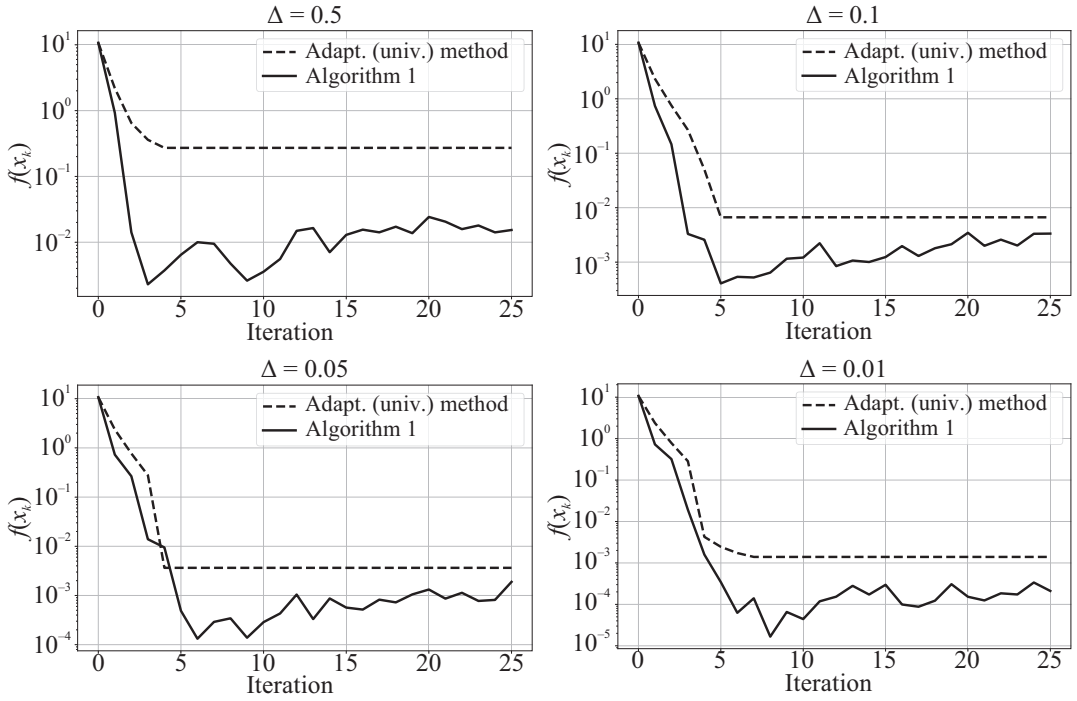

Fig. 1. Numerical solution of the logistic regression problem with (29), $n = 10^3$, and $m = 100$: Algorithm 1 vs. the adaptive universal gradient method [30].

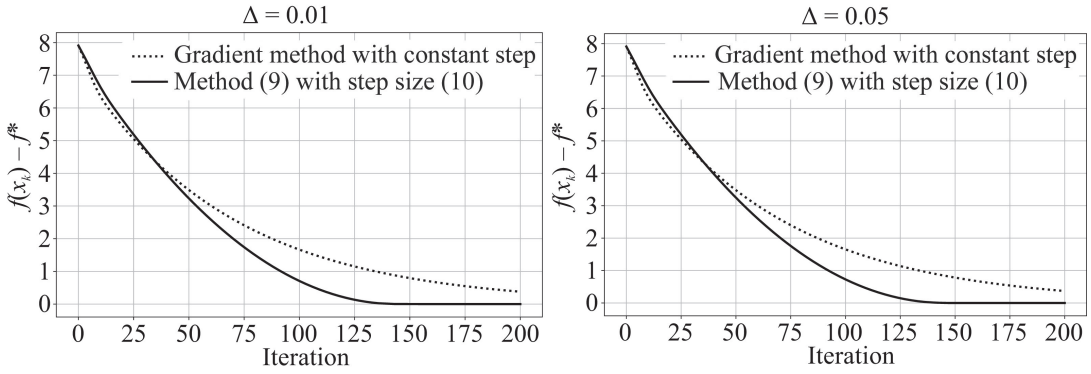

Fig. 2. The results of minimizing the objective function (30): algorithm (9) with the step size (10) vs. the gradient method with constant step size.

where $h(t) = t^2$ and

$$g(x_1, x_2) = 1 + \frac{1}{4N} \sum_{i=1}^{N} a_i \sin^2(b_i x_1) + c_i \cos^2(d_i x_2).$$

In the computational experiments, the elements of the sequences $\{a_i\}_i$ and $\{c_i\}_i$ were assigned independent uniform distributions on the interval $[0, 20)$. Similarly, the elements of the sequences $\{b_i\}_i$ and $\{d_i\}_i$ were independently and uniformly distributed on the interval $[-25, 25)$. All methods were initialized from the starting point $(1, 1)$ with the parameters $N = 10$ and $\alpha = 0.01$. Figure 2 shows the iteration-to-iteration dynamics of the objective function values for the algorithms compared. According to the experimental results, the proposed algorithm (9) with the step size (10) chosen based on the (L_0, L_1)-smoothness of the objective function has higher efficiency than the gradient method with a constant step size.

6. CONCLUSIONS

In this paper, we have proposed and analyzed several gradient methods for minimizing (L_0, L_1)-smooth objective functions under the additive inexactness in the gradient values at the iterations. Attention has been focused on investigating the convergence of these methods for some classes of nonconvex functions, such as ρ-quasar-convex ones and those satisfying the Polyak–Łojasiewicz (PL) condition.

The state-of-the-art research in this field has been reviewed. All key definitions and auxiliary results have been adapted to the case of a Δ-inexact gradient. Some lemmas characterizing the impact of the additive gradient error on the trajectory of the gradient method have been formulated and proven. The convergence rate of the gradient method on the class of quasar-convex functions satisfying the PL condition has been estimated theoretically. In particular, the method has been shown to converge at a nearly linear rate (up to some parameters related to the errors). It has also been proven that starting from some iteration, the method converges to a neighborhood of the exact solution on the class of functions with the PL condition. Separate consideration has been given to a detailed study of the case $L_0 = 0$ (the class of $(0, M)$-smooth functions), which is of significant interest for machine learning problems, such as training logistic regression models. An adaptive modification of the gradient method using inexact gradient values has been developed, and a theoretical estimate of its convergence rate has been derived. In the case of exact gradient values (for $\Delta = 0$), the method has been shown to converge at the rate of a geometric progression. The efficiency of the proposed methods has been confirmed by the results of computational experiments conducted both on the logistic regression problem and on a certain nonconvex quasar-convex problem.

Concerning directions for further research, it seems interesting to study accelerated methods for (L_0, L_1)-smooth problems with inexact gradient information and develop methods with adaptive tuning to the parameters L_0 and L_1.

FUNDING

This work was supported by the Ministry of Science and Higher Education of the Russian Federation within state assignment no. 075-03-2024-074 under the project “Study of Asymptotic Characteristics of Fluctuations of Differential Equations and Systems, and Optimization Methods.”

REFERENCES

1. Zhang, J., He, T., Sra, S., and Jadbabaie, A., Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity, *Proceedings of the International Conference on Learning Representations*, 2020, pp. 1–14.
2. Schmidt, M., Le Roux, N., and Bach, F., Minimizing Finite Sums with the Stochastic Average Gradient, *Mathematical Programming*, 2017, vol. 162, pp. 83–112.
3. Johnson, R. and Zhang, T., Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction, *Advances in Neural Information Processing Systems*, 2013, no. 26, pp. 315–323.
4. Defazio, A., Bach, F., and Lacoste-Julien, S., SAGA: A Fast Incremental Gradient Method with Support for Non-Strongly Convex Composite Objectives, *Advances in Neural Information Processing Systems*, 2014, no. 2, pp. 1646–1654.
5. Nguyen, L., Liu, J., Scheinberg, K., and Takač, M., Sarah: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient, *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2014, vol. 6, pp. 4009–4023.
6. Nguyen, L., Scheinberg, K., and Takač, M., Inexact Sarah Algorithm for Stochastic Optimization, *Optimization Methods and Software*, 2021, vol. 36, no. 1, pp. 237–258.

7. Beznosikov, A. and Takač, M., Random-Reshuffled SARAH Does Not Need Full Gradient Computations, *Optim. Lett.*, 2024, vol. 18, pp. 727–749.

8. Shi, Z., Sadiev, A., Loizou, N., et al., Ai-Sarah: Adaptive and Implicit Stochastic Recursive Gradient Methods, *Transactions on Machine Learning Research*, 2023, pp. 1–40.

9. Defazio, A. and Bottou, L., On the Ineffectiveness of Variance Reduced Optimization for Deep Learning, *Advances in Neural Information Processing Systems*, 2019, vol. 32, pp. 1753–1763.

10. Zhang, B., Jin, J., Fang, C., and Wang, L., Improved Analysis of Clipping Algorithms for Non-Convex Optimization, *Advances in Neural Information Processing Systems*, 2020, vol. 19, pp. 15511–15522.

11. Chen, Z., Zhou, Y., Liang, Y., and Lu, Z., Generalized-Smooth Nonconvex Optimization Is as Efficient as Smooth Nonconvex Optimization, *Proceedings of the International Conference on Machine Learning (PMLR)*, 2023, pp. 5396–5427.

12. Zhao, S.-Y., Xie, Y.-P., and Li, W.-J., On the Convergence and Improvement of Stochastic Normalized Gradient Descent, *Science China Information Sciences*, 2021, vol. 64, pp. 1–13.

13. Faw, M., Rout, L., Caramanis, C., and Shakkottai, S., Beyond Uniform Smoothness: A Stopped Analysis of Adaptive SGD, *Proceedings of the 36th Annual Conference on Learning Theory (PMLR)*, 2023, pp. 89–160.

14. Wang, B., Zhang, H., Ma, Z., and Chen, W., Convergence of Adagrad for Non-Convex Objectives: Simple Proofs and Relaxed Assumptions, *Proceedings of the 36th Annual Conference on Learning Theory (PMLR)*, 2023, pp. 161–190.

15. Li, H., Rakhlin, A., and Jadbabaie, A., Convergence of Adam under Relaxed Assumptions, *Advances in Neural Information Processing Systems*, 2024, pp. 1792–1804.

16. Hubler, F., Yang, J., Li, X., and He, N., Parameter-Agnostic Optimization under Relaxed Smoothness, *Proceedings of the International Conference on Artificial Intelligence and Statistics (PMLR)*, 2024, pp. 4861–4869.

17. Pascanu, R., Mikolov, T., and Bengio, Y., On the Difficulty of Training Recurrent Neural Networks, *Proceedings of the International Conference on Machine Learning*, 2013, vol. 28, pp. 1310–1318.

18. Polyak, B., *Introduction to Optimization*, New York: Optimization Software, 1987.

19. Koloskova, A., Hendrikx, H., and Stich, S., Revisiting Gradient Clipping: Stochastic Bias and Tight Convergence Guarantees, *Proceedings of the International Conference on Machine Learning*, 2023, pp. 17343–17363.

20. Takezawa, Y., Bao, H., Sato, R., et al., Polyak Meets Parameter-Free Clipped Gradient Descent, *arXiv:2405.15010*, 2024.

21. Li, H., Qian, J., Tian, Y., Rakhlin, A., and Jadbabaie, A., Convex and Non-Convex Optimization under Generalized Smoothness, *Advances in Neural Information Processing Systems*, 2024, vol. 36, pp. 2675–2686.

22. Gorbunov, E., Tupitsa, N., Choudhury, S., et al., Methods for Convex $(L0, L1)$-Smooth Optimization: Clipping, Acceleration, and Adaptivity, *arXiv:2409.14989*, 2024.

23. Vankov, D., Rodomanov, A., Nedich, A., et al., Optimizing $(L0, L1)$-Smooth Functions by Gradient Methods, *arXiv:2410.10800*, 2024.

24. Lobanov, A., Gasnikov, A., Gorbunov, E., and Takáč, M., Linear Convergence Rate in Convex Setup is Possible! Gradient Descent Method Variants under (L0, L1)-Smoothness, *arXiv:2412.17050*, 2024.

25. Stonyakin, F., Kuruzov, I., and Polyak, B., Stopping Rules for Gradient Methods for Non-Convex Problems with Additive Noise in Gradient, *Journal of Optimization Theory and Applications*, 2022, vol. 198, pp. 531–551.

26. Wang, J. and Wibisono, A., Continuized Acceleration for Quasar Convex Functions in Non-Convex Optimization, *arXiv.2302.07851*, 2023.

27. Hermant, J., Aujol, J.F., Dossal, C., and Rondepierre, A., Study of the Behaviour of Nesterov Accelerated Gradient in a Non Convex Setting: The Strongly Quasar Convex Case, *arXiv:2405.19809*, 2024.

28. Hinder, O., Sidford, A., and Sohoni, N., Near-Optimal Methods for Minimizing Star-Convex Functions and Beyond, *arXiv.1906.11985*, 2019.

29. Karimi, H., Nutini, J., and Schmidt, M., Linear Convergence of Gradient and Proximal-Gradient Methods under the Polyak–Lojasiewicz Condition, *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016, pp. 795–811.

30. Stonyakin, F., Tyurin, A., Gasnikov, A., et al., Inexact Relative Smoothness and Strong Convexity for Optimization and Variational Inequalities by Inexact Model, *arXiv:2402.06319*, 2024.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board