

# Accelerated gradient methods with absolute and relative noise in the gradient

Artem Vasin <sup>a</sup> and Alexander Gasnikov <sup>a,b,c</sup> and Pavel Dvurechensky<sup>d</sup> and Vladimir Spokoiny<sup>d,e</sup>

<sup>a</sup>Moscow Institute of Physics and Technology, Dolgoprudny, Russia; <sup>b</sup>Institute for Information Transmission Problems, Moscow, Russia; <sup>c</sup>ISP RAS Research Center for Trusted Artificial Intelligence, Russia; <sup>d</sup>Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany; <sup>e</sup>Humboldt-University of Berlin, Berlin, Germany.

## ARTICLE HISTORY

Compiled January 10, 2023

## ABSTRACT

In this paper, we investigate accelerated first-order methods for smooth convex optimization problems under inexact information on the gradient of the objective. The noise in the gradient is considered to be additive with two possibilities: absolute noise bounded by a constant, and relative noise proportional to the norm of the gradient. We investigate the accumulation of the errors in the convex and strongly convex settings with the main difference with most of the previous works being that the feasible set can be unbounded. The key to the latter is to prove a bound on the trajectory of the algorithm. We also give a stopping criterion for the algorithm and consider extensions to the cases of stochastic optimization and composite nonsmooth problems.

## 1. Introduction

We consider convex optimization problem on a closed convex (not necessarily bounded) set  $Q \subseteq \mathbb{R}^n$ :

$$\min_{x \in Q} f(x). \quad (1)$$

We assume that the objective  $f$  is  $L_f$ -smooth and strongly convex with the parameter  $\mu \geq 0$ , i.e., for all  $x, y \in Q$ :

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L_f \|y - x\|_2,$$

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 \leq f(y).$$

---

This work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

In the convergence rate analysis of different first-order methods these assumptions are typically used in the form of an upper and lower quadratic bounds [4, 6, 9, 14, 20, 23, 24, 27, 28, 37, 40, 51, 53] for the objective:

$$\begin{aligned} f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 &\leq f(y) \\ &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f}{2} \|y - x\|_2^2. \end{aligned} \quad (2)$$

Note that the last relation is a consequence of the  $L_f$ -smoothness and, in general, is not equivalent, to it [28, 52].

In many applications, instead of an access to the exact gradient  $\nabla f(x)$  an algorithm has access only to its inexact approximation  $\tilde{\nabla} f(x)$ . Typical examples include gradient-free (or zeroth-order) methods which use a gradient estimator based on finite differences [7, 11, 47], and optimization problems in infinite-dimensional spaces related to inverse problems [29, 34]. The two most popular definitions of gradient inexactness in practice are [45] as follows: for all  $x \in Q$  it holds that

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \delta, \quad (\text{absolute error}) \quad \text{or} \quad (3)$$

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \alpha \|\nabla f(x)\|_2, \quad \alpha \in [0, 1] \quad (\text{relative error}). \quad (4)$$

Under assumption (3), many results exist for non-accelerated and accelerated first-order methods, see, e.g., [1, 10, 12, 45]. These results are in a sense pessimistic in general with the explanation going back to the analysis in [44]. We can explain this by a very simple example. Consider the following problem

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} \sum_{i=1}^n \lambda_i \cdot (x^i)^2 \right\}, \quad (5)$$

where  $0 \leq \mu = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n = L_f$ ,  $L_f \geq 2\mu$ . Clearly, the solution of this problem is  $x^* = 0$ . Assume that the inexactness takes place only in the first component  $x_1$ , i.e., instead of  $\partial f(x)/\partial x^1 = \mu x^1$  we have access to  $\tilde{\partial} f(x)/\partial x^1 = \mu x^1 - \delta$ , where  $\delta$  is the error. For the simple gradient descent

$$x_k = x_{k-1} - \frac{1}{L_f} \tilde{\nabla} f(x_{k-1}),$$

we can conclude that if  $x_0^1 \geq 0$ , then for all  $k \in \mathbb{N}$  large enough, i.e.,  $k \gg L/\mu$ , it holds that

$$x_k^1 \geq \frac{\delta}{L} \frac{1 - (1 - \mu/L_f)^k}{1 - (1 - \mu/L_f)} \simeq \frac{\delta}{\mu}. \quad (6)$$

Hence,<sup>1</sup>

$$f(x_k) - f(x^*) \gtrsim \frac{\delta^2}{2\mu}.$$

From this result, we see that it may be problematic to approximate  $f(x^*)$  with any desired accuracy, especially in the ill-conditioned setting when the strong convexity constant  $\mu$  is smaller than the desired accuracy  $\varepsilon$ . For accelerated gradient methods the situation may be even worse since they are more sensitive to the gradient errors and such errors may even be accumulated by the algorithm [15, 21, 28]. This drawback may be overcome by proposing a certain stopping rule so that the algorithm does not try to minimize below some threshold given by the gradient error or by adding a strongly convex regularizer with coefficient  $\mu$  of the same order as the desired accuracy  $\varepsilon$ , see [28, 38, 44, 45]. Roughly speaking, for non-accelerated algorithms it was proved in [44, 45] that if  $\delta$  is of the order  $\varepsilon^2$ , then it is possible to reach  $\varepsilon$ -accuracy in the objective residual function in almost the same number of iterations as in the exact case  $\delta = 0$  by applying a computationally convenient stopping rule.

In this paper, we analyze an accelerated gradient method in both convex and strongly convex settings and estimate how the gradient error defined in (3) influences the convergence rate. An important part of our contribution is that our analysis is made without an assumption that the feasible set  $Q$  is bounded. The main key for this development is a recurrent estimate for the distance between the current iterates and the optimal solution closest to the starting point. In particular, our results imply that it is sufficient to assume that  $\delta$  is of the order  $\varepsilon$  in order to obtain objective residual of the order  $\varepsilon$ . We also present a stopping rule and prove that if it is satisfied at some iteration, the algorithm solves problem (1) with certain accuracy. Moreover, we prove that until this rule is fulfilled, the trajectory of the algorithm is bounded (which helps us to treat the setting of possibly unbounded set  $Q$ ) and that it is fulfilled for sure in a number of iterations which is optimal for the class of smooth convex optimization problems.

Under assumption (4), non-accelerated gradient method for strongly convex problems is shown in [45] to have linear convergence with condition number  $O\left(\frac{1}{1-\alpha} \cdot \frac{L_f}{\mu}\right)$ , i.e.  $\frac{1}{1-\alpha}$  times worse than in the exact case. Yet, convergence to any small error is guaranteed unlike the case of inexactness (3). This result holds also under the relaxed strong convexity assumption [28] known as Polyak–Lojasiewicz or gradient domination condition. We are not aware of any such results for accelerated gradient methods.

In this paper, we analyze an accelerated gradient method under inexact gradients satisfying (4) and answer the question of what is the maximum value of  $\alpha$  such that the accelerated algorithm with inexact gradients converges with the same rate as the exact accelerated algorithm. For the case  $\mu \neq 0$  our answer is that  $\alpha$  should satisfy  $\alpha = O\left(\frac{\mu}{L_f}\right)$ . We hypothesize that this bound can be improved to  $\alpha = O\left(\left(\frac{\mu}{L_f}\right)^{1/2}\right)$  and, for the case  $\mu = 0$ , the iteration-dependent value  $\alpha_k$  should satisfy  $\alpha_k = O\left(\left(\frac{1}{k}\right)^{3/2}\right)$ , where  $k$  is the iteration counter. Numerical experiments demonstrate that, in general, for  $\alpha$  larger than the mentioned above thresholds the convergence may slow down a lot up to divergence for the considered accelerated method.

---

<sup>1</sup>This bound corresponds to the worst-case philosophy, i.e., choosing the worst example for the considered class of methods [9, 28, 39, 40]. We expect more interesting results by considering average-case complexity [43, 49].

Close results with the bound  $\alpha = O\left(\left(\frac{\mu}{L_f}\right)^{5/4}\right)$  in the case  $\mu \gg \varepsilon$  were recently obtained using another techniques in stochastic optimization with decision dependent distribution [17] and policy evaluation in reinforcement learning via reduction to stochastic variational inequality with Markovian noise [36]. In [17, 36], the authors assumed that

$$\|\tilde{\nabla}f(x) - \nabla f(x)\|_2 \leq B\|x - x^*\|_2. \quad (7)$$

Since  $x^*$  is a solution, when  $Q = \mathbb{R}^n$ , we have  $\nabla f(x^*) = 0$ . Therefore,

$$\|\nabla f(x)\|_2 = \|\nabla f(x) - \nabla f(x^*)\|_2 \leq L_f\|x - x^*\|_2.$$

Thus, if (4) holds, then (7) also holds with  $B = \alpha L_f$ .

## 2. Ideas behind the results

### 2.1. Absolute noise

Important results on gradient error accumulation for first-order methods were developed in a series of works by O. Devolder, F. Glineur and Yu. Nesterov 2011–2014 [13–16]. In these works, the authors were motivated by inequalities (2). Their idea was to relax (2), assuming inexactness in the gradient, introducing the inexact gradient  $\tilde{\nabla}f(x)$ , satisfying for all  $x, y \in Q$

$$\begin{aligned} f(x) + \langle \tilde{\nabla}f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 - \delta &\leq f(y) \\ &\leq f(x) + \langle \tilde{\nabla}f(x), y - x \rangle + \frac{L_f}{2}\|y - x\|_2^2 + \delta. \end{aligned} \quad (8)$$

This assumption allows to develop a theory for error accumulation for first-order methods. In particular, they obtained the following convergence rates for non-accelerated gradient methods:

$$f(x_k) - f(x^*) = O\left(\min\left\{\frac{L_f R^2}{k} + \delta, L_f R^2 \exp\left(-\frac{\mu}{L_f}k\right) + \delta\right\}\right), \quad (9)$$

and for accelerated methods:

$$f(x_k) - f(x^*) = O\left(\min\left\{\frac{L_f R^2}{k^2} + k\delta, L_f R^2 \exp\left(-\sqrt{\frac{\mu}{L_f}}\frac{k}{2}\right) + \sqrt{\frac{L_f}{\mu}}\delta\right\}\right), \quad (10)$$

where  $R$  is such that  $\|x_{start} - x^*\|_2 \leq R$ , i.e., an estimate for the distance between the starting point  $x_{start}$  and a solution  $x^*$ . If  $x^*$  is not unique, one may take  $x^*$  to be the closest point to  $x_{start}$ . Both of these bounds are unimprovable [15, 16]. See also [14, 21, 35] for “intermediate” situations between accelerated and non-accelerated methods and extensions for stochastic optimization.

Following [16], it is possible to make a reduction of the “absolute noise” inexactness

in the sense of (3) to the inexactness in the sense of (8) by setting

$$\delta = \delta_{(8)} = \frac{\delta_{(3)}^2}{2L_f} + \frac{\delta_{(3)}^2}{\mu} \simeq \frac{\delta_{(3)}^2}{\mu} \quad (11)$$

and setting  $L_{f,(8)} = 2L_{f,(2)}$ ,  $\mu_{f,(8)} = \mu_{f,(2)}/2$ . The key observations here are that

$$\langle \tilde{\nabla} f(x) - \nabla f(x), y - x \rangle \leq \frac{1}{2L_f} \|\tilde{\nabla} f(x) - \nabla f(x)\|_2^2 + \frac{L_f}{2} \|y - x\|_2^2,$$

$$\langle \tilde{\nabla} f(x) - \nabla f(x), y - x \rangle \geq \frac{1}{\mu} \|\tilde{\nabla} f(x) - \nabla f(x)\|_2^2 - \frac{\mu}{4} \|y - x\|_2^2.$$

From this reduction, we see that when  $\mu > 0$ , for non-accelerated methods. the result (9) is almost the same as in the example in (5). We see also, that, if the error can be controlled, to guarantee that  $f(x_k) - f(x^*) \leq \varepsilon$  for non-accelerated method when <sup>2</sup>  $\mu = \Omega(\varepsilon)$  we should set  $\delta_{(3)} = O(\varepsilon)$ , which is an expected result. Unfortunately, for accelerated methods, such reduction leads to the bound  $\delta_{(3)} = O(\varepsilon^{3/2})$ , which is worse than our bound indicated in Section 1. The key to our improvement is a more refined version of (8).

In the works [15, 18, 19, 50, 51], the following refined version of (8) is used:

$$\begin{aligned} f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 - \delta_1 \|y - x\|_2 &\leq f(y) \\ &\leq f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{L_f}{2} \|y - x\|_2^2 + \delta_2. \end{aligned} \quad (12)$$

These inequalities lead to the following counterparts of (9) and (10), respectively, for non-accelerated gradient methods:

$$\begin{aligned} f(x_k) - f(x^*) \\ = O \left( \min \left\{ \frac{L_f R^2}{k} + \tilde{R} \delta_1 + \delta_2, L_f R^2 \exp \left( -\frac{\mu}{L_f} k \right) + \tilde{R} \delta_1 + \delta_2 \right\} \right), \end{aligned} \quad (13)$$

and for accelerated gradient methods [15, 19]:

$$\begin{aligned} f(x_k) - f(x^*) \\ = O \left( \min \left\{ \frac{L_f R^2}{k^2} + \tilde{R} \delta_1 + k \delta_2, L_f R^2 \exp \left( -\sqrt{\frac{\mu}{L_f}} \frac{k}{2} \right) + \tilde{R} \delta_1 + \sqrt{\frac{L_f}{\mu}} \delta_2 \right\} \right), \end{aligned} \quad (14)$$

where  $\tilde{R}$  is the maximum distance between the sequences of iterates generated by the algorithm and the solution  $x^*$  closest to the starting point.

---

<sup>2</sup>If  $\mu \lesssim \varepsilon$ , we can regularize the problem and guarantee that  $\mu = \Omega(\varepsilon)$ , see [28]. Another advantage of strong convexity is the possibility to use the norm of inexact gradient for the stopping criteria, see [28, 44]. Yet, regularization requires [28] some prior knowledge about the distance to the solution. Since we typically do not have such information the procedure becomes more difficult via applying the restart technique, see [27, 28].

From (13), (14), we see that if  $\tilde{R}$  is bounded,<sup>3</sup> then by setting

$$\delta_1 = \delta_{(3)}, \delta_2 = \frac{\delta_{(3)}^2}{2L_f},$$

we obtain the desired result: it is possible to guarantee  $f(x_k) - f(x^*) \leq \varepsilon$  with  $\delta_{(3)} = O(\varepsilon)$ .

Previous works mainly rely on the assumption that  $\tilde{R}$  is bounded. As we may see from example (5), in general, when the strong convexity parameter  $\mu$  is small compared to the desired accuracy  $\varepsilon$ , only a bound

$$\tilde{R} \simeq R + \frac{\delta_{(3)}}{\mu} \gtrsim R + \frac{\delta_{(3)}}{\varepsilon}$$

is possible to obtain [28]. This bound leads to very pessimistic estimates. Moreover, the growth of  $\tilde{R}$  is observed in different numerical experiments and in theoretical estimates caused by error accumulation. In our work, we investigate this problem and, in particular, propose an alternative to regularization<sup>4</sup> approach that is based on “early stopping”<sup>5</sup> of the considered iterative procedure by developing proper stopping rule.

## 2.2. Relative noise

We now explain a way of reduction of the relative inexactness in the sense of (4) to the inexactness in the sense of (8), which allows us to apply (10) when  $\mu \gg \varepsilon$ . Since  $f(x)$  has Lipschitz gradient, from (4), (8), we can derive that after  $k$  iterations (where  $k$  is greater than  $\sqrt{L_f/\mu}$  by a logarithmic factor  $\log(L_f R^2/\varepsilon)$  with  $\varepsilon$  being the desired accuracy in terms of the objective residual):

$$\begin{aligned} f(x_k) - f(x^*) &\stackrel{(10),(11)}{\simeq} \frac{\varepsilon}{2} + \sqrt{\frac{L_f}{\mu} \frac{\delta_{(3)}^2}{\mu}} \simeq \sqrt{\frac{L_f}{\mu} \frac{\delta_{(3)}^2}{\mu}} \\ &\stackrel{(4),(8)}{\simeq} \sqrt{\frac{L_f}{\mu} \frac{\alpha^2 \max_{t=1,\dots,k} \|\nabla f(x_t)\|_2^2}{\mu}} \leq \sqrt{\frac{L_f}{\mu} \frac{2L_f \alpha^2 \max_{t=1,\dots,k} (f(x_k) - f(x^*))}{\mu}} \\ &\lesssim \sqrt{\frac{L_f}{\mu} \frac{4L_f \alpha^2 (f(x_0) - f(x^*))}{\mu}}. \end{aligned} \quad (15)$$

Choosing  $\alpha = O\left(\left(\frac{\mu}{L_f}\right)^{3/4}\right)$ , we guarantee that the following restart condition holds

$$f(x_k) - f(x^*) \leq \frac{1}{2} (f(x_0) - f(x^*)).$$

When the restart condition holds, we restart the method. Then, after  $\log(\Delta f/\varepsilon)$  restarts we can guarantee the desired  $\varepsilon$ -accuracy in terms of the objective residual. In

<sup>3</sup>In many situations this is true. For example, when  $Q$  is bounded or when  $\mu \gg \varepsilon$ .

<sup>4</sup>By using regularization we can guarantee  $\mu \sim \varepsilon$  and therefore with  $\delta_{(3)} \sim \varepsilon$  we have the desired estimate  $\tilde{R} \simeq R$ .

<sup>5</sup>This terminology is popular also in Machine Learning community, where “early stopping” is used also as an alternative to regularization to prevent overfitting [31].

ill-conditioned setting, i.e., when  $\mu$  is small, the calculations are more involved. Yet, the main idea remains the same and replacing  $\sqrt{L_f/\mu}$  with  $k$  (cf. (10)) we obtain that the inequality  $\alpha_k \lesssim (\frac{1}{k})^{3/2}$  allows us to obtain the same convergence rate as in the exact gradients case.

Among many types of accelerated gradient methods, we choose to consider methods with one projection (Similar Triangles Methods (STM)), see [10, 23, 30, 32, 50] and references therein. We choose this type of accelerated methods since: 1) it is primal-dual [22, 30]; 2) it is possible to bound  $\tilde{R}$  in the absence of noise [30, 40, 50] and when the noise is present [32, 33]; 3) has previously been intensively investigated, see [23] and references therein.

### 3. Some motivation for inexact gradients

In this section, we describe two, among many others, research directions where inexact gradients play an important role. We emphasise that, although the results below are not new, the way they are presented is of some value in our opinion and can be useful for the specialists in these directions.

#### 3.1. Gradient-free methods

In this subsection, we consider convex optimization problem:

$$\min_{x \in Q \subseteq \mathbb{R}^n} f(x),$$

where  $Q$  is a convex and closed set. In some applications we do not have an access to the gradient  $\nabla f(x)$  of the objective function, but we can calculate the value <sup>6</sup> of  $f(x)$  with accuracy  $\delta_f$  [11], i.e., one can evaluate  $\tilde{f}(x)$  s.t.

$$|\tilde{f}(x) - f(x)| \leq \delta_f.$$

An interesting question in this setting is as follows. If the accuracy  $\delta_f$  of the approximation can be controlled, how should it be chosen in order to guarantee a desired accuracy  $\varepsilon$  when solving problem (1)? A related question is what is the largest level of noise  $\delta_f$  such that the algorithm can still achieve a desired accuracy  $\varepsilon$ ?

In the considered setting, a number of options exists for approximating the gradient, see, e.g., [7] and references therein. We consider the following examples, assuming that  $f$  has  $L_p$ -Lipschitz  $p$ -th order derivatives w.r.t. the Euclidean norm.

- **( $p$ -th order finite-differences).** In this case, the gradient approximation is constructed via finite differences of inexact values  $\tilde{f}(x)$ , which, e.g., in the case of  $p = 2$  lead to the following approximation to the  $i$ -th partial derivative

$$\tilde{\nabla}_i f(x) = \frac{\tilde{f}(x + he_i) - \tilde{f}(x - he_i)}{2h}, \quad i \in \{1, \dots, n\},$$

---

<sup>6</sup>The approach we describe requires that the function values are available not only in  $Q$ , but also in some (depends on a particular approach) vicinity of  $Q$ . This problem can be solved in two different ways. The first one is “slightly shrink the feasible set” approach [8]. The second one is “continuation” of  $f$  to  $\mathbb{R}^n$  preserving its convexity and Lipschitz continuity [47]:  $f_{new}(x) := f(\text{proj}_Q(x)) + \alpha \min_{y \in Q} \|x - y\|_2$ .

where  $e_i$  is the  $i$ -th coordinate vector and  $h > 0$  is a parameter. For general values of  $p$ , we have that (3) holds with

$$\delta = \sqrt{n}O\left(L_p h^p + \frac{\delta_f}{h}\right),$$

see [7]. The optimal choice of  $h$  guarantees that  $\delta = O\left(\sqrt{n}\delta_f^{\frac{p}{p+1}}\right)$ . From Section 1, we know that it is possible to solve problem (1) with accuracy  $\varepsilon = O(\delta)$  in terms of the objective value. Hence, in order to guarantee  $\varepsilon$ -accuracy, we should choose

$$\delta_f \sim \left(\frac{\varepsilon}{\sqrt{n}}\right)^{\frac{p+1}{p}}.$$

Unfortunately, such a simple idea does not allow one to reach the following lower bound in the class of algorithms that have sample complexity  $O\left(\frac{n^{c_1}}{\varepsilon^{c_2}}\right)$ , for some  $c_1, c_2 \geq 1$ : [47]

$$\delta_f \sim \max\left\{\frac{\varepsilon^2}{\sqrt{n}}, \frac{\varepsilon}{n}\right\}. \quad (16)$$

Note that, instead of the finite-difference approximation approach, in some applications we can use the kernel approach [3, 46] which has recently a got renowned interest [2, 42].

- **(Gaussian Smoothed Gradients).** In this case, the approximate gradient is formally defined as

$$\tilde{\nabla}f(x) = \frac{1}{h}\mathbb{E}\tilde{f}(x + he)e,$$

where  $e \in N(0, I_n)$  is the standard Gaussian random vector. This implies that (3) holds with

$$\delta = O\left(n^{p/2}L_p h^p + \frac{\sqrt{n}\delta_f}{h}\right),$$

see [7, 41]. The optimal choice of  $h$  guarantees that  $\delta = O\left((n\delta_f)^{\frac{p}{p+1}}\right)$ . Hence, in order to guarantee  $\varepsilon$ -accuracy, we should choose

$$\delta_f = O\left(\frac{\varepsilon^{\frac{p+1}{p}}}{n}\right).$$

This bound does not match the lower bound (16) as well. Moreover, here (and in the next approach) we have an additional difficulty since  $\tilde{\nabla}f(x)$ , in general, is not possible to evaluate exactly and only an inexact approximation is possible, for example, by the Monte Carlo approach [7], which leads to additional computational price for the better quality of approximation.

- **(Sphere Smoothed Gradients).** In this case, the approximate gradient is formally defined as

$$\tilde{\nabla}f(x) = \frac{n}{h}\mathbb{E}\tilde{f}(x + he)e,$$

where  $e$  is random vector with uniform distribution in the unit sphere in  $\mathbb{R}^n$  with the center at 0. This implies that (3) holds with

$$\delta = O\left(L_p h^p + \frac{n\delta_f}{h}\right),$$

see [7]. The optimal choice of  $h$  guarantees  $\delta = O\left((n\delta_f)^{\frac{p}{p+1}}\right)$ . Hence, in order to guarantee  $\varepsilon$ -accuracy, we should choose

$$\delta_f = O\left(\frac{\varepsilon^{\frac{p+1}{p}}}{n}\right).$$

This bound does not match the lower bound (16) as well. It may seem that the this and the previous approaches are almost the same, but below we give a more accurate result for the Sphere smoothing. We are not aware of a way to obtain such a result for the Gaussian smoothing. The result is as follows [15, 47]: for the Sphere smoothed gradient, we have that (8) holds with

$$\delta \simeq 2L_0 h + \frac{\sqrt{n}\delta_f \tilde{R}}{h}, \tag{17}$$

where  $L_0$  is the Lipschitz constant of  $f$  and in (8)  $L_f = \min\left\{L_1, \frac{7L_0^2}{h}\right\}$  when  $p = 1$  and  $L_f = \frac{7L_0^2}{h}$ , when  $p = 0$ . The bound (17) is more accurate than the previous bounds since it corresponds to the first part of the lower bound (16). Indeed, by choosing a proper  $h$  in (17) we obtain  $\varepsilon \sim \delta \sim n^{1/4}\delta_f^{1/2}$ . Hence, in order to guarantee  $\varepsilon$ -accuracy, we should choose

$$\delta_f = O\left(\frac{\varepsilon^2}{\sqrt{n}}\right).$$

The other part of the lower bound (16), i.e., the case when  $\delta_f = O\left(\frac{\varepsilon}{n}\right)$ , is also tight, see [5]. Here we can also repeat the remark that the sphere smoothed gradient approximation  $\tilde{\nabla}f(x)$ , in general, is not available and needs to be approximated by a stochastic inexact gradient. In Section 6, we describe an extension of our analysis of accelerated gradient method with absolute noise in the gradient to the setting of stochastic gradients.

The bound in (17) and its consequences additionally illustrate that the inexactness and algorithms we describe in Section 2 and develop below are also tight (optimal) enough. Otherwise, it would not be possible to achieve the lower bound using the reduction of gradient-free methods to gradient methods with inexact oracle and the proposed analysis of the error accumulation for gradient-type methods.

### 3.2. Inverse problems

Another rather important research direction where gradients are typically available only approximately is optimization in Hilbert spaces [54], arising, in particular, in inverse problems theory [34].

We start by recalling a way to calculate a derivative in a general Hilbert space. Let  $J(q) := J(q, u(q))$ , where  $u(q)$  is the unique solution of the equation  $G(q, u) = 0$ . Assume that the partial  $q$ -derivative  $G_q(q, u)$  of the operator  $G(q, u)$  is invertible. Then, we have

$$G_q(q, u) + G_u(q, u)\nabla u(q) = 0, \quad \text{and} \quad \nabla u(q) = -[G_u(q, u)]^{-1} G_q(q, u).$$

Therefore,

$$\nabla J(q) = J_q(q, u) + J_u(q, u)\nabla u(q) = J_q(q, u) - J_u(q, u) [G_u(q, u)]^{-1} G_q(q, u).$$

The same result can be obtained by considering the Lagrange functional

$$L(q, u; \psi) = J(q, u(q)) + \langle \psi, G(q, u) \rangle$$

with

$$L_u(q, u; \psi) = 0, \quad G_q(q, u) = 0, \quad \text{and} \quad \nabla J(q) = L_q(q, u; \psi).$$

Indeed, by simple calculations, we can connect these two approaches by setting

$$\psi(q, u) = -[G_u(q, u)^T]^{-1} J_u(q, u)^T.$$

Next, we demonstrate this technique on an inverse problem based on an elliptic initial-boundary-value problem. Let  $u(x, y)$  be the solution of the following problem, which we refer to as (P)

$$\begin{aligned} u_{xx} + u_{yy} &= 0, & x, y &\in (0, 1), \\ u(1, y) &= q(y), & y &\in (0, 1), \\ u_x(0, y) &= 0, & y &\in (0, 1), \\ u(x, 0) &= u(x, 1) = 0, & x &\in (0, 1). \end{aligned}$$

Here we use subscripts  $x, y$  to denote the corresponding partial derivatives. The first two relations constitute the system of equations  $G(q, u) = \bar{G} \cdot (q, u) = 0$ , and the last two ones constitute the feasible set  $Q$ .

Assume that the goal is to solve an inverse problem of estimating  $q(y) \in L_2(0, 1)$  by observing  $b(y) = u(0, y) \in L_2(0, 1)$ , where  $u(x, y) \in L_2((0, 1) \times (0, 1))$  is the (unique) solution of (P) [34]. We can reduce this problem to an optimization problem [34]:

$$\min_q \left\{ \mathfrak{J}(q) := \min_{u: \bar{G} \cdot (q, u) = 0, u \in Q} J(q, u) := J(u) = \int_0^1 |u(0, y) - b(y)|^2 dy \right\}, \quad (18)$$

which can be solved numerically since it is a convex quadratic optimization problem. We can also directly apply Lagrange multipliers principle to (18), see [54]. For that we

introduce Lagrange multipliers  $\psi := (\psi(x, y), \lambda(y))$  and write the Lagrange function:

$$L(q, u; \psi, \lambda) = J(u) + \langle \psi, \bar{G} \cdot (q, u) \rangle = \int_0^1 |u(0, y) - b(y)|^2 dy - \int_0^1 \int_0^1 (u_{xx} + u_{yy}) \psi(x, y) dx dy + \int_0^1 (q(y) - u(1, y)) \lambda(y) dy.$$

To obtain a conjugate problem for  $\psi$ , we need to vary  $L(q, u; \psi)$  in  $\delta u$  satisfying  $u \in Q$ :

$$\delta_u L(q, u; \psi) = 2 \int_0^1 (u(0, y) - b(y)) \delta u(0, y) dy - \int_0^1 \int_0^1 (\delta u_{xx} + \delta u_{yy}) \psi(x, y) dx dy - \int_0^1 \delta u(1, y) \lambda(y) dy, \quad (19)$$

where

$$\begin{aligned} \delta u_x(0, y) &= 0, & y \in (0, 1), \\ \delta u(x, 0) &= \delta u(x, 1) = 0, & x \in (0, 1). \end{aligned}$$

Using the integration by part, from (19), we derive

$$\begin{aligned} \delta_u L(q, u; \psi) &= \int_0^1 (2(u(0, y) - b(y)) - \psi_x(0, y)) \delta u(0, y) dy - \int_0^1 \psi(1, y) \delta u_x(1, y) dy - \int_0^1 \psi(x, 1) \delta u_y(x, 1) dx + \int_0^1 \psi(x, 0) \delta u_y(x, 0) dy + \\ &\quad \int_0^1 \int_0^1 (\psi_{xx} + \psi_{yy}) \delta u(x, y) dx dy + \int_0^1 (\psi_x(1, y) - \lambda(y)) \delta u(1, y) dy. \end{aligned}$$

Consider now the corresponding conjugate problem, which we refer to as (D):

$$\begin{aligned} \psi_{xx} + \psi_{yy} &= 0, & x, y \in (0, 1), \\ \psi_x(0, y) &= 2(u(0, y) - b(y)), & y \in (0, 1), \\ \psi_x(0, y) &= 2(u(0, y) - b(y)), & y \in (0, 1), \\ \psi(x, 0) &= \psi(x, 1) = 0, & x \in (0, 1) \end{aligned}$$

and additional relation between Lagrange multipliers

$$\lambda(y) = \psi_x(1, y), \quad y \in (0, 1). \quad (20)$$

These relations appear since  $\delta_u L(q, u; \psi) = 0$ , and  $\delta u(0, y), \delta u_x(1, y), \delta u(1, y) \in L_2(0, 1); \delta u_y(x, 1), \delta u_y(x, 0) \in L_2(0, 1); \delta u(x, y) \in L_2((0, 1) \times (0, 1))$  are arbitrary.

Since by [48] it holds that

$$\mathfrak{J}(q) = \min_{u: (q, u) \in (P)} J(u) = \min_{u: \bar{G} \cdot (q, u) = 0, u \in Q} J(u) = \min_{u \in Q} \max_{\psi \in (D)} L(q, u; \psi),$$

from the Demyanov–Danskin’s theorem [48], we have<sup>7</sup>

$$\nabla \mathfrak{J}(q) = \nabla_q \min_{u \in Q} \max_{\psi \in (D)} L(q, u; \psi) = L_q(q, u(q); \psi(q)),$$

where  $u(q)$  is the solution of (P) and  $\psi(q)$  is the solution of (D), where

$$\psi_x(0, y) = 2(u(0, y) - b(y)), \quad y \in (0, 1)$$

and  $u(0, y)$  depends on  $q(y)$  via (P) and, at the same time, the pair  $(u(q), \psi(q))$  is the solution of the saddle-point problem

$$\min_{u \in Q} \max_{\psi \in (D)} L(q, u; \psi).$$

Since  $\delta_\psi L(q, u; \psi) = 0$  entails  $\bar{G} \cdot (q, u) = 0$ , that is from (P), if we add  $u \in Q$  and  $\delta_u L(q, u; \psi) = 0$ , then  $u \in Q$  entails (D) as we have shown above. Note also that

$$L_q(q, u(q); \psi(q))(y) = \lambda(y), \quad y \in (0, 1).$$

Hence, by (20), we have that

$$\nabla \mathfrak{J}(q)(y) = \psi_x(1, y), \quad y \in (0, 1).$$

Thus we reduced the calculation of  $\nabla \mathfrak{J}(q)(y)$  to the solution of two correct initial-boundary-value problems for elliptic equation on a square, namely problems (P) and (D) [34].

This result can be also interpreted in a slightly different manner if we introduce a linear operator

$$A : q(y) := u(1, y) \mapsto u(0, y).$$

Here  $u(x, y)$  is the solution of problem (P). It was shown in [34] that

$$A : L_2(0, 1) \rightarrow L_2(0, 1).$$

The conjugate operator is [34]

$$A^* : p(y) := \psi_x(0, y) \mapsto \psi_x(1, y), \quad A^* : L_2(0, 1) \rightarrow L_2(0, 1).$$

Here  $\psi(x, y)$  is the solution of the conjugate problem (D). Thus, considering

$$\mathfrak{J}(q)(y) = \|Aq - b\|_2^2,$$

we have

$$\nabla \mathfrak{J}(q)(y) = A^*(2(Aq - b)),$$

---

<sup>7</sup>The same result in a more simple situation (without additional constraint  $u \in Q$ ) we considered at the beginning of this section. In that case we do not apply Demyanov–Danskin’s theorem and use the inverse function theorem.

which completely corresponds to the scheme as described above:

1. Based on  $q(y)$  we solve (P) and obtain  $u(0, y) = Aq(y)$  and define  $p(y) = 2(u(0, y) - b(y))$ .

2. Based on  $p(y)$  we solve (D) and calculate  $\nabla \mathfrak{J}(q)(y) = A^*p(y) = \psi_x(1, y)$ .

Summarizing, the inexactness in the gradient  $\nabla \mathfrak{J}(q)$  arises since we can solve (P) and (D) only numerically up to some accuracy.

The described above technique can be applied to many different inverse problems [34] and optimal control problems [54]. Note that, for optimal control problems, in practice another strategy is widely used. Namely, instead of approximate calculation of the gradient, optimization problem is replaced by an approximate one (for example, by using finite-differences schemes). For this approximate (finite-dimensional) problem the gradient is typically available precisely [25]. Moreover, in [25] the described above Lagrangian approach is used to explain the core of automatic differentiation, where the function calculation tree is represented as a system of explicitly solvable interlocking equations.

#### 4. Absolute noise in the gradient

In this section, we consider problem (1) in the absolute noise setting (see (3)), i.e., we assume that the inexact gradient  $\tilde{\nabla}f(x)$  satisfies uniformly in  $x \in Q$  the inequality

$$\|\tilde{\nabla}f(x) - \nabla f(x)\|_2 \leq \delta. \quad (21)$$

We underline that  $Q$  can be unbounded, for example  $\mathbb{R}^n$ . Under this assumption, we present several important relations concerning “inexact smoothness” and “inexact strong convexity”. Then, we present and analyze an accelerated gradient method, study its error accumulation, and propose a stopping rule.

##### 4.1. Auxiliary facts

We start with some auxiliary facts and assumptions. Let  $x_{start}$  be some starting point for an algorithm and assume that there is a constant  $R$  such that

$$\|x_{start} - x^*\|_2 \leq R,$$

where  $x^*$  is a solution to problem (1). If  $x^*$  is not unique we take  $x^*$  that is the closest to  $x_{start}$ . We assume that the function  $f$  has Lipschitz gradient with constant  $L_f$ , i.e., is  $L_f$ -smooth:

$$\forall x, y \in Q, \|\nabla f(x) - \nabla f(y)\|_2 \leq L_f \|x - y\|_2. \quad (22)$$

This implies the inequality

$$\forall x, y \in Q, f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f}{2} \|x - y\|_2^2. \quad (23)$$

In what follows, we use the following simple lemma.

**Lemma 4.1** (Fenchel inequality). *Let  $(\mathcal{E}, \langle \cdot, \cdot \rangle)$  be a Euclidean space, then  $\forall \lambda \in \mathbb{R}_+, \forall u, v \in \mathcal{E}$ ,*

$$\langle u, v \rangle \leq \frac{1}{2\lambda} \|u\|_{\mathcal{E}}^2 + \frac{\lambda}{2} \|v\|_{\mathcal{E}}^2.$$

Let us introduce several constants, which will be used below in this section:

$$L = 2L_f, \quad \delta_1 = \delta, \quad \delta_2 = \frac{\delta^2}{L}.$$

From the  $L_f$ -smoothness assumption, we obtain the following upper bound for the objective through the inexact oracle.

**Claim 4.1.** For all  $x, y \in Q$ , the following estimate holds:

$$f(y) \leq f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{L}{2} \|x - y\|_2^2 + \delta_2,$$

where  $L = 2L_f, \delta_2 = \frac{\delta^2}{2L_f}$ .

**Proof.** The proof is given by the following chain of inequalities:

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f}{2} \|x - y\|_2^2 \\ &\leq f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{1}{2L_f} \|\nabla f(x) - \tilde{\nabla} f(x)\|_2^2 + \frac{L_f}{2} \|x - y\|_2^2 + \frac{L_f}{2} \|x - y\|_2^2 \\ &\leq f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{L}{2} \|x - y\|_2^2 + \delta_2. \end{aligned}$$

□

We also assume that  $f$  is strongly convex with parameter  $\mu \geq 0$ , where the case  $\mu = 0$  corresponds to just convexity of  $f$ . This means that for all  $x, y \in Q$ :

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2 \leq f(y). \quad (24)$$

Based on this assumption and our assumption on the inexactness of the oracle, we can obtain two lower bounds for the objective. The first one is given by the following result.

**Claim 4.2.** For all  $x, y \in Q$ , the following estimate holds:

$$f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2 - \delta_1 \|x - y\|_2 \leq f(y),$$

where  $\delta_1 = \delta$ .

**Proof.** Using the Cauchy inequality and (24) we obtain:

$$\begin{aligned}
f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2 \\
&= f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2 - \langle \tilde{\nabla} f(x) - \nabla f(x), y - x \rangle \\
&\geq f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2 - \|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \|x - y\|_2 \\
&\geq f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2 - \delta_1 \|x - y\|_2
\end{aligned}$$

□

For the second estimate, we assume that  $\mu \neq 0$  and introduce

$$\delta_3 = \frac{\delta^2}{\mu}, \quad \mu \neq 0.$$

**Claim 4.3.** For all  $x, y \in Q$ , if in (24)  $\mu \neq 0$ , the following estimate holds:

$$f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{\mu}{4} \|y - x\|_2^2 - \delta_3 \leq f(y),$$

where  $\delta_3 = \frac{\delta^2}{\mu}$ .

**Proof.** Clearly,

$$\begin{aligned}
&f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{\mu}{4} \|x - y\|_2^2 - \delta_3 \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \langle \tilde{\nabla} f(x) - \nabla f(x), y - x \rangle + \frac{\mu}{4} \|x - y\|_2^2 - \delta_3.
\end{aligned}$$

Using Lemma 4.1, we obtain:

$$\begin{aligned}
&f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{\mu}{4} \|x - y\|_2^2 - \delta_3 \\
&\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\delta^2}{\mu} + \frac{\mu}{4} \|x - y\|_2^2 + \frac{\mu}{4} \|y - x\|_2^2 - \delta_3 \leq f(y).
\end{aligned}$$

□

To unify the derivations based on Claims 4.2 and 4.3, we use the notation  $\mu_\tau$ ,  $\tau \in \{1, 2\}$ , where  $\tau = 1$  and  $\mu_1 = \mu$  correspond to the bound in Claim 4.2 and  $\tau = 2$  and  $\mu_2 = \frac{\mu}{2}$  correspond to the bound in Claim 4.3 and the case when  $\mu \neq 0$ .

#### 4.2. Similar Triangles Method and its properties

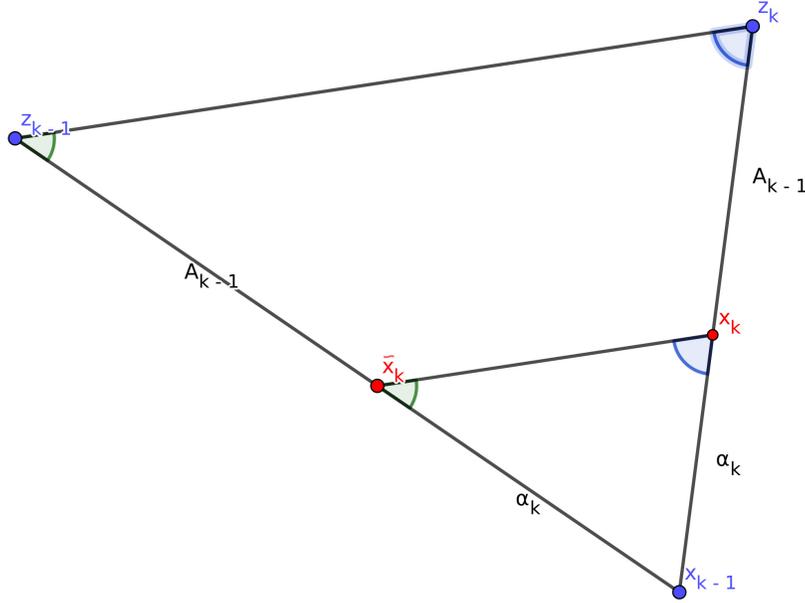
In this section, we introduce a variant of accelerated gradient method called Similar Triangles Method (STM). The design of STM is similar to that of the algorithm in [30] with the main difference being that here we use inexact gradient with absolute inexactness instead of exact gradient. This change required us to modify accordingly the analysis in order to take into account the presence of absolute inexactness in the gradient and possible unboundedness of the feasible set  $Q$ .

---

**Algorithm 1** STM  $(L, \mu_\tau, x_{start}), Q \subseteq \mathbb{R}^n$ 


---

- 1: **Input:** Starting point  $x_{start}$ , number of steps  $N$ .
  - 2: **Set**  $\tilde{x}_0 = x_{start}, \alpha_0 = \frac{1}{L}, A_0 = \frac{1}{L}$ .
  - 3: **Set**  $\psi_0(x) = \frac{1}{2}\|x - \tilde{x}_0\|_2^2 + \alpha_0 \left( f(\tilde{x}_0) + \langle \tilde{\nabla} f(\tilde{x}_0), x - \tilde{x}_0 \rangle + \frac{\mu_\tau}{2}\|x - \tilde{x}_0\|_2^2 \right)$ .
  - 4: **Set**  $z_0 = \operatorname{argmin}_{y \in Q} \psi_0(y), x_0 = z_0$ .
  - 5: **for**  $k = 1 \dots N$  **do**
  - 6:     Find  $\alpha_k$  from  $(1 + \mu_\tau A_{k-1})(A_{k-1} + \alpha_k) = L\alpha_k^2$ ,
  - 7:
  - 8:     or equivalently  $\alpha_k = \frac{1 + \mu_\tau A_{k-1}}{2L} + \sqrt{\frac{(1 + \mu_\tau A_{k-1})^2}{4L^2} + \frac{A_{k-1}(1 + \mu_\tau A_{k-1})}{L}}$ ,
  - 9:      $A_k = A_{k-1} + \alpha_k$ ,
  - 10:      $\tilde{x}_k = \frac{A_{k-1}x_{k-1} + \alpha_k z_{k-1}}{A_k}$ ,
  - 11:      $\psi_k(x) = \psi_{k-1}(x) + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), x - \tilde{x}_k \rangle + \frac{\mu_\tau}{2}\|x - \tilde{x}_k\|_2^2 \right)$ ,
  - 12:      $z_k = \operatorname{argmin}_{y \in Q} \psi_k(y)$ ,
  - 13:      $x_k = \frac{A_{k-1}x_{k-1} + \alpha_k z_k}{A_k}$ .
  - 14: **end for**
  - 15: **Output:**  $x_N$ .
- 



**Figure 1.** Geometry of Similar Triangles Method, Algorithm 1.

Figure 1 illustrates the iterates of the algorithm and justifies the name Similar Triangles Method (STM): by construction  $x_k - \tilde{x}_k = \frac{\alpha_k}{A_k}(z_k - z_{k-1})$ , i.e., the triangles  $(z_{k-1}, x_{k-1}, z_k)$  and  $(\tilde{x}_k, x_{k-1}, x_k)$  are similar. When  $Q = \mathbb{R}^n$ , the main step of the

algorithm can be simplified to

$$z_k = z_{k-1} - \frac{\alpha_k}{1 + A_k \mu_\tau} \left( \tilde{\nabla} f(\tilde{x}_k) + \mu_\tau (z_{k-1} - \tilde{x}_k) \right)$$

using the first-order optimality condition in the definition of the point  $z_k$ . This method is quite simple to implement, since it requires only one projection, which can be eliminated in the absence of constraints, and also has a geometric interpretation. Functions  $\psi_k(x)$  contain first-order information, and are also chosen in such a way that the inequalities guaranteed by convexity or strong convexity can be used to estimate the objective from below, providing an estimating functions sequence. Moreover, since the functions  $\psi_k(x)$  accumulate the first-order information from the previous iterations, the update of the variable  $z_k$  can be seen as a momentum step that leads to the accelerated convergence rate. As it will be seen in Remark 6.2 and Section 6.1, this method can be modified for composite nonsmooth optimization problems and stochastic problems.

In the analysis, we use the following identities that easily follow from the construction of the algorithm:

$$\begin{aligned} A_k(x_k - \tilde{x}_k) &= \alpha_k(z_k - \tilde{x}_k) + A_{k-1}(x_{k-1} - \tilde{x}_k), \\ \frac{1 + \mu_\tau A_{k-1}}{2A_k} \|z_k - z_{k-1}\|_2^2 &= \frac{L}{2} \|x_k - \tilde{x}_k\|_2^2, \\ A_{k-1} \|\tilde{x}_k - x_{k-1}\|_2 &= \alpha_k \|\tilde{x}_k - z_{k-1}\|_2. \end{aligned} \quad (25)$$

The following is the main technical result which will be used later in the analysis.

**Lemma 4.2.** *For all  $k \geq 1$ , the following inequality holds:*

$$\begin{aligned} \psi_k(z_k) &\geq \psi_{k-1}(z_{k-1}) + \frac{1 + \mu_\tau A_{k-1}}{2} \|z_k - z_{k-1}\|_2^2 \\ &\quad + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle + \frac{\mu_\tau}{2} \|z_k - \tilde{x}_k\|_2^2 \right). \end{aligned}$$

**Proof.** By the definition of  $\psi_k$ , we have

$$\begin{aligned} \psi_k(z_k) &= \psi_{k-1}(z_k) + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle + \frac{\mu_\tau}{2} \|z_k - \tilde{x}_k\|_2^2 \right) \\ &= \frac{1}{2} \|z_k - \tilde{x}_0\|_2^2 + \sum_{j=0}^{k-1} \alpha_j \left( f(\tilde{x}_j) + \langle \tilde{\nabla} f(\tilde{x}_j), z_k - \tilde{x}_j \rangle + \frac{\mu_\tau}{2} \|z_k - \tilde{x}_j\|_2^2 \right) \\ &\quad + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle + \frac{\mu_\tau}{2} \|z_k - \tilde{x}_k\|_2^2 \right). \end{aligned} \quad (26)$$

Further, by construction, the function  $\psi_{k-1}$  has its minimum at the point  $z_{k-1}$ , which implies, by the optimality condition,

$$\begin{aligned} \langle \nabla \psi_{k-1}(z_{k-1}), z_k - z_{k-1} \rangle &\geq 0 \\ \Leftrightarrow \langle z_{k-1} - \tilde{x}_0, z_k - z_{k-1} \rangle &\geq \sum_{j=0}^{k-1} \alpha_j \langle \tilde{\nabla} f(\tilde{x}_j) + \mu_\tau (z_{k-1} - \tilde{x}_j), z_{k-1} - z_k \rangle. \end{aligned} \quad (27)$$

We also have the identity

$$\frac{1}{2}\|z_k - \tilde{x}_0\|_2^2 = \frac{1}{2}\|z_{k-1} - z_k\|_2^2 + \frac{1}{2}\|z_{k-1} - \tilde{x}_0\|_2^2 + \langle z_{k-1} - \tilde{x}_0, z_k - z_{k-1} \rangle. \quad (28)$$

Combining the above, we have

$$\begin{aligned} \psi_k(z_k) &\stackrel{(26),(28)}{=} \frac{1}{2}\|z_{k-1} - \tilde{x}_0\|_2^2 + \langle z_{k-1} - \tilde{x}_0, z_k - z_{k-1} \rangle + \frac{1}{2}\|z_{k-1} - z_k\|_2^2 \\ &\quad + \sum_{j=0}^{k-1} \alpha_j \left( f(\tilde{x}_j) + \langle \tilde{\nabla} f(\tilde{x}_j), z_k - \tilde{x}_j \rangle + \frac{\mu_\tau}{2}\|z_k - \tilde{x}_j\|_2^2 \right) \\ &\stackrel{(27)}{\geq} \sum_{j=0}^{k-1} \alpha_j \left( \langle \tilde{\nabla} f(\tilde{x}_j) + \mu_\tau(z_{k-1} - \tilde{x}_j), z_{k-1} - z_k \rangle \right) \\ &\quad + \sum_{j=0}^{k-1} \alpha_j \left( f(\tilde{x}_j) + \langle \tilde{\nabla} f(\tilde{x}_j), z_k - \tilde{x}_j \rangle + \frac{\mu_\tau}{2}\|z_k - \tilde{x}_j\|_2^2 \right) \\ &\quad + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle + \frac{\mu_\tau}{2}\|z_k - \tilde{x}_k\|_2^2 \right) \\ &\quad + \frac{1}{2}\|z_{k-1} - \tilde{x}_0\|_2^2 + \frac{1}{2}\|z_{k-1} - z_k\|_2^2. \end{aligned}$$

Applying the identity

$$\langle z_{k-1} - \tilde{x}_j, z_{k-1} - z_k \rangle = \frac{1}{2}\|z_{k-1} - \tilde{x}_j\|_2^2 + \frac{1}{2}\|z_k - z_{k-1}\|_2^2 - \frac{1}{2}\|z_k - \tilde{x}_j\|_2^2,$$

and the definition of the sequence  $\{A_k\}$ , we finally get

$$\begin{aligned} \psi_k(z_k) &\geq \frac{1}{2}\|z_{k-1} - \tilde{x}_0\|_2^2 + \frac{1 + \mu_\tau A_{k-1}}{2}\|z_{k-1} - z_k\|_2^2 \\ &\quad + \sum_{j=0}^{k-1} \alpha_j \left( f(\tilde{x}_j) + \langle \tilde{\nabla} f(\tilde{x}_j), z_{k-1} - \tilde{x}_j \rangle + \frac{\mu_\tau}{2}\|z_{k-1} - \tilde{x}_j\|_2^2 \right) \\ &\quad + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle + \frac{\mu_\tau}{2}\|z_k - \tilde{x}_k\|_2^2 \right) \\ &= \psi_{k-1}(z_{k-1}) + \frac{1 + \mu_\tau A_{k-1}}{2}\|z_k - z_{k-1}\|_2^2 \\ &\quad + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle + \frac{\mu_\tau}{2}\|z_k - \tilde{x}_k\|_2^2 \right) \end{aligned}$$

□

**Remark 4.1.** In the case when  $\mu = 0$ , we obtain the following particular case of the result of Lemma 4.2:

$$\begin{aligned} \psi_k(z_k) &= \psi_{k-1}(z_{k-1}) + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle \right) \\ \Rightarrow \psi_k(z_k) &\geq \psi_{k-1}(z_{k-1}) + \frac{1}{2}\|z_k - z_{k-1}\|_2^2 + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle \right). \end{aligned}$$

We finish this subsection by a series of technical results that estimate the growth of the sequence  $\{A_k\}$  and related sequences.

**Claim 4.4.** If  $\mu \neq 0$ , then for all  $k \geq 1$  the following inequality holds:

$$A_k \geq A_{k-1} \lambda_{\mu_\tau, L},$$

where

$$\theta_{\mu_\tau, L} = \frac{\mu_\tau}{L}, \quad \lambda_{\mu_\tau, L} = \left(1 + \frac{1}{2}\theta_{\mu_\tau, L} + \sqrt{\theta_{\mu_\tau, L}}\right).$$

**Proof.** Using the relation between  $\alpha_k, A_k, A_{k-1}$ :

$$A_k(1 + \mu_\tau A_{k-1}) = L\alpha_k^2,$$

we obtain a quadratic equation for  $A_k$ :

$$\begin{aligned} A_k(1 + \mu_\tau A_{k-1}) &= L(A_k - A_{k-1})^2, \\ A_k(1 + \mu_\tau A_{k-1}) &= LA_k^2 - 2LA_{k-1}A_k + LA_{k-1}^2, \\ LA_k^2 - A_k(1 + \mu_\tau A_{k-1} + 2LA_{k-1}) &= 0 + LA_{k-1}^2. \end{aligned}$$

Solving this equation, we get

$$A_k \geq A_{k-1} \left(1 + \frac{\mu_\tau + 1}{L} + \sqrt{\frac{\mu_\tau}{L}}\right) \geq A_{k-1} \left(1 + \frac{\mu_\tau}{2L} + \sqrt{\frac{\mu_\tau}{L}}\right).$$

□

Using that, for  $x < 1$ ,  $1 + x \geq e^{\frac{x}{2}}$ , we obtain the following result.

**Corollary 4.3.**

$$\lambda_{\mu_\tau, L} = \left(1 + \frac{\mu_\tau}{2L} + \sqrt{\theta_{\mu_\tau, L}}\right) \geq \left(1 + \sqrt{\theta_{\mu_\tau, L}}\right) \geq e^{\frac{1}{2}\sqrt{\theta_{\mu_\tau, L}}}.$$

**Claim 4.5.** If  $\mu \neq 0$ , then for all  $k \geq 1$  the following inequality holds:

$$\frac{1}{A_k} \sum_{j=0}^k A_j \leq 1 + \sqrt{\frac{L}{\mu_\tau}}.$$

**Proof.** Using the previous claim, we get  $A_k \geq A_{k-j} \lambda_{\mu_\tau, L}^j$ , and, hence,  $\frac{A_{k-j}}{A_k} \leq \lambda_{\mu_\tau, L}^{-j}$ . This gives

$$\frac{1}{A_k} \sum_{j=0}^k A_j \leq \sum_{j=0}^k \lambda_{\mu_\tau, L}^{-j} = \frac{\lambda_{\mu_\tau, L}^{k+1} - 1}{\lambda_{\mu_\tau, L}^{k+1} - \lambda_{\mu_\tau, L}^k} \leq \frac{\lambda_{\mu_\tau, L}}{\lambda_{\mu_\tau, L} - 1} \leq 1 + \sqrt{\frac{L}{\mu_\tau}}.$$

□

**Claim 4.6.** If  $\mu = 0$ , then for all  $k \geq 1$ ,

$$A_k \geq \frac{(k+1)^2}{4L}.$$

**Proof.** If  $\mu = 0$ , then  $A_k = L\alpha_k^2$ , and, solving the quadratic equation, we get

$$\alpha_k = \frac{1 + \sqrt{1 + 4L^2\alpha_{k-1}^2}}{2L} \geq \frac{1 + 2L\alpha_{k-1}}{2L} = \frac{1}{2L} + \alpha_{k-1}.$$

Then, by induction, it is easy to see that

$$\alpha_k \geq \frac{k+1}{2L} \Rightarrow A_k = L\alpha_k^2 \geq \frac{(k+1)^2}{2L}.$$

□

**Claim 4.7.** If  $\mu = 0$ , then for all  $k \geq 1$

$$\frac{1}{A_k} \sum_{j=0}^k A_j \leq k + 1.$$

**Proof.** The proof follows from the simple calculations since  $\{A_k\}$  is non-decreasing:

$$\frac{1}{A_k} \sum_{j=0}^k A_j \leq \frac{1}{A_k} (k+1)A_k = k + 1.$$

□

### 4.3. Convergence rates under the absolute inexactness

In this section we obtain main convergence rate results for Algorithm 1. We will use the following sequence

$$\tilde{R}_k = \max_{0 \leq j \leq k} \{\|z_j - x^*\|_2, \|x_j - x^*\|_2, \|\tilde{x}_j - x^*\|_2\}. \quad (29)$$

**Proposition 4.4.** *The sequences  $\{x_k\}$ ,  $\{\tilde{x}_k\}$ ,  $\{z_k\}$  generated by Algorithm 1 satisfy for all  $k \geq 0$  the inequality*

$$A_k f(x_k) \leq \psi_k(z_k) + \delta_2 \sum_{j=0}^k A_j + \delta_1 \sum_{j=1}^k \alpha_j \|\tilde{x}_j - z_{j-1}\|_2.$$

**Proof.** We prove the result by induction. The induction basis for  $k = 0$  follows from the facts that  $A_0 = \alpha_0 = \frac{1}{L}$  and

$$\psi_0(x) = \alpha_0 \left( f(\tilde{x}_0) + \langle \tilde{\nabla} f(\tilde{x}_0), x - \tilde{x}_0 \rangle + \frac{\mu_\tau}{2} \|x - \tilde{x}_0\|_2^2 \right) + \frac{1}{2} \|x - \tilde{x}_0\|_2^2,$$

which imply, by Claim 4.1, and since  $x_0 = z_0$ , that

$$\begin{aligned} f(x_0) &\leq f(\tilde{x}_0) + \langle \tilde{\nabla} f(\tilde{x}_0), x_0 - \tilde{x}_0 \rangle + \frac{L}{2} \|x_0 - \tilde{x}_0\|_2^2 + \delta_2 \\ &= L\psi_0(z_0) - \frac{\mu_\tau}{2} \|z_0 - \tilde{x}_0\|_2^2 + \delta_2 \leq L\psi_0(z_0) + \delta_2. \end{aligned}$$

To make the induction step, we start from the following corollary of Claim 4.1:

$$\begin{aligned} &A_k f(x_k) - A_{k-1} \delta_1 \|x_{k-1} - \tilde{x}_k\|_2 \\ &\leq A_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), x_k - \tilde{x}_k \rangle + \frac{L}{2} \|x_k - \tilde{x}_k\|_2^2 + \delta_2 \right) - A_{k-1} \delta_1 \|x_{k-1} - \tilde{x}_k\|_2. \end{aligned}$$

Using equations (25), this gives

$$\begin{aligned} &A_k f(x_k) - A_{k-1} \delta_1 \|x_{k-1} - \tilde{x}_k\|_2 \\ &\leq A_{k-1} \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), x_{k-1} - \tilde{x}_k \rangle \right) \\ &\quad + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle \right) \\ &\quad + \frac{(1 + \mu_\tau A_{k-1})}{2} \|z_k - z_{k-1}\|_2^2 + A_k \delta_2 - A_{k-1} \delta_1 \|x_{k-1} - \tilde{x}_k\|_2 \\ &\leq A_{k-1} f(x_{k-1}) + \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle \right) \\ &\quad + \frac{1 + \mu_\tau A_{k-1}}{2} \|z_k - z_{k-1}\|_2^2 + A_k \delta_2. \end{aligned}$$

By the induction hypothesis and since  $\frac{\mu_\tau}{2} \|z_k - \tilde{x}_k\|_2^2 \geq 0$ , we further obtain

$$\begin{aligned} &A_k f(x_k) - A_{k-1} \delta_1 \|x_{k-1} - \tilde{x}_k\|_2 \leq \psi_{k-1}(z_{k-1}) + \delta_2 \sum_{j=0}^{k-1} A_j + \delta_1 \sum_{j=1}^{k-1} \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 \\ &+ \alpha_k \left( f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle + \frac{\mu_\tau}{2} \|z_k - \tilde{x}_k\|_2^2 \right) + \frac{1 + \mu_\tau A_{k-1}}{2} \|z_k - z_{k-1}\|_2^2 + A_k \delta_2. \end{aligned}$$

Using Lemma 4.2, we get

$$\begin{aligned} &A_k f(x_k) \leq A_{k-1} \delta_1 \|x_{k-1} - \tilde{x}_k\|_2 + \psi_k(z_k) + \delta_2 \sum_{j=0}^k A_j + \delta_1 \sum_{j=1}^{k-1} \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 \\ &\stackrel{(25)}{=} \psi_k(z_k) + \delta_2 \sum_{j=0}^k A_j + \delta_1 \sum_{j=1}^{k-1} \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 + \alpha_k \delta_1 \|\tilde{x}_k - z_{k-1}\|_2, \end{aligned}$$

which finishes the induction step and the proof.  $\square$

Using the definition of  $\{\tilde{R}_k\}$  and  $\{A_k\}$ , we obtain the following simple corollary of

the above proposition:

$$\begin{aligned}
A_k f(x_k) &\leq \psi_k(z_k) + \delta_2 \sum_{j=0}^k A_j + \delta_1 \sum_{j=1}^k \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 \\
&\leq \psi_k(z_k) + \delta_2 \sum_{j=0}^k A_j + \delta_1 \sum_{j=1}^k \alpha_j (\|z_{k-1} - x^*\|_2 + \|\tilde{x}_k - x^*\|_2) \\
&\leq \psi_k(z_k) + \delta_2 \sum_{j=0}^k A_j + 2\tilde{R}_k \delta_1 A_k. \tag{30}
\end{aligned}$$

We note that the above estimates hold both in the case of  $\mu \neq 0$  and in the case of  $\mu = 0$ .

The proof of the following result repeats verbatim the proof of Proposition 4.4, except for Claim 4.2 being replaced by Claim 4.3.

**Proposition 4.5.** *If  $\mu \neq 0$ , the sequences  $\{x_k\}$ ,  $\{\tilde{x}_k\}$ ,  $\{z_k\}$  generated by Algorithm 1 satisfy for all  $k \geq 0$  the inequality*

$$A_k f(x_k) \leq \psi_k(z_k) + \delta_2 \sum_{j=0}^k A_j + \delta_3 \sum_{j=0}^{k-1} A_j.$$

**Proposition 4.6.** *Assume that the oracle error  $\delta$  satisfies  $\delta = 0$  and that  $\|\tilde{x}_0 - x^*\|_2 \leq R$  for some  $R$ . Then, for all  $k \geq 1$ ,  $\tilde{R}_k \leq R$ .*

**Proof.** We first prove that, for all  $k \geq 0$ ,  $\|z_k - x^*\|_2 \leq R$ . Let us fix  $k \geq 0$ . By Proposition 4.4, we have  $A_k f(x_k) \leq \psi_k(z_k)$ . Further,  $\psi_k(x)$  is strongly convex with the constant at least 1. At the same time, by the strong convexity of  $f$ , we have

$$f(\tilde{x}_j) + \langle \nabla f(\tilde{x}_k), x^* - \tilde{x}_k \rangle + \frac{\mu_\tau}{2} \|x^* - \tilde{x}_k\|_2^2 \leq f(x^*) \leq f(x_k).$$

Using these three facts and the definition of  $\psi_k(x)$ , we obtain:

$$\begin{aligned}
\frac{1}{2}\|z_k - x^*\|_2^2 &= \frac{1}{2}\|z_k - x^*\|_2^2 + A_k f(x_k) - A_k f(x_k) \\
&\leq \psi_k(z_k) + \frac{1}{2}\|z_k - x^*\|_2^2 - A_k f(x_k) \\
&\leq \psi_k(x^*) - A_k f(x_k) \\
&\leq \sum_{j=0}^k \alpha_j \left( f(\tilde{x}_j) + \langle \nabla f(\tilde{x}_k), x^* - \tilde{x}_k \rangle + \frac{\mu_\tau}{2} \|x^* - \tilde{x}_k\|_2^2 \right) \\
&\quad + \frac{1}{2} \|x^* - \tilde{x}_0\|_2^2 - A_k f(x_k) \\
&\leq \sum_{j=0}^k \alpha_j f(x^*) + \frac{1}{2} R^2 - A_k f(x_k) \\
&= A_k (f(x^*) - f(x_k)) + \frac{1}{2} R^2 \leq \frac{1}{2} R^2.
\end{aligned}$$

For the remaining two sequences,  $\{\tilde{x}_k\}$  and  $\{x_k\}$  the proof is organized by induction. Clearly,  $\|\tilde{x}_0 - x^*\| \leq R$ . Since, by construction,  $x_0 = z_0$ , we have  $\|x_0 - x^*\| \leq R$ . Then, by construction of the algorithm and the induction hypothesis, we have

$$\begin{aligned}
\|x_k - x^*\|_2 &= \left\| \frac{A_{k-1}}{A_k} (x_{k-1} - x^*) + \frac{\alpha_k}{A_k} (z_k - x^*) \right\|_2 \\
&\leq \frac{A_{k-1}}{A_k} \|x_{k-1} - x^*\|_2 + \frac{\alpha_k}{A_k} \|z_k - x^*\|_2 \leq R.
\end{aligned}$$

In the same way, we obtain  $\|\tilde{x}_k - x^*\| \leq R$  using the definition

$$\tilde{x}_k = \frac{\alpha_k}{A_k} z_{k-1} + \frac{A_{k-1}}{A_k} x_{k-1}.$$

□

Using the above results, we obtain the following convergence rate result for the STM algorithm.

**Theorem 4.7** (Main Theorem). *Let  $\|\tilde{x}_0 - x^*\|_2 \leq R$  for some  $R$ . If  $\mu \neq 0$ , the sequences  $\{x_k\}$ ,  $\{\tilde{x}_k\}$ ,  $\{z_k\}$  generated by Algorithm 1 satisfy for all  $N \geq 0$  the inequalities*

$$\begin{aligned}
f(x_N) - f(x^*) &\leq LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu_1}{L}}N\right) + \left(1 + \sqrt{\frac{L}{\mu_1}}\right) \delta_2 + 3\tilde{R}_N \delta_1, \\
f(x_N) - f(x^*) &\leq LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu_2}{L}}N\right) + \left(1 + \sqrt{\frac{L}{\mu_2}}\right) \delta_2 + \left(1 + \sqrt{\frac{L}{\mu_2}}\right) \delta_3.
\end{aligned}$$

If  $\mu = 0$ , the sequences  $\{x_k\}$ ,  $\{\tilde{x}_k\}$ ,  $\{z_k\}$  generated by Algorithm 1 satisfy for all  $N \geq 0$

the inequality

$$f(x_N) - f(x^*) \leq \frac{4LR^2}{N^2} + 3\tilde{R}_N\delta_1 + N\delta_2, \quad (31)$$

where the sequence  $\{\tilde{R}_k\}$  is defined in (29).

**Proof.** The proofs of the first and second inequalities are nearly the same with the only difference that the proof of the first inequality is based on Proposition 4.4 and Claim 4.2, whereas the proof of the second inequality is based on Proposition 4.5 and Claim 4.3. Thus, we give only the proof of the first inequality. From (30), by the definition of  $\{z_N\}$  and  $\{\psi_N(\cdot)\}$ , and Claim 4.2, we have

$$\begin{aligned} A_N f(x_N) &\leq \psi_N(z_N) + \delta_2 \sum_{j=0}^N A_j + 2\tilde{R}_N\delta_1 A_N \\ &\leq \frac{1}{2} \|x^* - \tilde{x}_0\|_2^2 + \delta_2 \sum_{j=0}^N A_j \\ &\quad + 2\tilde{R}_N\delta_1 A_N + \sum_{j=0}^N \alpha_k (f(\tilde{x}_j) + \langle \tilde{\nabla} f(\tilde{x}_j), x^* - \tilde{x}_j \rangle + \frac{\mu_1}{2} \|x^* - \tilde{x}_j\|_2^2) \\ &\leq \delta_2 \sum_{j=0}^N A_j + 2\tilde{R}_N\delta_1 A_N + \sum_{j=0}^N \alpha_k (\tilde{R}_j\delta_1 + f(x^*)) + \frac{1}{2} R^2 \\ &= \delta_2 \sum_{j=0}^N A_j + 3\tilde{R}_N\delta_1 A_N + A_N f(x^*) + \frac{1}{2} R^2 \\ &\iff f(x_N) - f(x^*) \leq \frac{R^2}{2A_N} + \delta_2 \frac{1}{A_N} \sum_{j=0}^N A_j + 3\tilde{R}_N\delta_1. \end{aligned}$$

Using Claim 4.4 with Corollary 4.3 and Claim 4.5 we get:

$$f(x_N) - f(x^*) \leq LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu_1}{L}}N\right) + \left(1 + \sqrt{\frac{L}{\mu_1}}\right) \delta_2 + 3\tilde{R}_N\delta_1.$$

Repeating the same steps and using Claim 4.6 and 4.7, we prove the third inequality.  $\square$

Commenting on the results obtained in Theorem 4.7, we can conclude that in the case of strong convexity and the presence of absolute noise, STM converges in terms of the objective value up to some limiting accuracy. Namely, the convergence rate bound is the sum of the convergence rate of the optimal method for the class of strongly convex and Lipschitz-smooth problems and the term characterizing the limiting error caused by the noise is

$$\left(1 + \sqrt{\frac{L}{\mu_2}}\right) (\delta_2 + \delta_3).$$

In the case when  $\mu = 0$ , we obtain a weaker convergence rate statement since in the estimate we see a linear accumulation of the noise in the term  $N\delta_2$ , as well as in the term  $3\tilde{R}_N\delta_1$  (Note that Proposition 4.6 gives the estimate for  $\tilde{R}_N$  in the absence of noise). This motivates us to use the regularization technique to make a reduction of the convex case to the strongly convex case, which is considered in the next Remark 4.2. Another way to deal with the noise accumulation is to introduce a stopping rule, which is done below in Section 4.4.

**Remark 4.2.** We can make a reduction of the setting when  $\mu = 0$  to the setting when  $\mu \neq 0$ . Indeed, suppose that  $\mu = 0$  and consider the following regularized problem:

$$\min_{x \in Q} \left\{ f^\mu(x) := f(x) + \frac{\mu}{2} \|x - \tilde{x}_0\|_2^2 \right\}.$$

Then, we have

$$\begin{aligned} \nabla f^\mu(x) &= \nabla f(x) + \mu(x - \tilde{x}_0), \\ \tilde{\nabla} f^\mu(x) &= \tilde{\nabla} f(x) + \mu(x - \tilde{x}_0), \\ \|\tilde{\nabla} f^\mu(x) - \nabla f^\mu(x)\|_2 &= \|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \delta. \end{aligned}$$

Clearly,  $f^\mu(x)$  has Lipschitz gradient. Indeed,  $\forall x, y \in Q$ :

$$\begin{aligned} \|\nabla f^\mu(x) - \nabla f^\mu(y)\|_2 &= \|(\nabla f(x) - \nabla f(y)) + \mu(x - y)\|_2 \\ &\leq \|\nabla f(x) - \nabla f(y)\|_2 + \mu\|x - y\|_2 \\ &\leq L_f\|x - y\|_2 + \mu\|x - y\|_2 \leq (L_f + \mu)\|x - y\|_2. \end{aligned}$$

Since  $\mu \leq L$ , we have that  $f^\mu(x)$  is  $L^\mu$ -smooth with  $L^\mu = 4L_f = 2L$ . Moreover,  $f^\mu(x)$  is strongly convex and we can apply the derivations corresponding to the case  $\tau = 2$ . Using Theorem 4.7, and setting  $x_\mu^* = \operatorname{argmin}_{x \in Q} f^\mu(x)$ ,  $R_\mu$  s.t.  $\|x_\mu^* - \tilde{x}_0\|_2 \leq R_\mu$ , we obtain the following inequalities

$$\begin{aligned} f^\mu(x_k) - f^\mu(x_\mu^*) &\leq 2LR_\mu^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{4L}}k\right) + \left(1 + \sqrt{\frac{4L}{\mu}}\right) \left(\frac{1}{2L} + \frac{1}{\mu}\right) \delta^2, \\ f^\mu(x_\mu^*) &\leq f(x^*) + \frac{\mu}{2}R^2. \end{aligned}$$

Translating this to the original objective  $f$ , we obtain

$$\begin{aligned} f(x_k) - f(x^*) &\leq f^\mu(x_k) - f(x^*) \\ &\leq f^\mu(x_k) - f(x_\mu^*) + \frac{\mu}{2}R^2 \\ &\leq 2LR_\mu^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{4L}}k\right) + \left(1 + \sqrt{\frac{4L}{\mu}}\right) \left(\frac{1}{2L} + \frac{1}{\mu}\right) \delta^2 + \frac{\mu}{2}R^2. \end{aligned}$$

By the strong convexity of the function  $f^\mu$ , we get:

$$\begin{aligned} f(x^*) + \frac{\mu}{2}R_\mu^2 &\leq f(x_\mu^*) + \frac{\mu}{2}R_\mu^2 = f^\mu(x_\mu^*) \leq f^\mu(x^*) = f(x^*) + \frac{\mu}{2}R^2 \Rightarrow \\ &R_\mu \leq R. \end{aligned}$$

Finally, we get the convergence rate as follows:

$$f(x_k) - f(x^*) \leq 2LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2L}}k\right) + \left(1 + \sqrt{\frac{4L}{\mu}}\right) \left(\frac{1}{2L} + \frac{1}{\mu}\right) \delta^2 + \frac{\mu}{2}R^2.$$

To obtain an error  $\varepsilon$  in the r.h.s., we choose  $\mu = \frac{2}{3}\frac{\varepsilon}{R^2}$ .

#### 4.4. Stopping rule under the absolute inexactness

In this subsection, we consider the setting with  $\tau = 1$  and  $\mu = 0$ . In this case, a possible drawback of the convergence rate obtained in Theorem 4.7

$$f(x_N) - f(x^*) \leq \frac{4LR^2}{N^2} + 3\tilde{R}_N\delta_1 + N\delta_2$$

can be that the sequence  $\{\tilde{R}_N\}$  may increase as  $N$  increases. To overcome this, we formulate a certain condition (stopping rule) and prove that if it is satisfied at iteration  $N$ , the algorithm solves problem (1) with certain accuracy, and if it is not satisfied at iteration  $N$ , then  $\tilde{R}_N \leq R$ . Moreover, we estimate the maximum number of iterations to satisfy this condition.

**Theorem 4.8.** *Consider the setting  $\tau = 1$  and  $\mu = 0$  and assume that for some  $R$ ,  $\|\tilde{x}_0 - x^*\|_2 \leq R$ . Let  $\varepsilon > 0$  be the desired solution accuracy. Let  $N$  be the first iteration such that*

$$f(x_N) - f(x^*) \leq \frac{\delta_2}{A_N} \sum_{j=0}^N A_j + 3R\delta_1 + \varepsilon. \quad (32)$$

*Then, for all  $k \in \{0, \dots, N-1\}$ , we have that  $\tilde{R}_k \leq R$ . Moreover,*

$$N \leq N_{\max} := \left\lceil \sqrt{\frac{2LR^2}{\varepsilon}} \right\rceil. \quad (33)$$

**Proof.** Fixing any  $k \geq 0$ , applying Proposition 4.4, the fact that 1-strongly convex function  $\psi_k(\cdot)$  attains its minimum at the point  $z_k$ , the definition of this function, and

Claim 4.2, we obtain

$$\begin{aligned}
\frac{1}{2}\|z_k - x^*\|_2^2 + A_k f(x_k) &\leq \frac{1}{2}\|z_k - x^*\|_2 + \psi_k(z_k) + \delta_2 \sum_{j=0}^k A_j + \delta_1 \sum_{j=1}^k \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 \\
&\leq \psi_k(x^*) + \delta_2 \sum_{j=0}^k A_j + \delta_1 \sum_{j=1}^k \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 \\
&= \frac{1}{2}\|\tilde{x}_0 - x^*\|_2^2 + \sum_{j=0}^k \alpha_j (f(\tilde{x}_j) + \langle \tilde{\nabla} f(\tilde{x}_j), x^* - \tilde{x}_j \rangle) \\
&\quad + \delta_2 \sum_{j=0}^k A_j + \delta_1 \sum_{j=1}^k \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 \\
&\leq \frac{R^2}{2} + A_k f(x^*) + \delta_1 \sum_{j=0}^k \alpha_j \|\tilde{x}_j - x^*\|_2 + \delta_2 \sum_{j=0}^k A_j \\
&\quad + \delta_1 \sum_{j=1}^k \alpha_j \|\tilde{x}_j - z_{j-1}\|_2. \tag{34}
\end{aligned}$$

Whence,

$$\begin{aligned}
\frac{1}{2}\|z_k - x^*\|_2^2 &\leq \frac{R^2}{2} + A_k \left( f(x^*) - f(x_k) + \frac{\delta_1 \alpha_0}{A_k} \|\tilde{x}_0 - x^*\|_2 \right. \\
&\quad \left. + \frac{\delta_1}{A_k} \sum_{j=1}^k \alpha_j (2\|\tilde{x}_j - x^*\|_2 + \|z_{j-1} - x^*\|_2) + \frac{\delta_2}{A_k} \sum_{j=0}^k A_j \right). \tag{35}
\end{aligned}$$

Setting  $k = 0$ , since  $\|\tilde{x}_0 - x^*\|_2 \leq R$  and, by the Theorem assumption, inequality (32) does not hold for  $k \leq N - 1$ , we obtain

$$\begin{aligned}
\frac{1}{2}\|z_0 - x^*\|_2^2 &\leq \frac{R^2}{2} + A_0 \left( f(x^*) - f(x_0) + \frac{\delta_1 \alpha_0}{A_0} R + \delta_2 \right) \\
&\leq \frac{R^2}{2} + A_0 \left( -\delta_2 - 3R\delta_1 - \varepsilon + \frac{\delta_1 \alpha_0}{A_0} R + \delta_2 \right) \leq \frac{R^2}{2},
\end{aligned}$$

where we also used that  $\alpha_0 = A_0$ . Thus, we obtain that  $\|z_0 - x^*\|_2 \leq R$ , and, since  $x_0 = z_0$ , that  $\|x_0 - x^*\|_2 \leq R$ . Hence,  $\tilde{R}_0 \leq R$ . Let us now assume that for some  $k \leq N - 1$ ,  $\tilde{R}_{k-1} \leq R$  (see (29) for the definition of  $\{\tilde{R}_k\}$ ). Then, by the definition of  $\tilde{x}_k$  in Algorithm 1 and convexity of the norm, we have that  $\|\tilde{x}_k - x^*\|_2 \leq R$ . Further,

since  $k \leq N - 1$ , we have that inequality (32) does not hold. Thus, from (35), we have:

$$\begin{aligned} \frac{1}{2} \|z_k - x^*\|_2^2 &\leq \frac{R^2}{2} + A_k \left( f(x^*) - f(x_k) + \frac{\delta_1 \alpha_0 R}{A_k} + \frac{\delta_1}{A_k} \sum_{j=1}^k \alpha_j \cdot 3R + \frac{\delta_2}{A_k} \sum_{j=0}^k A_j \right) \\ &\leq \frac{R^2}{2} + A_k \left( -\frac{\delta_2}{A_k} \sum_{j=0}^k A_j - 3R\delta_1 - \varepsilon + 3R\delta_1 + \frac{\delta_2}{A_k} \sum_{j=0}^k A_j \right) \leq \frac{R^2}{2}. \end{aligned}$$

This implies that  $\|z_k - x^*\|_2 \leq R$ , and, by the definition of  $x_k$  and the convexity of the norm, that  $\|x_k - x^*\|_2 \leq R$ . Hence,  $\tilde{R}_k \leq R$ . In summary, we obtain by induction that for all  $k \in \{0, \dots, N - 1\}$ ,  $\tilde{R}_k \leq R$ . This also implies that  $\|\tilde{x}_N - x^*\|_2 \leq R$ .

We now prove the second statement of the Theorem. Let us assume the opposite, i.e.,  $N > N_{\max}$ . We use (34) with  $k = N - 1$  and obtain, since  $\tilde{R}_{N-1} \leq R$ , that

$$\begin{aligned} f(x_{N-1}) - f(x^*) &\leq \frac{R^2}{2A_{N-1}} + 3R\delta_1 + \frac{\delta_2}{A_{N-1}} \sum_{j=0}^{N-1} A_j \\ &\leq \frac{2LR^2}{N^2} + 3R\delta_1 + \frac{\delta_2}{A_{N-1}} \sum_{j=0}^{N-1} A_j \\ &< \varepsilon + 3R\delta_1 + \frac{\delta_2}{A_{N-1}} \sum_{j=0}^{N-1} A_j, \end{aligned}$$

where we used Claim 4.6 and that  $N > N_{\max}$ . Thus, we see that after  $N - 1$  iterations, inequality (32) holds. This is a contradiction with the definition of  $N$  as the first iteration number for which this inequality holds. This finishes the proof.  $\square$

Combining (32) with Claim 4.7 and the fact that  $N \leq N_{\max}$ , we obtain that

$$f(x_N) - f(x^*) \leq \delta_2(N_{\max} + 1) + 3R\delta_1 + \varepsilon.$$

Thus, if we redefine  $\varepsilon \rightarrow \frac{\varepsilon}{3}$ , and set  $\delta_2 \leq \frac{\varepsilon}{3(N_{\max} + 1)}$ ,  $\delta_1 \leq \frac{\varepsilon}{9R}$ , we guarantee that  $f(x_N) - f(x^*) \leq \varepsilon$ .

**Remark 4.3.** In some situations we have at our disposal the value of  $f(x^*)$  or its estimate. For example, when solving systems of linear equations by reformulating them as minimization problems:

$$\begin{aligned} Ax &= b, \\ \min_x \{f(x) &= \|Ax - b\|_2^2\}, \end{aligned}$$

if a solution exists, we have  $f^* = f(x^*) = 0$ . This allows us, based on (34), to change the inequality (32) to a more adaptive version, which can be checked online and which can be fulfilled much earlier than (32). Such counterpart of (32) reads as

$$f(x_N) - f(x^*) \leq \frac{\delta_2}{A_N} \sum_{j=0}^N A_j + R\delta_1 + \delta_1 \sum_{j=1}^N \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 + \varepsilon. \quad (36)$$

If this inequality is not fulfilled at iteration  $k$ , we have that  $\tilde{R}_k \leq R$ . If it is fulfilled at iteration  $k$ , we obtain that

$$f(x_k) - f(x^*) \leq \delta_2(k+1) + 3R\delta_1 + \varepsilon.$$

Moreover, we also obtain that (36) holds after no more than  $N_{\max}$  iterations.

## 5. Relative noise in the gradient

In this section, we consider problem (1) in the relative noise setting (see (4)), i.e., we assume that the inexact gradient  $\tilde{\nabla}f(x)$  satisfies uniformly in  $x \in Q$  the inequality

$$\|\tilde{\nabla}f(x) - \nabla f(x)\| \leq \alpha \|\nabla f(x)\|_2.$$

As in the previous section, we assume that  $f$  is  $L_f$ -smooth. We also assume that  $f$  is strongly convex with  $\mu \neq 0$  and that  $Q = \mathbb{R}^n$ . For this setting, we analyze a slightly different version of accelerated gradient method, adopted from [50].

---

**Algorithm 2** STM2 ( $L, \mu_\tau, x_{start}$ ),  $Q \subseteq \mathbb{R}^n$

---

- 1: **Input:** Starting point  $x_{start}$ , number of steps  $N$
  - 2: **Set**  $y_0 = u_0 = x_0 = x_{start}$ ,
  - 3: **Set**  $A_0 = \frac{1}{L}$ ,  $\alpha_0 = A_0$ .
  - 4: **for**  $k = 1 \dots N$  **do**
  - 5:     Find  $\alpha_k$  from  $(1 + \mu_2 A_{k-1})(A_{k-1} + \alpha_k) = L\alpha_k^2$ ,
  - 6:
  - 7:     or equivalently  $\alpha_k = \frac{1 + \mu_2 A_{k-1}}{2L} + \sqrt{\frac{(1 + \mu_2 A_{k-1})^2}{4L^2} + \frac{A_{k-1}(1 + \mu_2 A_{k-1})}{L}}$ ,
  - 8:      $A_k = A_{k-1} + \alpha_k$ ,
  - 9:      $y_k = \frac{A_{k-1}x_{k-1} + \alpha_k u_{k-1}}{A_k}$ ,
  - 10:
  - 11:      $\phi_k(x) = \alpha_k \langle \tilde{\nabla}f(y_k), x - y_k \rangle + \frac{1 + \mu_2 A_{k-1}}{2} \|u_{k-1} - x\|_2^2 + \frac{\mu_2 \alpha_k}{2} \|y_k - x\|_2^2$ ,
  - 12:
  - 13:      $u_k = \operatorname{argmin}_{u \in Q} \phi_k(u)$ ,
  - 14:      $x_k = \frac{A_{k-1}x_{k-1} + \alpha_k u_k}{A_k}$ .
  - 15: **end for**
  - 16: **Output:**  $x_N$ .
- 

Since  $Q = \mathbb{R}^n$ , the main step of the algorithm can be simplified to

$$u_{k+1} = \frac{1 + \mu_2 A_k}{1 + \mu_2 A_{k+1}} u_k + \frac{\mu_2 \alpha_{k+1}}{1 + \mu_2 A_{k+1}} y_{k+1} - \frac{\alpha_{k+1}}{1 + \mu_2 A_{k+1}} \tilde{\nabla}f(y_{k+1}).$$

Combining Definition 3.3 of [50] with Claims 4.1, 4.3 and particular choice  $V[y](x) = \frac{1}{2} \|x - y\|_2^2$ , we have that  $\delta$  in Definition 3.3 of [50] can be set to  $\delta = \frac{3}{2} \frac{\delta^2}{\mu_2} \geq \frac{\delta^2}{2L_f} + \frac{\delta^2}{\mu} = \delta_2 + \delta_3$ , where we used that  $\mu \leq L_f$  and that  $\mu_2 = \mu/2$ . Further,  $L$  in Definition 3.3 of [50] can be set to  $L$  in our paper, and  $\mu$  in Definition 3.3 of [50] can be set to  $\mu_2 = \mu/2$  in our paper. Algorithm 2 is a particular case of Algorithm 2 in [50]. Since in this section, we are in the setting of relative inexactness (4), in each iteration of

this algorithm we have a different error  $\delta_k = \alpha \|\nabla f(y_k)\|_2$ , which gives the following expression for  $\delta_k$  in Algorithm 2 in [50]:  $\delta_k = \frac{3\alpha^2 \|\nabla f(y_k)\|_2^2}{2\mu_2}$ .

Applying Theorem 3.4 of [50], we obtain the following convergence rate for all  $N \geq 0$ :

$$\begin{aligned} f(x_N) - f(x^*) &\leq \frac{R^2}{A_N} + \sum_{k=1}^N \frac{3\alpha^2 A_k \|\nabla f(y_k)\|_2^2}{\mu_2 A_N} := \frac{\kappa}{A_N}, \\ \|u_N - x^*\|_2^2 &\leq \frac{1}{1 + A_N \mu_2} \left[ R^2 + \sum_{k=1}^N \frac{3\alpha^2 A_k \|\nabla f(y_k)\|_2^2}{\mu_2} \right] := \frac{\kappa}{1 + A_N \mu_2}. \end{aligned}$$

Since we assumed that  $Q = \mathbb{R}^n$ , we have that  $\nabla f(x^*) = 0$  and that, for all  $x \in Q$ ,  $f(x) - f(x^*) \leq \frac{L}{4} \|x - x^*\|_2^2$ , where we used (23) and our definition  $L = 2L_f$ . Then, using convergence rate for  $\{u_k\}$ , we obtain

$$f(u_k) - f(x^*) \leq \frac{L}{4} \|u_k - x^*\|_2^2 \leq \frac{L\kappa}{4(1 + A_k \mu_2)}.$$

Using the convexity of  $f$  and the definition of the sequence  $\{y_k\}$  we get:

$$\begin{aligned} f(y_{N+1}) - f(x^*) &\leq \frac{\alpha_{N+1}}{A_{N+1}} [f(u_N) - f(x^*)] + \frac{A_N}{A_{N+1}} [f(x_N) - f(x^*)] \\ &\leq \frac{\alpha_{N+1}}{A_{N+1}} \cdot \frac{L\kappa}{4(1 + A_k \mu_2)} + \frac{A_N}{A_{N+1}} \cdot \frac{\kappa}{A_N}. \end{aligned}$$

Our next goal is to estimate  $\frac{\alpha_{N+1}}{A_{N+1}} \cdot \frac{L}{4(1 + A_k \mu_2)}$  from above. Using the inequalities  $\frac{A_k}{1 + \mu_2 A_k} \leq \frac{1}{\mu_2}$  and  $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}$ , and the definition of the sequence  $\{\alpha_k\}$ :

$$\alpha_k = \frac{1 + \mu_2 A_{k-1}}{2L} + \sqrt{\frac{(1 + \mu_2 A_{k-1})^2}{4L^2} + \frac{A_{k-1}(1 + \mu_2 A_{k-1})}{L}},$$

we have

$$\begin{aligned} &\frac{\alpha_{N+1}}{A_{N+1}} \cdot \frac{L}{4(1 + A_k \mu_2)} \\ &= \frac{L}{4A_{N+1}} \frac{1}{1 + \mu_2 A_N} \left( \frac{1 + \mu_2 A_N}{2L} + \sqrt{\frac{(1 + \mu_2 A_N)^2}{4L^2} + \frac{A_N(1 + \mu_2 A_N)}{L}} \right) \\ &\leq \frac{L}{4} \frac{1}{A_{N+1}} \left( \frac{1}{2L} + \frac{1}{2L} + \sqrt{\frac{1}{L\mu_2}} \right) \leq \frac{1}{4A_{N+1}} \sqrt{\frac{L}{\mu_2}}. \end{aligned}$$

This gives us the following estimate

$$f(y_{N+1}) - f^* \leq \frac{\kappa}{4A_{N+1}} \sqrt{\frac{L}{\mu_2}} + \frac{\kappa}{A_{N+1}} \leq \frac{1}{4} \sqrt{\frac{L}{\mu_2}} \left( \frac{5R^2}{A_{N+1}} + \sum_{k=1}^N \frac{15\alpha^2 A_k \|\nabla f(y_k)\|_2^2}{\mu_2 A_{N+1}} \right),$$

where we used that  $L/\mu_2 \geq 1$  and the definition of  $\kappa$ .

Since  $f$  is  $L_f$ -smooth,  $L = 2L_f$  and  $\nabla f(x^*) = 0$ , we obtain for any  $x \in Q$  that

$$\|\nabla f(x)\|_2^2 \leq L(f(x) - f(x^*)).$$

Whence, using the previous bound,

$$f(y_{N+1}) - f(x^*) \leq \frac{1}{4} \sqrt{\frac{L}{\mu_2}} \left( \frac{5R^2}{A_{N+1}} + \sum_{k=1}^N \frac{15\alpha^2 A_k L(f(y_k) - f^*)}{\mu_2 A_{N+1}} \right).$$

Introducing the following notations  $\lambda = \frac{5R^2}{4} \sqrt{\frac{L}{\mu_2}}$ ,  $\theta = \frac{15\alpha^2}{4} \sqrt{\frac{L^3}{\mu_2^3}}$ ,  $\Delta_k = f(y_k) - f(x^*)$ , we obtain the following recurrence

$$\Delta_N \leq \frac{\lambda}{A_N} + \theta \sum_{k=0}^{N-1} \frac{A_k}{A_N} \Delta_k,$$

where we add the term corresponding to  $k = 0$  to the sum to simplify the proof that will follow. Analyzing this recurrence, we obtain.

**Claim 5.1.** For all  $k \geq 1$  it holds that

$$\Delta_k \leq \frac{(1 + \theta)^{k-1}}{A_k} \lambda + \theta \frac{A_0 (1 + \theta)^{k-1}}{A_k} \Delta_0.$$

**Proof.** The induction basis  $k = 1$  is obvious. Induction step:

$$\begin{aligned} \Delta_k &\leq \frac{\lambda}{A_k} + \theta \sum_{j=0}^{k-1} \frac{A_j}{A_k} \Delta_j \\ &\leq \frac{\lambda}{A_k} + \theta \sum_{j=1}^{k-1} \frac{A_j}{A_k} \Delta_j + \theta \frac{A_0}{A_k} \Delta_0 \\ &\leq \frac{\lambda}{A_k} + \theta \sum_{j=1}^{k-1} \left( \frac{A_j (1 + \theta)^{j-1}}{A_k} \lambda + \theta \frac{A_0 (1 + \theta)^{j-1}}{A_k} \Delta_0 \right) + \frac{A_0}{A_k} \Delta_0 \\ &\leq \frac{\lambda}{A_k} + \theta \sum_{j=0}^{k-2} \left( \frac{\lambda (1 + \theta)^j}{A_k} + \theta \frac{A_0 (1 + \theta)^j}{A_k} \Delta_0 \right) + \frac{A_0}{A_k} \Delta_0 \\ &= \frac{1}{A_k} \left( \lambda + \lambda \left[ (1 + \theta)^{k-1} - 1 \right] + \theta A_0 \Delta_0 \left[ (1 + \theta)^{k-1} - 1 \right] + A_0 \Delta_0 \right) \\ &= \frac{(1 + \theta)^{k-1}}{A_k} \lambda + \theta \frac{A_0 (1 + \theta)^{k-1}}{A_k} \Delta_0. \end{aligned}$$

□

This gives us the following result

$$f(y_k) - f(x^*) \leq \frac{\lambda (1 + \theta)^k}{A_k} + \theta \frac{A_0 (1 + \theta)^k}{A_k} (f(y_0) - f(x^*)).$$

By the definition of  $\theta = \frac{15\alpha^2}{4}\sqrt{\frac{L^3}{\mu_2^3}}$ , we obtain, that, if we choose  $\alpha$  as

$$\alpha \leq \frac{1}{7} \frac{\mu_2}{L} = \Theta\left(\frac{\mu}{L}\right), \quad (37)$$

then

$$\begin{aligned} 1 + \sqrt{\frac{\mu_2}{L}} \left( \frac{1}{2} + \frac{15}{196} + \frac{15}{392} \right) &\leq 1 + \sqrt{\frac{\mu_2}{L}} \sqrt{\frac{1}{2}} \\ \Leftrightarrow 1 + \theta + \frac{1}{2} \sqrt{\frac{\mu_2}{L}} + \frac{1}{2} \sqrt{\frac{\mu_2}{L}} \theta &\leq 1 + \sqrt{\frac{\mu_2}{2L}} \\ \Leftrightarrow \frac{1 + \theta}{1 + \sqrt{\frac{\mu_2}{2L}}} &\leq \frac{1}{1 + \frac{1}{2} \sqrt{\frac{\mu_2}{L}}}. \end{aligned}$$

Combining this with Claim 4.4 and Corollary 4.3, we obtain that

$$\frac{(1 + \theta)^k}{A_k} \leq \left( \frac{1 + \theta}{1 + \sqrt{\frac{\mu_2}{2L}}} \right)^k \frac{1}{A_0} \leq \left( \frac{1}{1 + \frac{1}{2} \sqrt{\frac{\mu_2}{L}}} \right)^k \frac{1}{A_0} \leq L \exp\left(-\frac{k}{4} \sqrt{\frac{\mu_2}{L}}\right). \quad (38)$$

As a result, we get the following theorem.

**Theorem 5.1.** *Assume that the objective  $f$  is  $L_f$ -smooth and strongly convex with  $\mu \neq 0$ , that the inexactness in the gradient is described by (4), and that  $Q = \mathbb{R}^n$ . Also assume that  $\alpha$  is chosen according to (37). Then, for all  $k \geq 1$ , the sequence  $\{y_k\}$  generated by Algorithm 2 satisfies*

$$f(y_k) - f(x^*) \leq \left( \frac{5LR^2}{4} + \frac{15}{196} \sqrt{\frac{2L}{\mu}} [f(y_0) - f(x^*)] \right) \exp\left(-\frac{k}{4} \sqrt{\frac{\mu}{2L}}\right).$$

## 6. Extensions

In this section, we extend the analysis of Algorithm 1 with absolute noise to two settings. The first extension is an extension to stochastic optimization setting where the error in the gradient has stochastic nature. The second one is the extension to structured nonsmooth setting of composite minimization, where the objective is given as a sum of smooth part with inexact gradient and a simple convex function. In both cases, the analysis mainly follows the lines of Section 4. Thus, we underline the differences and skip in the proofs some steps that are similar to the analysis in that section.

### 6.1. Random additive noise in the gradient

In this subsection, we extend the analysis of Algorithm 1 for the setting of random absolute noise in the gradient. We assume that an algorithm can use the stochastic gradient  $\tilde{\nabla}f(x, \xi)$ , which is assumed to have bounded variance for all, possibly random,

$x \in Q$ :

$$\mathbb{E}_\xi \left[ \|\tilde{\nabla} f(x, \xi) - \nabla f(x)\|_2^2 \mid x \right] \leq \delta^2. \quad (39)$$

Similarly to Section 4, we assume  $L_f$ -smoothness and  $\mu$ -strong convexity of  $f$ , i.e., that (23),(24) hold. As before, we set  $L = 2L_f$  and

$$\delta_1 = \delta, \quad \delta_2 = \frac{\delta^2}{L}, \quad \delta_3 = \frac{\delta^2}{\mu},$$

where the latter quantity is defined whenever  $\mu > 0$ .

One of the main motivations for such stochastic problems is machine learning. For example, Empirical Risk Minimization problem with the finite-sum structure of the objective

$$f(x) = \frac{1}{M} \sum_{i=1}^M f_i(x)$$

can be considered as a stochastic optimization problem with stochastic gradient

$$\tilde{\nabla} f(x, \xi) = \frac{1}{m} \sum_{i \in \xi} \nabla f_i(x); \quad \xi \subset \{1 \dots M\}, \quad |\xi| = m, \quad m < M,$$

where  $\xi$  is a random subset of  $\{1 \dots M\}$ . It should be noted that the error  $\delta^2$  can be reduced by the use of mini-batches. Namely increasing the size of  $\xi$  from 1 to  $m$  decreases the variance from  $\delta^2$  to  $\frac{\delta^2}{m}$ .

The first step of the analysis is to obtain the counterparts of Claims 4.1, 4.2, and 4.3 in the stochastic setting.

**Claim 6.1.** Assume that  $x, y$  are random vectors. Then,

$$\mathbb{E}[f(y)] \leq \mathbb{E} \left( f(x) + \langle \tilde{\nabla} f(x, \xi), y - x \rangle + \frac{L}{2} \|x - y\|_2^2 \right) + \delta_2.$$

**Proof.** Using the  $L_f$ -smoothness, we obtain

$$\begin{aligned} \mathbb{E} \left[ f(y) \mid x \right] &\leq \mathbb{E} \left[ f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f}{2} \|y - x\|_2^2 \mid x \right] \\ &\stackrel{(39)}{\leq} \mathbb{E} \left[ f(x) + \langle \tilde{\nabla} f(x, \xi), y - x \rangle + \frac{L}{2} \|y - x\|_2^2 \mid x \right] + \delta_2, \end{aligned}$$

where  $L = 2L_f$ . Taking the full expectation of both sides, we get the required.  $\square$

Using the same steps as in the proof of Claim 6.1, we get the counterparts of Claims 4.2, 4.3.

**Claim 6.2.** Assume that  $x, y$  are random vectors. Then,

$$\mathbb{E} \left[ f(x) + \langle \tilde{\nabla} f(x, \xi), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2 - \delta_1 \|x - y\|_2 \mid x \right] \leq \mathbb{E} \left[ f(y) \mid x \right].$$

**Claim 6.3.** Assume that  $x, y$  are random vectors and that  $\mu \neq 0$ . Then,

$$\mathbb{E} \left[ f(x) + \langle \tilde{\nabla} f(x, \xi), y - x \rangle + \frac{\mu}{4} \|y - x\|_2^2 \mid x \right] - \delta_3 \leq \mathbb{E} \left[ f(y) \mid x \right].$$

The following sequence is the counterpart of the sequence  $\{\tilde{R}_k\}$ :

$$\tilde{B}_k = \max_{0 \leq j \leq k} \{ \mathbb{E} \|z_j - x^*\|_2, \mathbb{E} \|x_j - x^*\|_2, \mathbb{E} \|\tilde{x}_j - x^*\|_2 \}. \quad (40)$$

Using the above, we obtain the following counterparts of Propositions 4.4 and 4.5 under the assumptions of this subsection.

**Proposition 6.1.** *The sequences generated by Algorithm 1 satisfy for all  $k \geq 0$  the inequality:*

$$A_k \mathbb{E} [f(x_k)] \leq \mathbb{E} [\psi_k(z_k)] + \delta_2 \sum_{j=0}^k A_j + \delta_1 \sum_{j=1}^k \alpha_j \mathbb{E} [\|\tilde{x}_j - z_{j-1}\|_2].$$

**Proposition 6.2.** *If  $\mu \neq 0$ , the sequences generated by Algorithm 1 satisfy for all  $k \geq 0$  the inequality:*

$$A_k \mathbb{E} [f(x_k)] \leq \mathbb{E} [\psi_k(z_k)] + \delta_2 \sum_{j=0}^k A_j + \delta_3 \sum_{j=0}^{k-1} A_j.$$

The proofs of these propositions repeat the same induction steps as in the proofs of Propositions 4.4, 4.5, but using the new Claims 6.1, 6.2, and 6.3. Using the last two propositions, we finally obtain the following counterpart of convergence Theorem 4.7 for the stochastic setting.

**Theorem 6.3** (Convergence rate of stochastic STM).

Let  $\|\tilde{x}_0 - x^*\|_2 \leq R$  for some  $R$ , function  $f$  be  $L_f$ -smooth and strongly convex with parameter  $\mu \geq 0$ . Let the stochastic gradient  $\tilde{\nabla} f(x, \xi)$  satisfy

$$\mathbb{E}_\xi \left[ \|\tilde{\nabla} f(x, \xi) - \nabla f(x)\|_2^2 \mid x \right] \leq \delta^2. \quad (41)$$

Then, if  $\mu \neq 0$ , the sequence  $\{x_N\}$  generated by Algorithm 1 satisfy for all  $N \geq 0$  the inequalities

$$\begin{aligned} \mathbb{E} f(x_N) - f(x^*) &\leq LR^2 \exp\left(-\frac{1}{2} \sqrt{\frac{\mu_1}{L}} N\right) + \left(1 + \sqrt{\frac{L}{\mu_1}}\right) \delta_2 + 3\tilde{B}_N \delta_1, \\ \mathbb{E} f(x_N) - f(x^*) &\leq LR^2 \exp\left(-\frac{1}{2} \sqrt{\frac{\mu_2}{L}} N\right) + \left(1 + \sqrt{\frac{L}{\mu_2}}\right) \delta_2 + \left(1 + \sqrt{\frac{L}{\mu_2}}\right) \delta_3. \end{aligned}$$

If  $\mu = 0$ , the sequences  $\{x_k\}$ ,  $\{\tilde{x}_k\}$ ,  $\{z_k\}$  generated by Algorithm 1 satisfy for all  $N \geq 0$

the inequality

$$\mathbb{E}f(x_N) - f(x^*) \leq \frac{4LR^2}{N^2} + 3\tilde{B}_N\delta_1 + N\delta_2.$$

Here the sequence  $\{\tilde{B}_k\}$  is defined in (40).

As we see, Algorithm 1 has the same convergence rate in the stochastic setting as in the deterministic setting. The proof of the above theorem repeats the same steps as the proof of Theorem 4.7. Thus, we omit the proof.

**Remark 6.1.** Usually, in the context of stochastic optimization, the analysis of algorithms relies also on the assumption of unbiased stochastic gradient:

$$\mathbb{E} \left[ \tilde{\nabla} f(x, \xi) \mid x \right] = \nabla f(x).$$

Our analysis does not require this assumption.

## 6.2. Nonsmooth objective

In this subsection, we consider the problem of structured nonsmooth optimization, usually referred to as composite minimization,

$$\min_{x \in Q} \{f(x) = \mathcal{L}(x) + r(x)\}. \quad (42)$$

We assume that the function  $\mathcal{L}$  is  $L_f$ -smooth and  $\mu$ -strongly-convex (see (22), (24)), the function  $r(x)$  is convex and relatively simple. We further assume that inexact gradient  $\tilde{\nabla}\mathcal{L}(x)$  with absolute noise (cf. (3)) is available for  $\mathcal{L}$ .

This setting is motivated, in particular, by machine learning problems, for example, logistic regression loss minimization problem with the  $l_1$  regularization and dataset  $\{(X_i, y_i)\}_{i=1}^K$ , where  $y_i \in \{0, 1\}$  for  $i \in \{1, K\}$ . For this problem, we have

$$\begin{aligned} \mathcal{L}(x) &= \sum_{i=1}^K y_i \ln p_i(x) + (1 - y_i) \ln(1 - p_i(x)), \\ p_i(x) &= \sigma(\langle x, X_i \rangle), \sigma(z) = \frac{1}{1 + \exp(-z)}, \\ r(x) &= \lambda_1 \|x\|_1. \end{aligned}$$

In the setting of composite minimization, Algorithm 1 requires only one change in the definition of the function sequence  $\{\psi_k(\cdot)\}$  as follows:

$$\begin{aligned} \psi_0(x) &= \frac{1}{2} \|x - \tilde{x}_0\|_2^2 + \alpha_0 \left( \mathcal{L}(\tilde{x}_0) + \langle \tilde{\nabla}\mathcal{L}(\tilde{x}_0), x - \tilde{x}_0 \rangle + \frac{\mu_\tau}{2} \|x - \tilde{x}_0\|_2^2 + r(x) \right), \\ \psi_k(x) &= \psi_{k-1}(x) + \alpha_k \left( \mathcal{L}(\tilde{x}_k) + \langle \tilde{\nabla}\mathcal{L}(\tilde{x}_k), x - \tilde{x}_k \rangle + \frac{\mu_\tau}{2} \|x - \tilde{x}_k\|_2^2 + r(x) \right). \end{aligned} \quad (43)$$

For such modified algorithm, in the concept of absolute noise (3), the convergence result remains the same. However, some intermediate statements, such as Lemma 4.2,

require a different analysis. Therefore, we make a different analysis to obtain an estimate in the spirit of Proposition 4.4.

**Lemma 6.4** (auxiliary statement for  $\psi_k$ 's). *Under the assumptions of this subsection, for the modified sequence  $\{\psi_k(\cdot)\}$ , we have*

$$\begin{aligned} \psi_{k+1}(z_{k+1}) \geq & \psi_k(z_k) + \frac{1 + \mu_\tau A_k}{2} \|z_k - z_{k+1}\|_2^2 \\ & + \alpha_k \left( \mathcal{L}(\tilde{x}_k) + \langle \tilde{\nabla} \mathcal{L}(\tilde{x}_k), x - \tilde{x}_k \rangle + \frac{\mu_\tau}{2} \|x - \tilde{x}_k\|_2^2 + r(x) \right). \end{aligned}$$

**Proof.** The function  $\psi_k$  defined in (43) is  $\frac{1 + \mu_\tau A_k}{2}$ -strongly-convex. Thus, since  $z_k$  is its minimizer, we have

$$\psi_k(z_{k+1}) \geq \psi_k(z_k) + \frac{1 + \mu_\tau A_k}{2} \|z_k - z_{k+1}\|_2^2.$$

Using the recurrent definition of  $\psi_{k+1}$  we obtain the required by induction.  $\square$

Using Lemma 6.4 instead of Lemma 4.2 and convexity of the function  $r(x)$ , we can obtain a result similar to Proposition 4.4.

**Proposition 6.5.** *The sequences  $\{x_k\}$ ,  $\{\tilde{x}_k\}$ ,  $\{z_k\}$  generated by Algorithm 1 modified for structured nonsmooth optimization satisfy for all  $k \geq 0$  the inequality*

$$A_k f(x_k) \leq \psi_k(z_k) + \delta_2 \sum_{j=0}^k A_j + \delta_1 \sum_{j=1}^k \alpha_j \|\tilde{x}_j - z_{j-1}\|_2.$$

**Proof.** The induction basis  $k = 0$  is obvious and repeats the proof of Proposition 4.4 since

$$f(x_0) \leq L\psi_0(z_0) + \delta_2.$$

Let us consider iteration  $k > 0$ . Since  $r(x)$  is convex, by the definition of  $x_k$ , we get:

$$A_k r(x_k) \leq A_{k-1} r(x_{k-1}) + \alpha_k r(z_k).$$

By this inequality, Claim 4.1 applied to  $\mathcal{L}(x)$ , the definition of the sequences  $\{x_k\}$ ,  $\{\tilde{x}_k\}$ , we have

$$\begin{aligned} A_k f(x_k) & \leq A_k \left( \mathcal{L}(\tilde{x}_k) + \langle \tilde{\nabla} \mathcal{L}(\tilde{x}_k), x_k - \tilde{x}_k \rangle + \frac{L}{2} \|x_k - \tilde{x}_k\|_2^2 + \delta_2 + r(x_k) \right) \\ & \leq A_{k-1} \left( \mathcal{L}(\tilde{x}_k) + \langle \tilde{\nabla} \mathcal{L}(\tilde{x}_k), x_{k-1} - \tilde{x}_k \rangle + r(x_{k-1}) \right) \\ & \quad + \alpha_k \left( \mathcal{L}(\tilde{x}_k) + \langle \tilde{\nabla} \mathcal{L}(\tilde{x}_k), z_k - \tilde{x}_k \rangle + r(z_k) \right) + A_k \delta_2 + \frac{L\alpha_k^2}{A_k} \|z_k - z_{k-1}\|_2^2 \\ & \leq A_{k-1} f(x_{k-1}) + \alpha_k \left( \mathcal{L}(\tilde{x}_k) + \langle \tilde{\nabla} \mathcal{L}(\tilde{x}_k), z_k - \tilde{x}_k \rangle + r(z_k) \right) \\ & \quad + \frac{1 + \mu_\tau A_{k-1}}{2} \|z_k - z_{k-1}\|_2^2 + A_k \delta_2 + A_{k-1} \delta_1 \|x_{k-1} - \tilde{x}_k\|_2, \end{aligned}$$

where in the last inequality we used the equation in Step 6 of Algorithm 1 and Claim 4.2 applied to  $\mathcal{L}$ . By the induction hypothesis and since  $\frac{\mu_\tau}{2}\|z_k - \tilde{x}_k\|_2^2 \geq 0$ , we further obtain

$$\begin{aligned} A_k f(x_k) - A_{k-1} \delta_1 \|x_{k-1} - \tilde{x}_k\|_2 &\leq \psi_{k-1}(z_{k-1}) + \delta_2 \sum_{j=0}^{k-1} A_j + \delta_1 \sum_{j=1}^{k-1} \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 \\ &\quad + \alpha_k \left( \mathcal{L}(\tilde{x}_k) + \langle \tilde{\nabla} \mathcal{L}(\tilde{x}_k), z_k - \tilde{x}_k \rangle + \frac{\mu_\tau}{2} \|z_k - \tilde{x}_k\|_2^2 + r(z_k) \right) \\ &\quad + \frac{1 + \mu_\tau A_{k-1}}{2} \|z_k - z_{k-1}\|_2^2 + A_k \delta_2. \end{aligned}$$

Using Lemma 6.4, we can finish the proof in a similar way as in the proof of Proposition 4.4.  $\square$

We finally obtain the following counterpart of Theorem 4.7 for composite minimization problems.

**Theorem 6.6.** *Let the modified Algorithm 1 be applied to composite problem (42), where the function  $\mathcal{L}(x)$  is  $L_f$ -smooth and  $\mu$ -strongly-convex and the function  $r(x)$  is convex. If  $\mu \neq 0$ , the sequences  $\{x_k\}$ ,  $\{\tilde{x}_k\}$ ,  $\{z_k\}$  generated by the modified Algorithm 1 satisfy for all  $N \geq 0$  the inequalities*

$$\begin{aligned} f(x_N) - f(x^*) &\leq LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu_1}{L}}N\right) + \left(1 + \sqrt{\frac{L}{\mu_1}}\right) \delta_2 + 3\tilde{R}_N \delta_1, \\ f(x_N) - f(x^*) &\leq LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu_2}{L}}N\right) + \left(1 + \sqrt{\frac{L}{\mu_2}}\right) \delta_2 + \left(1 + \sqrt{\frac{L}{\mu_2}}\right) \delta_3. \end{aligned}$$

If  $\mu = 0$ , the sequences  $\{x_k\}$ ,  $\{\tilde{x}_k\}$ ,  $\{z_k\}$  generated by the modified Algorithm 1 satisfy for all  $N \geq 0$  the inequality

$$f(x_N) - f(x^*) \leq \frac{4LR^2}{N^2} + 3\tilde{R}_N \delta_1 + N\delta_2,$$

where the sequence  $\{\tilde{R}_k\}$  is defined in (29).

As we see, for composite problems, modulo a small modification of the algorithm, the main result is the same as in the smooth case.

## 7. Conclusions and observations

In this section, we give a number of remarks in order to discuss the obtained results. In particular, the convergence rate results obtained so far explicitly include the oracle inexactness, and we can look at these results from a little bit different angle of controlling the inexactness. In particular, if the oracle error can be controlled, we can estimate how small should be the oracle error if our goal is to obtain an  $\varepsilon$ -solution to the problem. Such bound also give an estimate for the largest tolerable error not preventing the algorithm from obtaining an  $\varepsilon$ -solution.

**Remark 7.1.** In Sections 6.1, 6.2, we considered the extensions of Algorithm 1 with absolute noise to the settings of stochastic optimization and structured nonsmooth optimization. We strongly believe that it is possible to combine these two extensions into one since the analysis in both cases follows the same lines as the analysis in Section 4. We believe that the same can be also done with the analysis of Algorithm 2 under the relative noise in the gradient (see stochastic version of this condition in [55]). We leave these developments for the future work.

**Remark 7.2.** The results of Theorem 4.7 and Proposition 4.6 are obtained for possibly unbounded feasible set  $Q$ . If this set is compact, we can set  $R = \text{diam}(Q)$ , i.e., the diameter of the set  $Q$ . This simplifies the results and derivations since, in this case, by the construction of Algorithm 1,  $\tilde{R}_k \leq R$  for all  $k \geq 0$ .

**Remark 7.3.** When considering the absolute noise, in Section 4.1, we had two possibilities for dealing with “inexact strong convexity”: according to Claim 4.2 when  $\mu \geq 0$  and according to Claim 4.3 when  $\mu > 0$ . This resulted in two different bounds in Theorem 4.7 in the setting when  $\mu > 0$ . Recalling that

$$\delta_1 = \delta, \quad \delta_2 = \frac{\delta^2}{L}, \quad \delta_3 = \frac{\delta^2}{\mu},$$

and comparing the two bounds in Theorem 4.7, we see that if

$$\delta < \frac{3\tilde{R}}{\frac{1 + \sqrt{\frac{L}{\mu}}}{\mu} + \frac{\sqrt{\frac{L}{\mu}}(\sqrt{2}-1)}{L}},$$

then the model corresponding to  $\tau = 2$ , that is described in Claim 4.3, leads to a smaller term in the convergence rate bound due to the error accumulation than the model corresponding to  $\tau = 1$ , that is described in Claim 4.2

The above results are valid for uncontrolled and unknown values of the error  $\delta$  in the model of absolute noise. At the same time, in some cases, it may happen that the error  $\delta$  can be controlled and made as small as one desires. For example, in the setting of Section 3.2, the gradient can be approximated using finite-difference solution of primal and adjoint systems of equations, and  $\delta$  can be decreased by decreasing the discretization step. In the setting of Section 6.1, the error  $\delta$  can be made smaller by the means of using mini-batches of stochastic gradients. Thus, a natural question is how small should one choose the accuracy  $\delta$  if the goal is to find an  $\varepsilon$ -approximate solution, i.e., guarantee  $f(x_k) - f(x^*) \leq \varepsilon$ ? A similar question could be as follows: given a target accuracy  $\varepsilon$ , how large is the error  $\delta$  that can be tolerated by an algorithm still guaranteeing the target accuracy  $\varepsilon$ ? This, in particular, allows one to compare the robustness of different algorithms with respect to the noise. In the following series of remarks, we address these questions by deriving the relations between  $\delta$  and  $\varepsilon$ .

**Remark 7.4.** Let us consider the “inexact strong convexity” model corresponding to  $\tau = 2$ ,  $\mu > 0$ ,  $\mu_2 = \frac{\mu}{2}$ , and  $\delta_2 = \frac{\delta^2}{2L_f}$ ,  $L = 2L_f$ ,  $\delta_3 = \frac{\delta^2}{\mu}$  (see Claims 4.1, 4.3). In this case, we can write explicit expressions for the dependence of the error  $\delta$  and the iteration number  $N$  on the target accuracy  $\varepsilon$ . Substituting the above values into the

bound in Theorem 4.7, we obtain

$$\begin{aligned} f(x_N) - f(x^*) &\leq LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2L}}N\right) + \left(1 + \sqrt{\frac{L}{\mu_2}}\right) \delta_2 + \left(1 + \sqrt{\frac{L}{\mu_2}}\right) \delta_3 \\ &= 2L_f R^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{4L_f}}N\right) + \left(1 + \sqrt{\frac{4L_f}{\mu}}\right) \left(\frac{2L_f + \mu}{2\mu L_f}\right) \delta^2. \end{aligned}$$

Thus, choosing

$$\begin{aligned} \delta &\leq \sqrt{\varepsilon} \sqrt{\frac{\mu L_f}{\mu + 2L_f}} \left(1 + \sqrt{\frac{4L_f}{\mu}}\right)^{-\frac{1}{2}} \leq \left(\frac{L_f \varepsilon}{1 + 2\frac{L_f}{\mu} + 2\sqrt{\frac{L_f}{\mu}} + 4\frac{L_f \sqrt{L_f}}{\mu \sqrt{\mu}}}\right)^{\frac{1}{2}} = O\left(\sqrt{\mu \varepsilon} \left(\frac{\mu}{L_f}\right)^{\frac{1}{4}}\right); \\ N &\geq 2\sqrt{\frac{4L_f}{\mu}} (\ln 4L_f R^2 + \ln \varepsilon^{-1}) = O\left(\sqrt{\frac{L_f}{\mu}} \ln \frac{L_f R^2}{\varepsilon}\right), \end{aligned}$$

we guarantee that

$$f(x_N) - f(x^*) \leq \varepsilon.$$

**Remark 7.5.** Let us consider the setting of Remark 4.2, where we made a reduction of the convex case  $\mu = 0$  to the strongly convex case by introducing a quadratic regularization with regularization parameter  $\mu$ . Recall that this led to the bound

$$f(x_N) - f(x^*) \leq LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2(L+2)}}N\right) + \left(1 + \sqrt{\frac{2L+4}{\mu}}\right) \left(\frac{1}{L} + \frac{1}{\mu}\right) \delta^2 + \frac{\mu}{2} R^2,$$

where  $R$  is such that  $\|\tilde{x}_0 - x^*\|_2 \leq R$ . We choose the regularization parameter  $\mu$ , the error level  $\delta$ , and the number of iterations  $N$  such that each of the three terms in this bound are smaller than  $\frac{\varepsilon}{3}$ . Then, choosing

$$\mu = \frac{2}{3} \frac{\varepsilon}{R^2},$$

$$\delta \leq \left(\frac{2}{243}\right)^{\frac{1}{4}} \frac{1}{\sqrt{1 + \sqrt{2L+4}}} R^{-\frac{3}{2}} \varepsilon^{\frac{5}{4}} = O\left(L^{-\frac{1}{4}} R^{-\frac{3}{2}} \varepsilon^{\frac{5}{4}}\right),$$

$$N \geq \sqrt{12L + 24} R \ln 2LR^2 + 2\sqrt{2L+4} \frac{1}{\sqrt{\varepsilon}} \ln \frac{1}{\varepsilon} = O\left(\sqrt{\frac{LR^2}{\varepsilon}} \ln \frac{LR^2}{\varepsilon}\right),$$

we guarantee that

$$f(x_N) - f(x^*) \leq \varepsilon.$$

**Remark 7.6.** Let us apply Theorem 4.8 for solving linear inverse problems. Let  $A \in \mathbb{R}^{n \times n}$  be such that  $\det(A) \neq 0$  and consider the following linear system for finding  $x \in \mathbb{R}^n$ :  $Ax = b$ . Solving this problem is equivalent to solving the convex optimization problem:

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} \|Ax - b\|_2^2 \right\}.$$

If we solve the latter problem with accuracy  $\varepsilon = \frac{\varepsilon_0^2}{2}$ , then we guarantee that  $\|Ax - b\|_2 \leq \varepsilon_0$ .

Let us assume that the solution  $x^*$  satisfies  $\|x^*\|_2 \leq R_*$  and that Algorithm 1 starts from the point 0. Then, we can take  $R = R_*$ . According to Theorem 4.8, given a target accuracy  $\zeta > 0$ , we have that Algorithm 1 stops after  $N_{\text{stop}}$  iterations such that  $N_{\text{stop}} \leq \sqrt{\frac{2LR_*^2}{\zeta}} + 1$ . Moreover, we have that

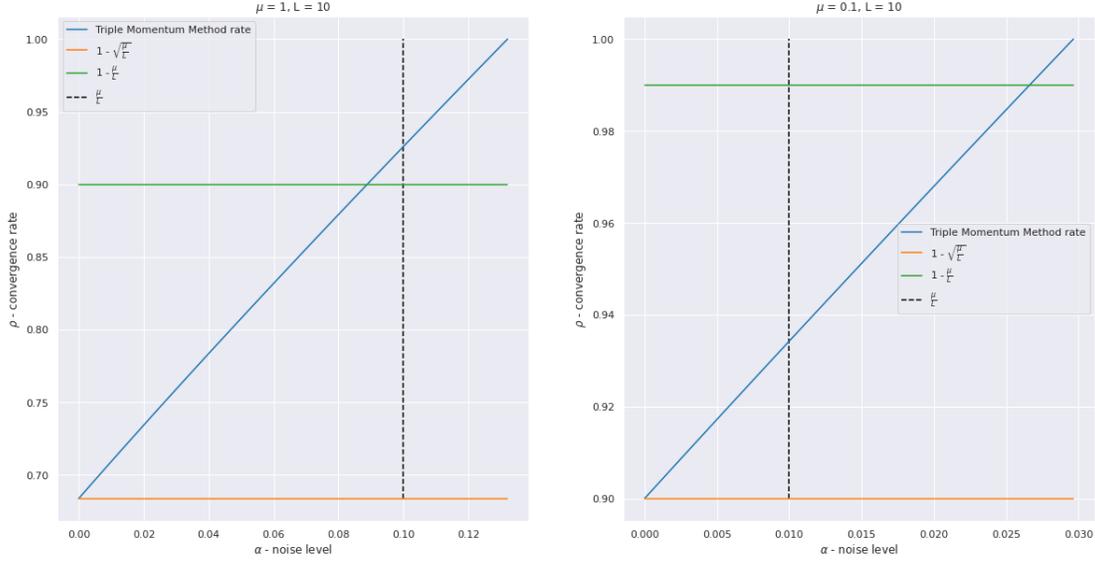
$$f(x_{N_{\text{stop}}}) - f(x^*) \leq \frac{\delta_2}{A_N} \sum_{j=0}^{N_{\text{stop}}} A_j + 3\delta_1 R_* + \zeta \leq N_{\text{stop}} \delta_2 + 3\delta_1 R_* + \zeta$$

since  $\{A_j\}$  is an increasing sequence.

Choosing  $\zeta \leq \frac{\varepsilon}{3}$  and  $\delta \leq \min \left\{ \left( \frac{L^{\frac{1}{4}}}{6\sqrt{3}R_*} \right) \varepsilon^{\frac{3}{4}}, \frac{\varepsilon}{9R_*} \right\}$ , we guarantee that  $f(x_{N_{\text{stop}}}) - f(x^*) \leq \varepsilon$  and, hence,  $\|Ax - b\|_2 \leq \varepsilon_0$ . Moreover, the number of iterations to guarantee such a solution is bounded as

$$N_{\varepsilon_0} = \frac{\sqrt{6LR_*^2}}{\varepsilon_0} + 1.$$

**Remark 7.7.** In the setting of relative noise in the gradient, Theorem 5.1 says that whenever  $\alpha \leq O\left(\frac{\mu}{L}\right)$ , STM converges linearly in the same way as accelerated gradient method in the exact setting, i.e., with the rate  $O\left(\left(1 - \sqrt{\frac{\mu}{L}}\right)^k\right)$  which is faster than the convergence rate  $O\left(\left(1 - \frac{\mu}{L}\right)^k\right)$  of gradient descent. Here  $k$  is the iteration counter. The paper [26] considers, in particular, accelerated method, called the Triple Momentum Method, in the presence of relative noise in the gradient. They show that when  $\alpha < \frac{\sqrt{\chi}+1}{4\chi-3\sqrt{\chi}+1} = O\left(\sqrt{\frac{\mu}{L}}\right)$ , where  $\chi = \frac{L}{\mu}$ , the Triple Momentum Method converges with a linear rate as well. At the same time, their convergence rate depends on the noise level  $\alpha$ , is no better than the accelerated rate, and is equal to it only in the case  $\alpha = 0$ . Figure 2 illustrates the situation for two different values of the condition number  $\chi$ . The black dashed line shows the threshold, below which STM with relative inexactness in the gradient has linear convergence rate similar to exact STM, and the latter rate is denoted by the orange line. Green line shows the convergence rate of the gradient method. Finally, the blue line shows the dependence of the convergence rate in [26] on the inexactness level  $\alpha$ . As we see, it can be even worse than that of the gradient method for large values of  $\alpha$ .



**Figure 2.** Comparison of the convergence rate of Triple Momentum Method and STM

As our experiments show, STM is more robust in the relative noise setting, that is, numerically estimating the dependence of the largest possible  $\alpha = \alpha^*$  for given problem parameters  $\mu, L$ , we get a larger upper bound. More detailed information can be found in Section 8. This leads us to the hypothesis that the condition  $\alpha \leq O\left(\frac{\mu}{L}\right)$  for inexact STM may be weakened.

## 8. Numerical experiments

In this section, we provide a series of numerical experiments to illustrate the practical performance of the considered algorithms under absolute and relative noise. The noise was generated as independent random uniform and unbiased.

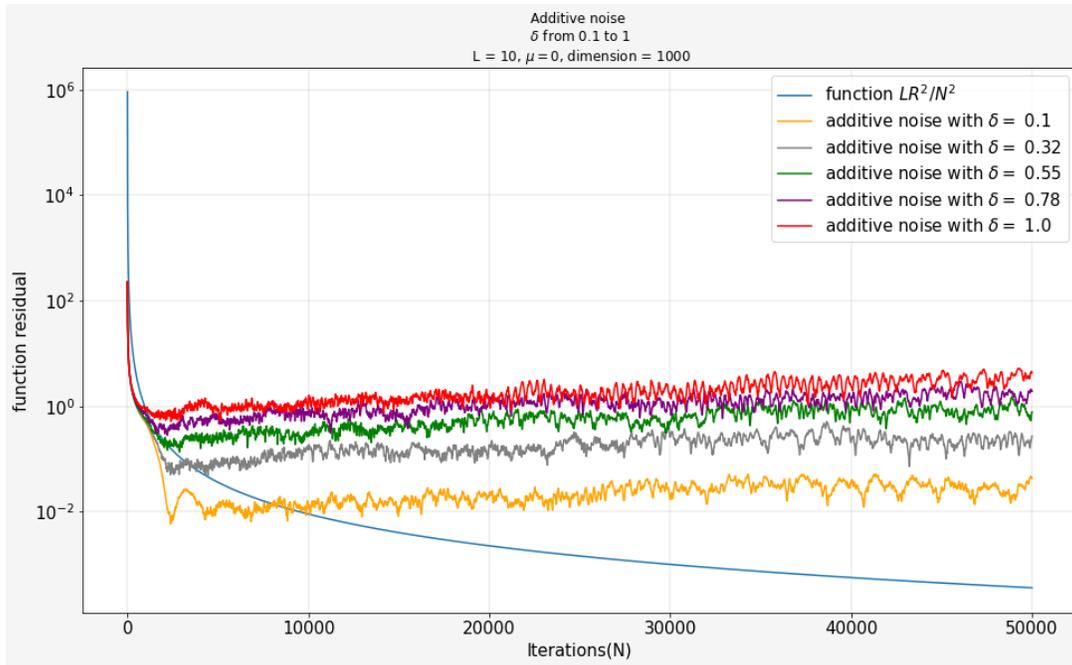
We start with the experiments in the setting of  $\mu = 0$  using the following objective function described in [40, p. 69] and known as the worst-case function for first-order methods:

$$f(x) = \frac{L}{8} \left( x_1^2 + \sum_{j=0}^{k-1} (x_j - x_{j+1})^2 + x_k^2 \right) - \frac{L}{4} x_1,$$

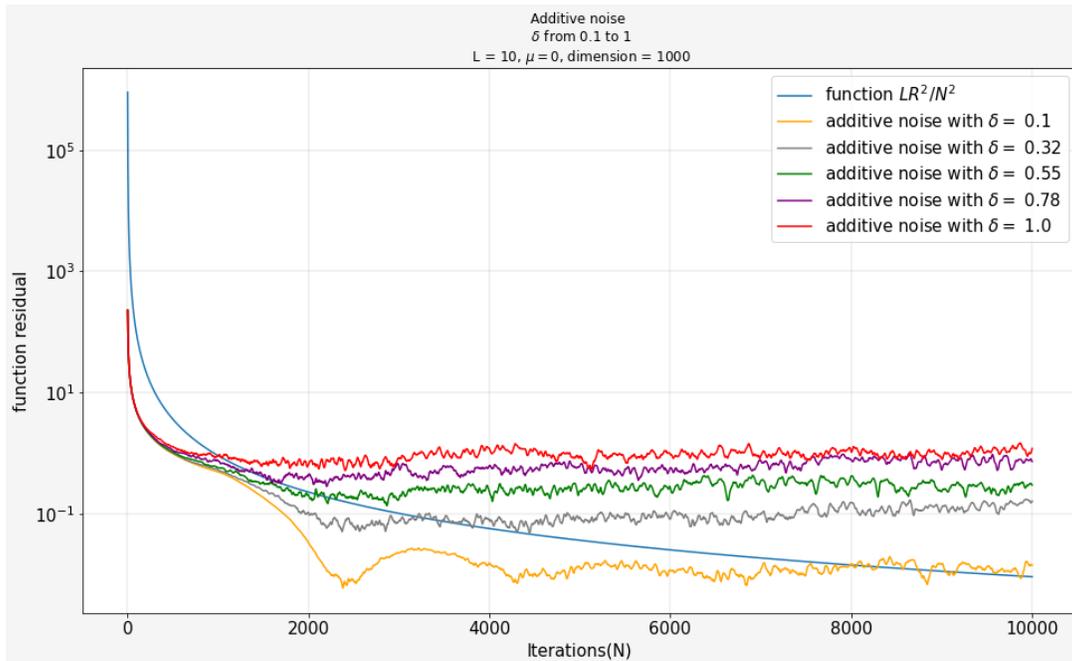
$$x^* = \left( 1 - \frac{1}{k+1}, \dots, 1 - \frac{k}{k+1}, 0, \dots, 0 \right)^T,$$

$$1 \leq k \leq \dim x.$$

The next two plots show the convergence of STM at the first 50 000 and 10 000 iterations, respectively, in the absolute noise setting with different values of  $\delta$ .



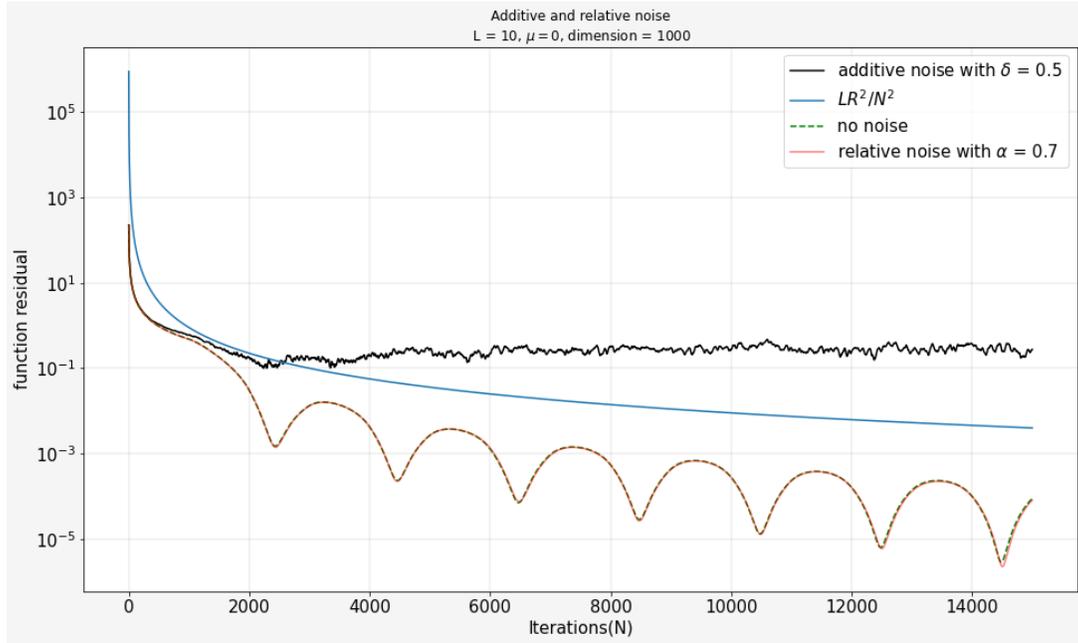
**Figure 3.** First test – the performance of STM for  $\mu = 0$  for the first 50 000 iterations.



**Figure 4.** First test – the performance of STM for  $\mu = 0$  for the first 10 000 iterations.

We can observe that, as predicted by Theorem 4.7, we see that the increasing third term in the convergence rate (31) at some point starts to overweight the first decreasing term.

We further compare the convergence in two different settings of the noise: absolute and relative.



**Figure 5.** Second test – the performance of STM for  $\mu = 0$  with relative and additive types of noise.

As was expected from the theory, for sufficiently small  $\alpha$ , the convergence of inexact method is very close to the convergence of the exact method. Since in this experiment the noise is stochastic, this effect can be possibly explained using the theoretical results obtained in [55]: under the strong growth condition (SGC)

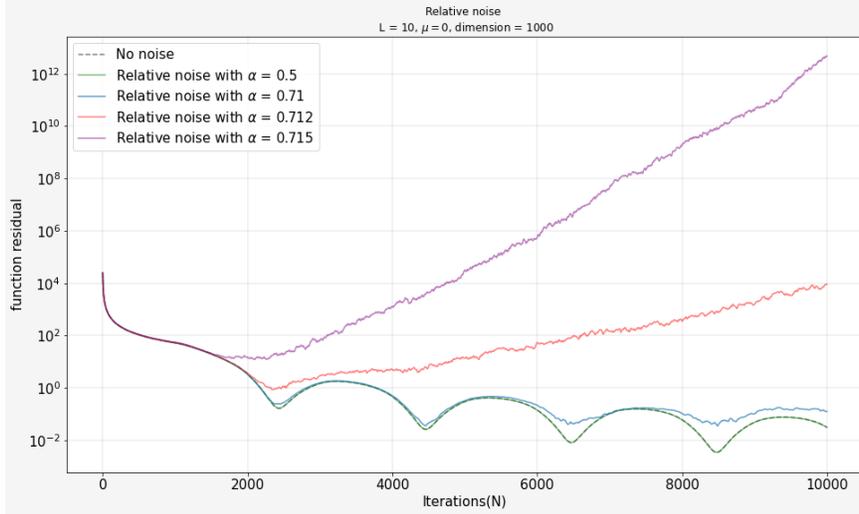
$$\mathbb{E}_{\xi} \|\tilde{\nabla} f(x, \xi)\|_2^2 \leq \rho \|\nabla f(x)\|_2^2,$$

$L_f$ -smoothness and convexity, SGD with Nesterov’s acceleration has the following convergence rate:

$$\mathbb{E} f(x^k) - f(x^*) \leq \frac{2\rho^2 L_f}{k^2} \|x_0 - x^*\|_2^2,$$

i.e., similar to the deterministic method despite that the gradients are stochastic. SGC can be translated into the relative noise condition (4), making them related. Although a different method is used in our paper, the obtained results make it reasonable to expect a similar convergence in the concept of relative noise as in the absence of any noise.

The next plot illustrates the convergence of STM in the setting of  $\mu = 0$  and relative noise in the gradient for different values of the parameter  $\alpha$ .



**Figure 6.** Third test – the performance of STM with relative noise and  $\mu = 0$  for different values of  $\alpha$ .

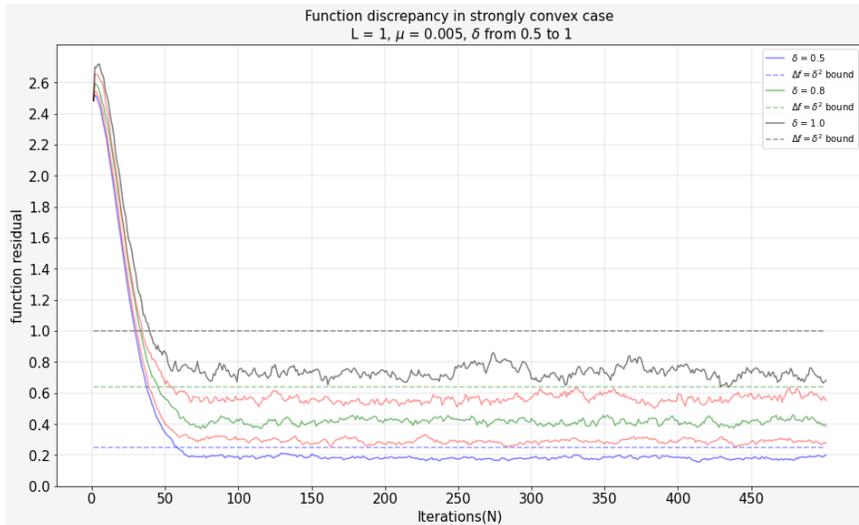
As we see, for  $\alpha \leq 0.71$ , the convergence of the method does not deteriorate and the value  $\alpha^* \approx 0.71$  can be seen as a threshold above which the method diverges.

We next explore the strongly convex setting with  $\mu > 0$  using the worst-case function [40, p.78]:

$$f(x) = \frac{\mu(\chi - 1)}{8} \left( x_1^2 + \sum_{j=1}^{n-1} (x_j - x_{j+1})^2 - 2x_1 \right) + \frac{\mu}{2} \|x\|_2^2,$$

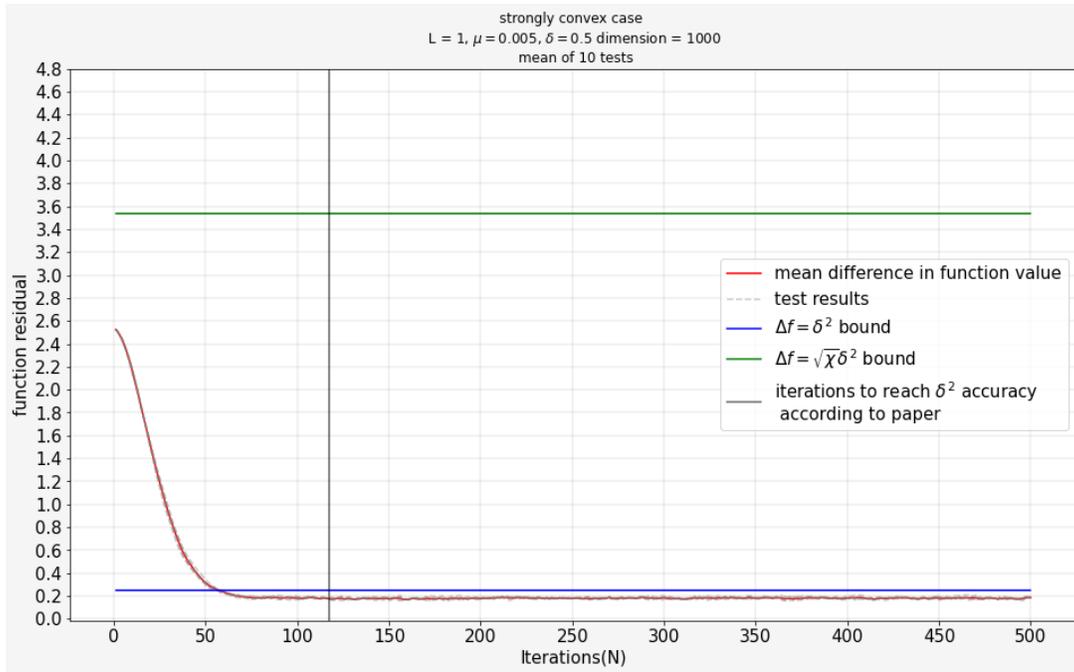
$$\chi = \frac{L}{\mu}.$$

We first consider the performance of STM with absolute noise for different values of  $\delta$ . Dashed lines represent the corresponding theoretical bound.



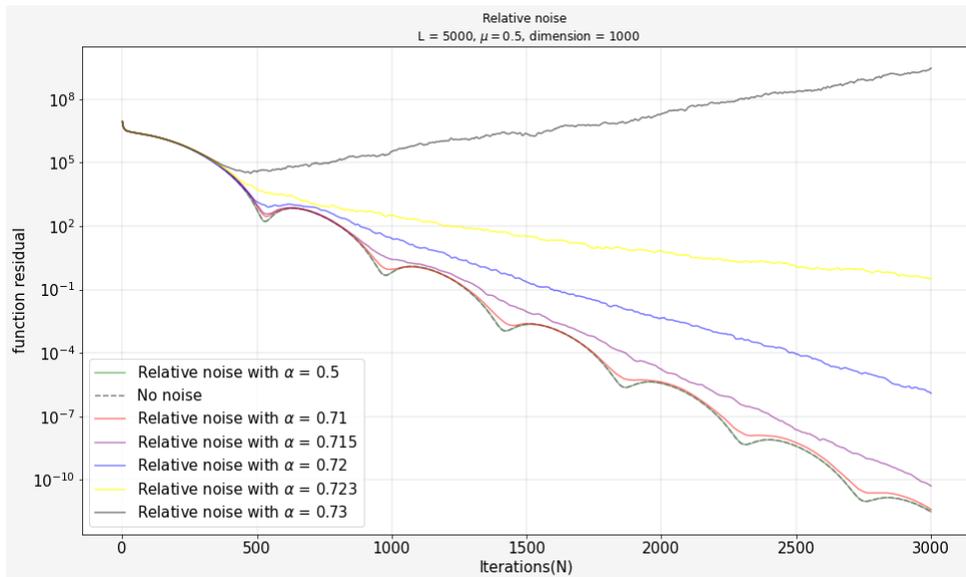
**Figure 7.** Fourth test – the performance of STM for  $\mu > 0$  and absolute noise  $\delta \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ .

The next plot confirms Theorem 4.7 and Remark 7.4.



**Figure 8.** Fifth test – mean of 30 tests, level of approximation and required number of steps.

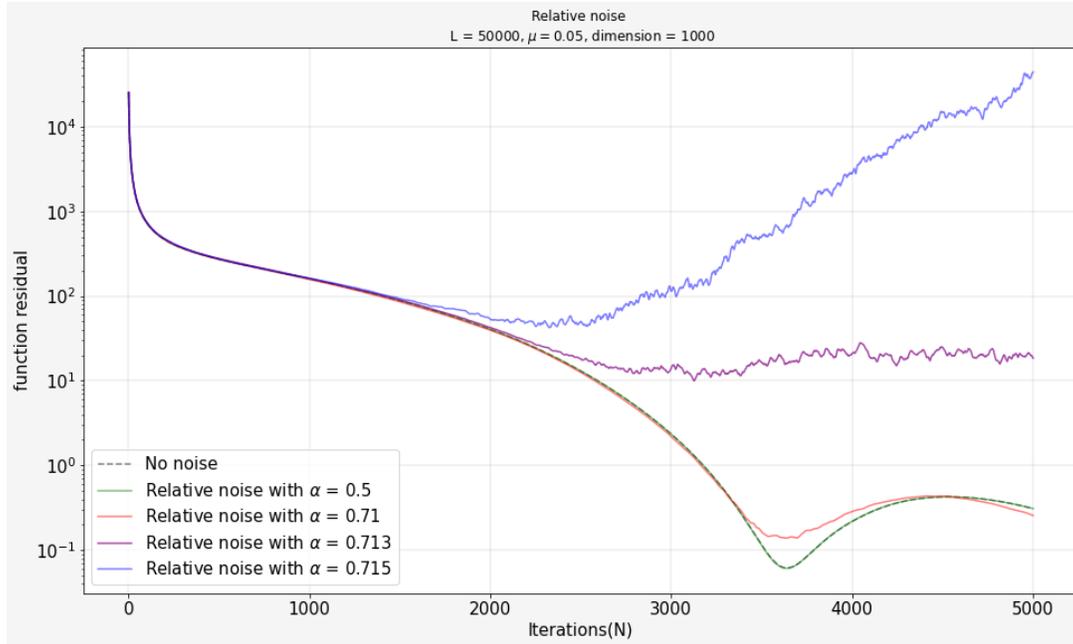
Next, similarly to the degenerate case  $\mu = 0$ , we consider the behavior of the method for different parameters  $\alpha$  when a relative noise is present in the gradient.



**Figure 9.** Sixth test – the performance of STM with relative noise and  $\mu > 0$  for different values of  $\alpha$ .

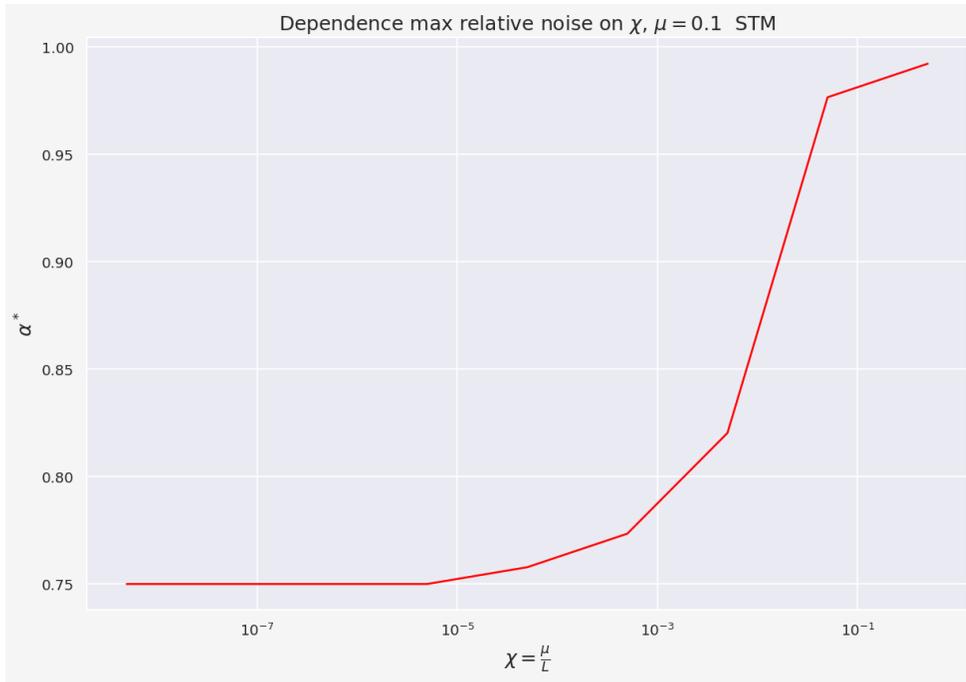
Note that, in the strongly convex case, we observe a similar effect as in the degenerate case: the algorithm converges for  $\alpha$ -values smaller than a certain threshold value

$\alpha^*$ .

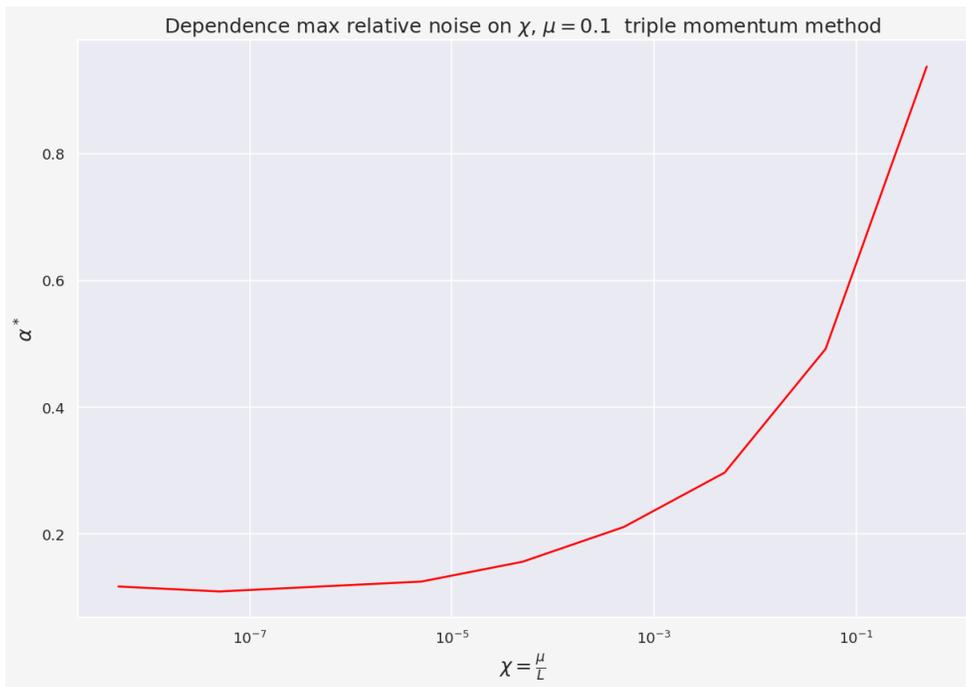


**Figure 10.** Seventh test – the performance of STM with relative noise and  $\mu > 0$  for different values of  $\alpha$ .

Finally, we compare STM and triple momentum method. Figures 11 and 12 show, that for the same parameters of the problem, STM is capable of converging at a much higher noise level than triple momentum algorithm.



**Figure 11.** Eighth test – threshold  $\alpha^*$  for different  $L$  and  $\mu = 0.1$ , for STM



**Figure 12.** Ninth test – threshold  $\alpha^*$  for different  $L$  and  $\mu = 0.1$ , for the Triple Momentum Method

## Acknowledgments

The authors are grateful to Eduard Gorbunov for useful discussions.

## References

- [1] A. Ajalloeian and S.U. Stich, *Analysis of sgd with biased gradient estimators*, deepai (2020).
- [2] A. Akhavan, M. Pontil, and A. Tsybakov, *Exploiting higher order smoothness in derivative-free optimization and continuous bandits*, Advances in Neural Information Processing Systems 33 (2020), pp. 9017–9027.
- [3] F. Bach and V. Perchet, *Highly-Smooth Zero-th Order Online Optimization*, in *29th Annual Conference on Learning Theory*, V. Feldman, A. Rakhlin, and O. Shamir, eds., Proceedings of Machine Learning Research Vol. 49, 23–26 Jun, Columbia University, New York, New York, USA. PMLR, 2016, pp. 257–283. Available at <http://proceedings.mlr.press/v49/bach16.html>.
- [4] A. Beck, *First-order methods in optimization*, SIAM, 2017.
- [5] A. Belloni, T. Liang, H. Narayanan, and A. Rakhlin, *Escaping the Local Minima via Simulated Annealing: Optimization of Approximately Convex Functions*, in *Proceedings of The 28th Conference on Learning Theory*, P. Grünwald, E. Hazan, and S. Kale, eds., Proceedings of Machine Learning Research Vol. 40, 03–06 Jul, Paris, France. PMLR, 2015, pp. 240–265. Available at <http://proceedings.mlr.press/v40/Belloni15.html>.
- [6] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization (Lecture Notes)*, Personal web-page of A. Nemirovski, 2015.
- [7] A.S. Berahas, L. Cao, K. Choromanski, and K. Scheinberg, *A theoretical and empirical comparison of gradient approximations in derivative-free optimization*, Foundations of Computational Mathematics (2021), pp. 1–54.
- [8] A. Beznosikov, A. Sadiev, and A. Gasnikov, *Gradient-Free Methods with Inexact Oracle for Convex-Concave Stochastic Saddle-Point Problem*, in *International Conference on Mathematical Optimization Theory and Operations Research*. Springer, 2020, pp. 105–119.
- [9] S. Bubeck, *et al.*, *Convex optimization: Algorithms and complexity*, Foundations and Trends® in Machine Learning 8 (2015), pp. 231–357.
- [10] M. Cohen, J. Diakonikolas, and L. Orecchia, *On acceleration with noise-corrupted gradients*, in *International Conference on Machine Learning*. PMLR, 2018, pp. 1019–1028.
- [11] A. Conn, K. Scheinberg, and L. Vicente, *Introduction to Derivative-Free Optimization*, Society for Industrial and Applied Mathematics, 2009, Available at <http://epubs.siam.org/doi/abs/10.1137/1.9780898718768>.
- [12] A. d’Aspremont, *Smooth optimization with approximate gradient*, SIAM Journal on Optimization 19 (2008), pp. 1171–1183.
- [13] O. Devolder, *Stochastic first order methods in smooth convex optimization*, CORE Discussion Paper 2011/70 (2011).
- [14] O. Devolder, *Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization*, Ph.D. diss., ICTEAM and CORE, Université Catholique de Louvain, 2013.
- [15] O. Devolder, F. Glineur, and Y. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Mathematical Programming 146 (2014), pp. 37–75. Available at <http://dx.doi.org/10.1007/s10107-013-0677-5>.
- [16] O. Devolder, F. Glineur, Y. Nesterov, *et al.*, *First-order methods with inexact oracle: the strongly convex case*, CORE Discussion Papers 2013016 (2013), p. 47.
- [17] D. Drusvyatskiy and L. Xiao, *Stochastic optimization with decision-dependent distributions*, Mathematics of Operations Research (2022).
- [18] D. Dvinskikh and A. Gasnikov, *Decentralized and parallelized primal and dual accelerated*

- methods for stochastic convex programming problems*, Journal of Inverse and Ill-posed Problems (2021).
- [19] D.M. Dvinskikh, A.I. Turin, A.V. Gasnikov, and S.S. Omelchenko, *Accelerated and non accelerated stochastic gradient descent in model generality*, Matematicheskie Zametki 108 (2020), pp. 515–528.
- [20] P. Dvurechensky, *Numerical methods in large-scale optimization: inexact oracle and primal-dual analysis*, HSE. Habilitation (2020).
- [21] P. Dvurechensky and A. Gasnikov, *Stochastic intermediate gradient method for convex problems with stochastic inexact oracle*, Journal of Optimization Theory and Applications 171 (2016), pp. 121–145. Available at <http://dx.doi.org/10.1007/s10957-016-0999-6>.
- [22] P. Dvurechensky, A. Gasnikov, and A. Kroshnin, *Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn’s Algorithm*, in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, eds., Proceedings of Machine Learning Research Vol. 80. 2018, pp. 1367–1376. arXiv:1802.04367.
- [23] P. Dvurechensky, S. Shtern, and M. Staudigl, *First-order methods for convex optimization*, EURO Journal on Computational Optimization 9 (2021), p. 100015.
- [24] A. d’Aspremont, D. Scieur, A. Taylor, *et al.*, *Acceleration methods*, Foundations and Trends® in Optimization 5 (2021), pp. 1–245.
- [25] Y.G. Evtushenko, *Optimization and fast automatic differentiation*, Computing Center of RAS, Moscow (2013).
- [26] O. Gannot, *A frequency-domain analysis of inexact gradient methods*, Mathematical Programming 194 (2022), pp. 975–1016. Available at <https://doi.org/10.1007/s10107-021-01665-8>.
- [27] A.V. Gasnikov, E.V. Gasnikova, Y.E. Nesterov, and A.V. Chernov, *Efficient numerical methods for entropy-linear programming problems*, Computational Mathematics and Mathematical Physics 56 (2016), pp. 514–524. Available at <http://dx.doi.org/10.1134/S0965542516040084>.
- [28] A. Gasnikov, *Universal gradient descent*, arXiv preprint arXiv:1711.00394 (2017).
- [29] A. Gasnikov, S. Kabanikhin, A. Mohammed, and M. Shishlenin, *Convex optimization in hilbert space with applications to inverse problems*, arXiv preprint arXiv:1703.00267 (2017).
- [30] A.V. Gasnikov and Y.E. Nesterov, *Universal method for stochastic composite optimization problems*, Computational Mathematics and Mathematical Physics 58 (2018), pp. 48–64.
- [31] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, Vol. 1, MIT press Cambridge, 2016.
- [32] E. Gorbunov, D. Dvinskikh, and A. Gasnikov, *Optimal decentralized distributed algorithms for stochastic convex optimization*, arXiv preprint arXiv:1911.07363 (2019).
- [33] E. Gorbunov, P. Dvurechensky, and A. Gasnikov, *An accelerated method for derivative-free smooth stochastic convex optimization*, arXiv preprint arXiv:1802.09022 (2018).
- [34] S.I. Kabanikhin, *Inverse and ill-posed problems: theory and applications*, Vol. 55, Walter De Gruyter, 2011.
- [35] D. Kamzolov, P. Dvurechensky, and A.V. Gasnikov, *Universal intermediate gradient method for convex problems with inexact oracle*, Optimization Methods and Software (2020), pp. 1–28.
- [36] G. Kotsalis, G. Lan, and T. Li, *Simple and optimal methods for stochastic variational inequalities, i: operator extrapolation*, SIAM Journal on Optimization 32 (2022), pp. 2041–2073.
- [37] G. Lan, *First-order and Stochastic Optimization Methods for Machine Learning*, Springer, 2020.
- [38] A.S. Nemirovski, *Regularizing properties of the conjugate gradient method for ill-posed problems*, Zhurnal Vychislitel’noi Matematiki i Matematicheskoi Fiziki 26 (1986), pp. 332–347.

- [39] A. Nemirovsky and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, J. Wiley & Sons, New York, 1983.
- [40] Y. Nesterov, *Lectures on convex optimization*, Vol. 137, Springer, 2018.
- [41] Y. Nesterov and V. Spokoiny, *Random gradient-free minimization of convex functions*, *Found. Comput. Math.* 17 (2017), pp. 527–566. Available at <https://doi.org/10.1007/s10208-015-9296-2>, First appeared in 2011 as CORE discussion paper 2011/16.
- [42] V. Novitskii and A. Gasnikov, *Improved exploiting higher order smoothness in derivative-free optimization and continuous bandit*, arXiv preprint arXiv:2101.03821 (2021).
- [43] F. Pedregosa and D. Scieur, *Average-case acceleration through spectral density estimation*, arXiv preprint arXiv:2002.04756 (2020).
- [44] B. Poljak, *Iterative algorithms for singular minimization problems*, in *Nonlinear Programming 4*, Elsevier, 1981, pp. 147–166.
- [45] B. Polyak, *Introduction to Optimization*, New York, Optimization Software, 1987.
- [46] B.T. Polyak and A.B. Tsybakov, *Optimal order of accuracy of search algorithms in stochastic optimization*, *Problemy Peredachi Informatsii* 26 (1990), pp. 45–53.
- [47] A. Risteski and Y. Li, *Algorithms and matching lower bounds for approximately-convex optimization*, *Advances in Neural Information Processing Systems* 29 (2016), pp. 4745–4753.
- [48] R.T. Rockafellar, *Convex analysis*, Vol. 36, Princeton university press, 1970.
- [49] D. Scieur and F. Pedregosa, *Universal Asymptotic Optimality of Polyak Momentum*, in *International Conference on Machine Learning*. PMLR, 2020, pp. 8565–8572.
- [50] F. Stonyakin, A. Tyurin, A. Gasnikov, P. Dvurechensky, A. Agafonov, D. Dvinskikh, M. Alkousa, D. Pasechnyuk, S. Artamonov, and V. Piskunova, *Inexact model: A framework for optimization and variational inequalities*, *Optimization Methods and Software* (2021). Available at <https://doi.org/10.1080/10556788.2021.1924714>.
- [51] F. Stonyakin, *Adaptive methods for variational inequalities, minimization problems and functional with generalized growth condition*, MIPT. Habilitation (2020).
- [52] A.B. Taylor, J.M. Hendrickx, and F. Glineur, *Smooth strongly convex interpolation and exact worst-case performance of first-order methods*, *Mathematical Programming* 161 (2017), pp. 307–345.
- [53] A. Tyurin, *Development of a method for solving structural optimization problems*, HSE. PhD Thesis (2020).
- [54] F. Vasilyev, *Optimization Methods*, Moscow, Russia: FP, 2002.
- [55] S. Vaswani, F. Bach, and M. Schmidt, *Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron*, in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1195–1204.