Noname manuscript No.
(will be inserted by the editor)

Stopping rules for accelerated gradient methods with additive noise in gradient

Vasin Artem · Alexander Gasnikov · Pavel Dvurechensky · Vladimir Spokoiny

Received: date / Accepted: date

Abstract In this article, we investigate an accelerated first-order method, namely, the method of similar triangles, which is optimal in the class of convex (strongly convex) problems with a Lipschitz gradient. The paper considers a model of additive noise in a gradient and a Euclidean prox-structure for not necessarily bounded sets. Convergence estimates are obtained in the case of strong convexity and its absence, and a stopping criterion is proposed for not strongly convex problems.

The research of A. Gasnikov was supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) 075-00337-20-03, project no. 0714-2020-0005. The work of A. Vasin was supported by Andrei M. Raigorodskii Scholarship in Optimization.

A. Vasin
Moscow Institute of Physics and Technology, Russia
A. Gasnikov
Moscow Institute of Physics and Technology, Russia
Institute for Information Transmission Problems RAS, Russia
Weierstrass Institute for Applied Analysis and Stochastics, Germany
P. Dvurechensky
Weierstrass Institute for Applied Analysis and Stochastics, Germany
Institute for Information Transmission Problems RAS, Russia
V. Spokoiny
Weierstrass Institute for Applied Analysis and Stochastics, Germany
Institute for Information Transmission Problems RAS, Russia

1 Introduction

We consider L-smooth (μ-strongly) convex optimization problem ($\mu \geq 0$):

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

This means that Q is convex set, and for all $x, y \in Q$:

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

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2.$$

In the analysis of the rates of convergence of different first-order methods these relations are typically rewrite as follows [15,9,6,28,4,26,38,35,51,21,48,23,13]

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

Note, that the last relation is a consequence of the previous ones and in general is not equivalent to them [50,26].

In many applications, especially for gradient-free methods (when estimating the gradient by finite differences [11,45,7]) optimization problems in infinite dimensional spaces (such examples arise when solving inverse problems [32,27]) instead of an access to $\nabla f(x)$ we have an access to its inexact approximation $\tilde{\nabla} f(x)$.

The two most popular conception of inexactness of gradient in practice are [43]: for all $x \in Q$

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \delta, \tag{2}$$

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \alpha\|\nabla f(x)\|_2, \quad \alpha \in [0, 1). \tag{3}$$

For the first conception (2) several results about the accumulation of error can be found in [43,12,10,1], but all these results are still far from to be optimistic in general. The reason was described in [42]. We can explain this reason by very simple example:

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{2}\sum_{i=1}^{n} \lambda_i \cdot (x^i)^2, \tag{4}$$

where $0 \geq \mu = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n = L$, $L \geq 2\mu$. The solution of this problem is $x_* = 0$. Assume that inexactness takes place only in the first component. That is instead of $\partial f(x)/\partial x^1 = \mu x^1$ we have an access to $\tilde{\partial} f(x)/\partial x^1 = \mu x^1 - \delta$. For simple gradient dynamic

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

we can conclude that for all $k \in \mathbb{N}$

$$x_k^1 \geq \frac{\delta}{L}\frac{1-(1-\mu/L)^k}{1-(1-\mu/L)} \geq \frac{\delta}{8\mu}. \tag{5}$$

Hence*

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

So we have a problem with (5), since μ can be to small ($\mu \lesssim \varepsilon$ – degenerate regime, where ε – desired accuracy in function value) in denominator of the RHS. We may expect even more serious troubles for accelerated gradient methods, since they are more sensitive to the level of noise [16,26]. The solution of this problem is well known (see, for example, [42,43,36]): to propose a stopping rule for the considered algorithm or to use regularization $\mu \sim \varepsilon$ [26]. Roughly speaking, for non accelerated algorithms in [42,43] it was proved that if $\delta \sim \varepsilon^2$, then it's possible to reach ε-accuracy in function value (with almost the same number of iterations as for no noise case $\delta = 0$) by applying computationally convenient stopping rule.

In this paper we show that it's sufficient to have $\delta \sim \varepsilon$ both for primal-dual non accelerated and accelerated gradient type methods [38, 26]. Primal-duality of methods is used to build computationally convenient stopping rule in degenerate regime. We emphasize, that the results $\delta \sim \varepsilon$ has a simple explanation (see section 2) and one might think that it is well known. But to the best of our knowledge the best results for accelerated methods require $\delta \sim \varepsilon^{3/2}$. So we consider our observation (that $\delta \sim \varepsilon$) to be an important part of this paper, although it has rather simple explanation.

The situation with the second criteria (3) is significantly better. For non accelerated algorithms inexactness in this case lead only to the deceleration of convergence $\sim (1-\alpha)^{-1}$-times [43]. This result holds true with the relaxed strong convexity assumption [26] (Polyak–Lojasiewicz condition). For accelerated case to the best of our knowledge this is an open problem to estimate accumulation of an error [26].

In this paper we show that if $\alpha \lesssim \left(\frac{\mu}{L}\right)^{3/4}$ in μ-strongly convex case and (on k-th iteration) $\alpha_k \lesssim \left(\frac{1}{k}\right)^{3/2}$ in degenerate regime we do not have any deceleration. Numerical experiments demonstrate that in general for α larger than mentioned above thresholds the convergence may slow down a lot up to divergence for considered accelerated method.

Note, that close results (with the requirement $\alpha \lesssim \left(\frac{\mu}{L}\right)^{5/4}$) in the case $\mu \gg \varepsilon$ were recently obtained by using another techniques in Stochastic Optimization with decision dependent distribution [18] and Policy Evaluation in Reinforcement Learning via reduction to stochastic Variational Inequality with Markovian noise [34]. In [34,18] it was assumed that

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq B\|x - x_*\|_2, \quad \alpha \in [0, 1). \tag{6}$$

*This bound corresponds to the worst-case philosophy concerning the choice of considered example for considered class of methods [37,38,9,26]. We expect more interesting results here by considering average-case complexity [47,41] (spectrum $\{\lambda_i\}$ average).

Since x_* is a solution, from Fermat's principle $\nabla f(x_*) = 0$. Therefore,

$$\|\nabla f(x)\|_2 = \|\nabla f(x) - \nabla f(x_*)\|_2 \leq L\|x - x_*\|_2.$$

So if (3) holds true then 6 also holds true with $B = \alpha L$.

2 Ideas behind the results

Important results in gradient error accumulation for first-order methods were developed in the cycle of works of O. Devolder, F. Glineur and Yu. Nesterov 2011–2014 [14,16,17,15]. In these works authors were motivated by (1). The idea is to "relax" (1), assuming inexactness in gradient. So they introduce inexact gradient $\tilde{\nabla} f(x)$, satisfying for all $x, y \in Q$

$$f(x) + \langle \tilde{\nabla} f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2 - \delta \leq f(y) \leq f(x) + \langle \tilde{\nabla} f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 + \delta. \quad (7)$$

Such a definition allows to develop precise theory for error accumulation for first-order methods.

Namely, it was proved that for non-accelerated gradient methods

$$f(x_k) - f(x_*) = O\left(\min\left\{\frac{LR^2}{k} + \delta, LR^2\exp\left(-\frac{\mu}{L}k\right) + \delta\right\}\right), \quad (8)$$

and for accelerated ones [16,20]

$$f(x_k) - f(x_*) = O\left(\min\left\{\frac{LR^2}{k^2} + k\delta, LR^2\exp\left(-\sqrt{\frac{\mu}{L}}\frac{k}{2}\right) + \sqrt{\frac{L}{\mu}}\delta\right\}\right), \quad (9)$$

where $R = \|x_{start} - x_*\|_2$ – the distance between starting point and the solution x_*. If x_* is not unique we take such x_* that is the closest to x_{start}. Both of these bounds are unimprovable [16,17]. See also [15,22,33] for "indermediate" situations between accelerated and non-accelerated methods.

Following to [17] we may reduce conception (2) to (7) by putting

$$\delta = \delta_{(7)} = \frac{\delta_{(2)}^2}{2L} + \frac{\delta_{(2)}^2}{\mu} \simeq \frac{\delta_{(2)}^2}{\mu} \quad (10)$$

and changing 2-times constant μ, L. The key observations here are

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

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

So, when $\mu > 0$ for non-accelerated methods this result is almost the same as we've obtained by considering example (4). To reach $f(x_k) - f(x_*) =$

ε when[†] $\mu \gtrsim \varepsilon$ we should put $\delta_{(2)} \sim \varepsilon$ that is good and rather expected. Unfortunately, for accelerated methods from this approach we will have $\delta_{(2)} \sim \varepsilon^{3/2}$. That is far from what we've declared in section 1. To improve this it's worth to propose more detailed conception rather then (7).

In the following works [16,19,20,49,48] the conception (7) was further developed

$$f(x) + \langle \tilde{\nabla} f(x), y - x\rangle + \frac{\mu}{2}\|y-x\|_2^2 - \delta_1\|y-x\|_2 \le f(y) \\ \le f(x) + \langle \tilde{\nabla} f(x), y - x\rangle + \frac{L}{2}\|y-x\|_2^2 + \delta_2. \qquad (11)$$

In this case (8) and (9) take a form for non-accelerated gradient methods

$$f(x_k) - f(x_*) \\ = O\left(\min\left\{\frac{LR^2}{k} + \tilde{R}\delta_1 + \delta_2, LR^2\exp\left(-\frac{\mu}{L}k\right) + \tilde{R}\delta_1 + \delta_2\right\}\right), \qquad (12)$$

and for accelerated ones [16,20]

$$f(x_k) - f(x_*) \\ = O\left(\min\left\{\frac{LR^2}{k^2} + \tilde{R}\delta_1 + k\delta_2, LR^2\exp\left(-\sqrt{\frac{\mu}{L}}\frac{k}{2}\right) + \tilde{R}\delta_1 + \sqrt{\frac{L}{\mu}}\delta_2\right\}\right), \qquad (13)$$

where $\tilde{R}$ is the maximal distance between generated points and the solution.

Thus from (12), (13) we may conclude that if $\tilde{R}$ is bounded,[‡] then by choosing

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

we will have the desired result: it is possible to reach $f(x_k) - f(x_*) = \varepsilon$ with $\delta_{(2)} \sim \varepsilon$.

But in general situation there is a problem in the assumption "if $\tilde{R}$ is bounded". As we may see from example (4) in general degenerate regime only such bound

$$\tilde{R} \simeq R + \frac{\delta_{(2)}}{\mu} \gtrsim R + \frac{\delta_{(2)}}{\varepsilon}$$

takes place [26]. This dependence spoils the result. The growth of $\tilde{R}$ we observe in different experiments. In the paper below we investigate this problem. In

[†]If $\mu \lesssim \varepsilon$, we can regularize the problem and guarantee the required condition [26]. Another advantage of strong convexity is possibility to use the norm of inexact gradient for the stopping criteria [26], like in [42]. But regularization requires some prior knowledge about the size of the solution [26]. Since we typically don not have such information the procedure becomes more difficult via applying the restarts [28,26].

[‡]In many situations this is true. For example, when Q is bounded, when $\mu \gg \varepsilon$.

particular, we propose an alternative approach to regularization[§] that is based on "early stopping"[¶] of considered iterative procedure by developing proper stopping rule.

Now we explain how to reduce relative inexactness (3) to (7) and to apply (9) when $\mu \gg \varepsilon$. Since $f(x)$ has Lipschitz gradient from (3), (7) we may derive that after k iterations (where k is greater than $\sqrt{L/\mu}$ on a logarithmic factor $\log\left(LR^2/\varepsilon\right)$, where ε – accuracy in function value)

$$f(x_k) - f(x_*) \overset{(9),(10)}{\simeq} \frac{\varepsilon}{2} + \sqrt{\frac{L}{\mu}}\frac{\delta_{(2)}^2}{\mu} \simeq \sqrt{\frac{L}{\mu}}\frac{\delta_{(2)}^2}{\mu}$$

$$\overset{(3),(7)}{\simeq} \sqrt{\frac{L}{\mu}}\frac{\alpha^2 \max_{t=1,\ldots,k} \|\nabla f(x_t)\|_2^2}{\mu} \leq \sqrt{\frac{L}{\mu}}\frac{2L\alpha^2 \max_{t=1,\ldots,k}(f(x_k) - f(x_*))}{\mu}$$

$$\lesssim \sqrt{\frac{L}{\mu}}\frac{4L\alpha^2\left(f(x_0) - f(x_*)\right)}{\mu}. \tag{14}$$

To guarantee that (restart condition)

$$f(x_k) - f(x_*) \leq \frac{1}{2}\left(f(x_0) - f(x_*)\right)$$

we should have $\alpha \lesssim \left(\frac{\mu}{L}\right)^{3/4}$. Then we restart the method. After $\log\left(\Delta f/\varepsilon\right)$ restarts we can guarantee the desired ε-accuracy in function value. In degenerate case the calculations are more tricky, but the idea remains the same with the replacing $\sqrt{L/\mu}$ to k (see (9)) that lead to $\alpha_k \lesssim \left(\frac{1}{k}\right)^{3/2}$. More accurate analysis in the subsequent part of the paper allows to formally obtain these bounds:

$$\alpha \lesssim \left(\frac{\mu}{L}\right)^{3/4},\ \alpha_k \lesssim \left(\frac{1}{k}\right)^{3/2}.$$

Below we'll concentrate only on accelerated method and choose the method with one projection (Similar Triangles Method (STM)), see [29,10,31,49,23] and reference there in. We decided to choose this method because: 1) it's primal-dual [29]; 2) has a nice theory of how to bound $\tilde{R}$ in no noise regime [29,38] ($\tilde{R} \leq R$) and noise one [31]; 3) and has previously been intensively investigated, see [23] and references there in.

3 Some motivation for inexact gradients

In this section we describe only two directions where inexact gradient 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 specialist in these directions.

[§]By using regularization we can guarantee $\mu \sim \varepsilon$ and therefore with $\delta_{(2)} \sim \varepsilon$ we have the desired $\tilde{R} \simeq R$.

[¶]This terminology is popular also in Machine Learning community, where "early stopping" is used also as alternative to regularization to prevent overfitting [30].

3.1 **Gradient-free methods**

In this section we consider convex optimization problem:

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

In some applications we do not have an access to gradient $\nabla f(x)$ of target function, but can calculate the value of[‖] $f(x)$ with accuracy δ_f [11]:

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

In this case there exist different conceptions for full gradient estimation (see [7] and references there in). For example (below we assuming that f has L_p-Lipschitz p-order derivatives in 2-norm),

– **(p-order finite-differences)**

$$\tilde{\nabla}_i f(x) = \frac{\tilde{f}(x + he_i) - \tilde{f}(x - he_i)}{2h} \text{ for } p = 2,$$

where e_i is i-th ort. Here we have

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

in the conception (2), see [7]. Optimal choice of h guarantees $\delta \sim \sqrt{n}\delta_f^{\frac{p}{p+1}}$. From section 1 we know that it is possible to solve the problem with accuracy (in function value) $\varepsilon \sim \delta$. Hence,

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

Unfortunately, such simple idea does not give tight lower bound in the class of algorithm that has sample complexity $\text{Poly}(n, \frac{1}{\varepsilon})$ [45] (obtained for $p = 0$, that is only Lipschitz-continuity of f required):

$$\delta_f \sim \max\left\{\frac{\varepsilon^2}{\sqrt{n}}, \frac{\varepsilon}{n}\right\}. \tag{15}$$

Note, that instead of finite-difference approximation approach in some applications we can use kernel approach [44,3]. The interest to this alternative has grown last time [2,40].

[‖] Note, that the approach describe above required that function values should be available not only in Q, but also in some (depends on approach we used) vicinity of Q. This problem can be solved in a two different ways. The first one is "margins inward approach" [8]. The second one is "continuation" f to $\mathbb{R}^n$ with preserving of convexity and Lipschitz continuity [45]: $f_{new}(x) := f\left(\text{proj}_Q(x)\right) + \alpha \min_{y \in Q} \|x - y\|_2$.

– **(Gaussian Smoothed Gradients)**

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

where $e \in N(0, I_n)$ is standard normal random vector. Here we have

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

in the conception (2), see [39,7]. Optimal choice of h guarantees $\delta \sim (n\delta_f)^{\frac{p}{p+1}}$. Hence,

$$\delta_f \sim \frac{\varepsilon^{\frac{p+1}{p}}}{n}.$$

That is also does not match the lower bound. Moreover, here (and in the approach below) we have additional difficulty: how to estimate $\tilde{f}(x)$. We can do it only roughly, for example, by using Monte Carlo approach [7]. This is a payment for the better quality of approximation!

– **(Sphere Smoothed Gradients)**

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

where e is random vector with uniform distribution in a unit sphere (with center at 0) in $\mathbb{R}^n$. Here we have

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

in the conception (2), see [7]. Optimal choice of h guarantees $\delta \sim (n\delta_f)^{\frac{p}{p+1}}$. Hence,

$$\delta_f \sim \frac{\varepsilon^{\frac{p+1}{p}}}{n}.$$

That is also does not match the lower bound. One can consider that the last two approach are almost the same, but below we describe more accurate result concerning Sphere smoothing. We do not know how to obtain such a result for Gaussian smoothing. The results is as follows [16,45]: For Sphere smoothed gradient in conception (7) we have

$$\delta \simeq 2L_0 h + \frac{\sqrt{n}\delta_f \tilde{R}}{h}, \tag{16}$$

where L_0 is Lipschitz constant of f and $L = \min\left\{L_1, \frac{7L_0^2}{h}\right\}$ in (7), when $p = 1$ and $L = \frac{7L_0^2}{h}$, when $p = 0$. The bound (16) is more accurate than the previous ones, since it corresponds to the first part of the lower bound (15). Indeed, by choosing properly h in (16) we obtain $\varepsilon \sim \delta \sim n^{1/4}\delta_f^{1/2}$. Hence,

$$\delta_f \sim \frac{\varepsilon^2}{\sqrt{n}}.$$

The rest part ($\delta_f \sim \frac{\varepsilon}{n}$) of lower bound (15) is also tight, see [5].

The last calculations (see (16)) additionally confirm that the conception of inexactness and algorithms we use and develop in section 2 are also tight (optimal) enough. Otherwise, it'd be hardly possible to reach lower bound by using gradient-free methods reduction to gradient ones and proposed analysis of an error accumulation for gradient-type methods.

3.2 **Inverse problems**

Another rather big direction of research where gradients are typically available only approximately is optimization in a Hilbert spaces [52]. Such optimization problems arise, in particular, in inverse problems theory [32].

We start with the reminder of how to calculate a derivative in general Hilbert space. Let

$$J(q) := J(q, u(q)),$$

where $u(q)$ is determine as unique solution of

$$G(q, u) = 0.$$

Assume that $G_q(q, u)$ is invertible, then

$$G_q(q, u) + G_u(q, u)\nabla u(q) = 0,$$

hence

$$\nabla u(q) = -\left[G_u(q, u)\right]^{-1} G_q(q, u).$$

Therefore

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

The same result could be obtained by considering Lagrange functional

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

with

$$L_u(q, u; \psi) = 0, G_q(q, u) = 0$$

and

$$\nabla J(q) = L_q(q, u; \psi).$$

Indeed, by simple calculations we can relate these two approaches, where

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

Now we demonstrate this technique on inverse problem for elliptic initial-boundary value problem.

Let u be the solution of the following problem (P)

$$u_{xx} + u_{yy} = 0, \quad x, y \in (0, 1),$$

$$u(1, y) = q(y), \quad y \in (0, 1),$$

$$u_x(0,y)=0,\quad y\in(0,1),$$
$$u(x,0)=u(x,1)=0,\quad x\in(0,1).$$

The first two relations

$$-u_{xx}-u_{yy}=0,\quad x,y\in(0,1),$$
$$q(y)-u(1,y)=0,\quad y\in(0,1),$$

we denote as $G(q,u)=\bar{G}\cdot(q,u)=0$ and the last two ones as $u\in Q$.

Assume that we want to estimate $q(y)\in L_2(0,1)$ 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) [32]. This is an inverse problem. We can reduce this problem to optimization one [32]:

$$\min_q \mathfrak{J}(q):=\min_{u:\ \bar{G}\cdot(q,u)=0,u\in Q} J(q,u):=J(u)=\int_0^1|u(0,y)-b(y)|^2dy. \quad (17)$$

We can solve (17) numerically. This problem is convex quadratic optimization problem. We can directly apply Lagrange multipliers principle to (17), see [52]:

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

To obtain conjugate problem for ψ we should vary $L(q,u;\psi)$ on δu satisfying $u\in Q$:

$$\delta_u L(q,u;\psi)=2\int_0^1(u(0,y)-b(y))\,\delta u(0,y)dy-$$
$$\int_0^1\int_0^1(\delta u_{xx}+\delta u_{yy})\,\psi(x,y)dxdy-\int_0^1\delta u(1,y)\lambda(y)dy, \quad (18)$$

where

$$\delta u_x(0,y)=0,\quad y\in(0,1),$$
$$\delta u(x,0)=\delta u(x,1)=0,\quad x\in(0,1).$$

Using integration by part, from (18) we can derive

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

Consider corresponding conjugate problem (D)

$$\psi_{xx} + \psi_{yy} = 0, \quad x, y \in (0,1)\,,$$

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

$$\psi\,(1,y) = 0, \quad y \in (0,1)\,,$$

$$\psi\,(x,0) = \psi\,(x,1) = 0, \quad x \in (0,1)$$

and additional relation between Lagrange multipliers

$$\lambda(y) = \psi_x(1,y), \quad y \in (0,1)\,. \tag{19}$$

These relations appears 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 [46]

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

from the Demyanov–Danskin's formula [46]**

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

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

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

saddle-point problem. Since $\delta_\psi L(q,u;\psi) = 0$ entails $\bar{G}\cdot(q,u) = 0$ that is form (P) if we add $u \in Q$ and $\delta_u L_(q,u;\psi) = 0$, when $u \in Q$ entails (D) as we've shown above.

Note also that

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

Hence, due to (19)

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

So we reduce $\nabla\mathfrak{J}(q)(y)$ calculation to the solution of two correct initial-boundary value problem for elliptic equation in a square (P) and (D) [32].

**The same result in more simple situation (without additional constraint $u \in Q$) we consider at the beginning of this section. We don't apply Demyanov–Danskin's formula and use inverse function theorem.

This result can be also interpreted in a little bit different manner. We introduce a linear operator

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

Here $u(x,y)$ is the solution of problem (P). It was shown in [32] that

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

Conjugate operator is [32]

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

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

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

we can write

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

that completely corresponds to the same scheme as described above:

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

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

So inexactness in gradient $\nabla\mathfrak{J}(q)$ arises since we can solve (P) and (D) only numerically.

The described above technique can be applied to many different inverse problems [32] and optimal control problems [52]. Note that for optimal control problems in practice another strategy widely used. Namely, instead of approximate calculation of gradient, optimization problem replaced by approximate one (for example, by using finite-differences schemes). For this reduced (finite-dimensional) problem the gradient is typically available precisely [24]. Moreover, in [24] the described above Lagrangian approach is based to explain the core of automatic differentiation where the function calculation tree represented as system of explicitly solvable interlocking equations.

4 Basic assumptions and problem description

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

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

Assume that

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leqslant \delta, \tag{20}$$

where $\tilde{\nabla} f(x)$ oracle gradient value. We consider two cases: Q is a compact set and Q is unbounded, for example $\mathbb{R}^n$. We define the constant:

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

to be the distance between the solution x^* and starting point x_{start}, if x^* is not unique we take such x^* that is the closest to x_{start}. We assume that function f has Lipschitz gradient with constant L_f:

$$\forall x, y \in Q,\ \|\nabla f(x) - \nabla f(y)\|_2 \leqslant L_f \|x - y\|_2. \tag{21}$$

This implies inequality:

$$\forall x, y \in Q,\ f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_f}{2}\|x - y\|_2^2. \tag{22}$$

We will use following lemma:

Lemma 1 (Fenchel inequality) *Let $(\mathcal{E}, \langle\cdot,\cdot\rangle)$ – euclidean space, then $\forall \lambda \in \mathbb{R}_+, \forall u, v \in \mathcal{E}$ the inequality holds:*

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

From previous assumptions we can get upper bound with inexact oracle.

Claim 1 *$\forall x, y \in Q$, the following estimate holds:*

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

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

Proof The proof follows from

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

We also assume strong convexity of f with parameter μ, however μ may equal zero – this corresponds to the ordinary convexity, supposed initially. Further we will use only a consequence of this:

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|_2^2 \leqslant f(y). \tag{23}$$

We obtain similar to claim 1 two lower bounds with inexact oracle.

Claim 2 *$\forall x, y \in Q$, the following estimate holds:*

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

where $\delta_1 = \delta$.

Proof Using Cauchy inequality and (23) we obtain:

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

Claim 3 *$\forall x, y \in Q$, if in* (23) *$\mu \neq 0$, the following estimate holds,*

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

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

Proof Trivial calculations bring

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

Using lemma 1 we obtain:

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

The last two inequalities give different results in convergence under certain conditions. We will study two models based on statements 2, 3 and we will denote them by the index $\boldsymbol{\tau}$, that is denote:

$$\begin{aligned} \mu_1 &= \mu, \\ \mu_2 &= \frac{\mu}{2}. \end{aligned} \tag{24}$$

Further in the text, we will use statements 3 and 2 in the notation corresponding to (24).

5 Similar Triangles Method and its properties

In this section we describe an accelerated method we choose to investigate gradient-error accumulation.

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

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

Figure 5 describes the position of the vertices. On the sides, not their lengths are marked, but the relationships in the corresponding sides in the similarity of triangles. In the case $Q = \mathbb{R}^n$, we can simplify the step of the algorithm by replacing it with:

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

This equality could be obtained explicitly solving the optimization problem, that is:

$$\psi_k(x) \to \min_{x \in \mathbb{R}^n},$$

$$\nabla\psi_k(x) = \nabla\psi_{k-1}(x) + \alpha_k\left(\tilde{\nabla} f(\tilde{x}_k) + \mu_\tau(x - \tilde{x}_k)\right),$$

$$\nabla\psi_n(x) = \sum_{z=0}^{n} \alpha_k\left(\tilde{\nabla} f(\tilde{x}_k) + \mu_\tau(x - \tilde{x}_k)\right) + (x - \tilde{x}_0),$$

$$\nabla\psi(z_k) = 0,$$

$$(1 + \mu A_k)z_k = -\left(\sum_{j=0}^{k} \alpha_j\left(\tilde{\nabla} f(\tilde{x}_j) - \mu\tilde{x}_j\right) + \tilde{x}_0\right)$$

From last equation for k and $k-1$ we get suggested step.
We define sequence:

$$\tilde{R}_N = \max_{0 \leqslant k \leqslant N}\{\|z_k - x^*\|_2, \|x_k - x^*\|_2, \|\tilde{x}_k - x^*\|_2\}.$$

We will also write down several identities that we will need in the proofs

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

Some of the identities can be obtained from geometric considerations, for example, from a figure, others by direct substitution into the definitions of the sequences $x_k, \tilde{x}_k, z_k$. Also very important are the estimates for the sequence A_k.

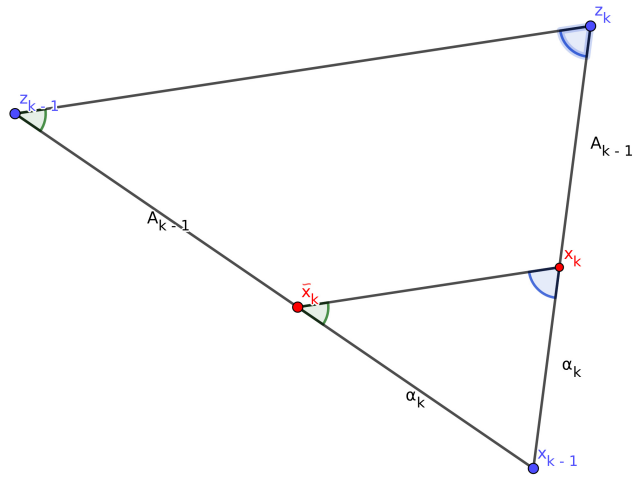

Fig. 1 Geometry of Similar Triangles method [29]

Claim 4 *If $\mu \neq 0$ and $\forall k \in \mathbb{N}$ the following inequality holds:*

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

where

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

Proof Using relation between

$$A_k(1 + \mu_\tau) = L\alpha_k^2$$

We get quadratic equation for A_k

$$LA_k^2 - A_k\left(1 + \mu_\tau A_{k-1} + 2LA_{k-1}\right) + LA_{k-1}^2$$

Solving it, we get:

$$\mathcal{D} = \left(1 + \mu_\tau A_{k-1} + 2LA_{k-1}\right)^2 - 4L^2 A_{k-1},$$
$$A_k \geqslant A_{k-1}\left(1 + \frac{\mu_\tau + 1}{L} + \sqrt{\frac{\mu_\tau}{L}}\right) \geqslant A_{k-1}\left(1 + \frac{\mu_\tau}{2L} + \sqrt{\frac{\mu_\tau}{L}}\right)$$

Corollary 1 *From inequality for $x < 1$*

$$1 + x \geqslant e^{\frac{x}{2}}$$

We get:

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

Claim 5 *If $\mu \neq 0$ $\forall k \in \mathbb{N}$ the following inequality holds:*

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

Proof According to the previous designations:

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

Using previous claim we can reduce this amount exponentially.

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

Claim 6 *If $\mu = 0$ then:*

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

Proof If $\mu = 0$, then $A_k = L\alpha_k^2$, and, solving the quadratic equation and using $A_{k-1} = L\alpha_{k-1}^2$ 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 get that:

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

Claim 7 *If $\mu = 0$ we have:*

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

Proof The proof follows from the simple calculations:

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

Lemma 2 *$\forall k \geqslant 1$ the following inequality holds:*

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

Proof From the definition of the ψ_{k-1} function, it has a minimum at the point z_{k-1}, then:

$$\langle\nabla\psi_{k-1}(z_{k-1}), z_k - z_{k-1}\rangle \geqslant 0,\ \nabla\psi_{k-1}(z_{k-1}) = (z_{k-1} - \tilde{x}_0) + \\ + \sum_{j=0}^{k-1}\alpha_j\left(\tilde{\nabla} f\left(\tilde{x}_j\right) + \mu_\tau\left(z_{k-1} - \tilde{x}_j\right)\right) \Rightarrow$$

$$\Rightarrow \psi_k(z_k) = \psi_{k-1}(z_k) + \alpha_k\left(f\left(\tilde{x}_k\right) + \langle\tilde{\nabla} f\left(\tilde{x}_k\right), z_k - \tilde{x}_k\rangle + \frac{\mu_\tau}{2}\|z_k - \tilde{x}_k\|_2^2\right) =$$

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

From equality:

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

and from fact:

$$\langle\nabla\psi_{k-1}(z_{k-1}), z_k - z_{k-1}\rangle \geqslant 0$$

$$\langle z_{k-1} - \tilde{x}_0, z_k - z_{k-1}\rangle \geqslant \sum_{j=0}^{k-1}\alpha_j\langle\tilde{\nabla} f(\tilde{x}_j) + \mu_\tau(z_{k-1} - \tilde{x}_j), z_{k-1} - z_k\rangle.$$

We get:

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

Using the linearity of the dot product, we split the sum by two and 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 reducing the terms we finally get:

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

Remark 1

In the case $\mu = 0$, we obtain a corollary from the strongly convexity of functions ψ_k and their definition, that is:

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

6 Main results

Here we will describe some results based on the previously presented lemmas and statements.

6.1 **Additive noise and main theorems.**

Theorem 1 *For algorithm* (1) $\forall k \in \mathbb{N}$ *the following inequality holds:*

$$A_k f(x_k) \leqslant \psi_k(z_k) + \delta_2 \sum_{j=0}^{k} A_j + 2\tilde{R}_k \delta_1 A_k.$$

Proof Base, $k = 0$:
Remind, that:

$$\alpha_0 = \frac{1}{L},$$
$$\psi_0(x) = \alpha_0 \left(f(x_0) + \langle \tilde{\nabla} f(\tilde{x}_0), x - \tilde{x}_0 \rangle + \frac{\mu_\tau}{2} \|x - \tilde{x}_0\|_2^2 \right) + \frac{1}{2} \|x - \tilde{x}_0\|_2^2.$$

Then we get:

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

Induction step:

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

Using equations (25) we obtain:

$$A_k f(x_k) - A_{k-1}\delta_1 \|x_{k-1} - \tilde{x}_k\|_2 \leqslant$$
$$\leqslant A_{k-1} \left(f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), x_{k-1} - \tilde{x}_k \rangle \right) + \alpha_k \left(f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle \right) +$$
$$+ \frac{(1 + \mu_1 A_{k-1})}{2} \|z_k - z_{k-1}\|_2^2 + A_k \delta_2 - A_{k-1}\delta_1 \|x_{k-1} - \tilde{x}_k\|_2 \leqslant$$
$$\leqslant A_{k-1} f(x_{k-1}) + \alpha_k (f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle) +$$
$$+ \frac{1 + \mu_1 A_{k-1}}{2} \|z_k - z_{k-1}\|_2^2 + A_k \delta_2.$$

Using the induction hypothesis, we obtain:

$$A_k f(x_k) - A_{k-1}\delta_1 \|x_{k-1} - \tilde{x}_k\|_2 \leqslant \psi_{k-1}(z_{k-1}) + \delta_2 \sum_{j=0}^{k-1} A_j + 2\tilde{R}\delta_1 A_{k-1} +$$
$$+ \frac{1 + \mu_1 A_{k-1}}{2} \|z_k - z_{k-1}\|_2^2 + \alpha_k \left(f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), z_k - \tilde{x}_k \rangle \right) + A_k \delta_2.$$

Using lemma 2 we can get:

$$A_k f(x_k) \leqslant A_{k-1}\delta_1\|x_{k-1} - \tilde{x}_k\|_2 + \psi_k(z_k) + \delta_2\sum_{j=0}^{k} A_j + 2\tilde{R}\delta_1 A_{k-1} =$$

$$= \psi_k(z_k) + \delta_2\sum_{j=0}^{k} A_j + 2\tilde{R}\delta_1 A_{k-1} + \alpha_k\|\tilde{x}_k - z_{k-1}\|_2 \leqslant$$

$$\leqslant \psi_k(z_k) + \delta_2\sum_{j=0}^{k} A_j + 2\tilde{R}\delta_1 A_{k-1} + \alpha_k\left(\|z_{k-1} - x^*\|_2 + \|\tilde{x}_k - x^*\|_2\right)\delta_1 \leqslant$$

$$\leqslant \psi_k(z_k) + \delta_2\sum_{j=0}^{k} A_j + 2\tilde{R}\delta_1 A_{k-1} + 2\alpha_k\tilde{R}\delta_1 \Rightarrow$$

$$\Rightarrow A_k f(x_k) \leqslant \psi(z_k) + \delta_2\sum_{j=0}^{k} A_j + 2\tilde{R}\delta_1 A_k.$$

Remark 2

We should note that this inequality is true both in the case of $\mu \neq 0$ and in the case of $\mu = 0$.

Theorem 2 *If $\mu \neq 0$ $\forall k \in \mathbb{N}$ the following inequality holds:*

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

The proof repeats verbatim theorem 1, except for claim 2, replaced by claim 3.

Theorem 3 *If $\delta = 0$ then $(\forall k \in \mathbb{N})$ $\tilde{R}_k \leqslant R$, where $R = \|\tilde{x}_0 - x^*\|_2$.*

Proof Using Theorem 1 we get $A_k f(x_k) \leqslant \psi_k(z_k)$ then, using:

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

then:

$$\frac{1}{2}\|z_k - x^*\|_2^2 = \frac{1}{2}\|z_k - x^*\|_2^2 + A_k f(x_k) - A_k f(x_k)$$

$$\leqslant \psi_k(z_k) + \frac{1}{2}\|z_k - x^*\|_2^2 - A_k f(x_k) \leqslant \psi_k(x^*) + \frac{1}{2}\|z_k - x^*\|_2^2 - A_k f(x_k)$$

$$\leqslant \sum_{j=0}^{k}\alpha_j\left(f(\tilde{x}_j) + \langle\nabla f(\tilde{x}_k), x^* - \tilde{x}_k\rangle + \frac{\mu}{2}\|x^* - \tilde{x}_k\|_2^2\right) + \frac{1}{2}\|x^* - \tilde{x}_0\|_2^2 - A_k f(x_k)$$

$$\leqslant A_k f(x_k) - A_k f(x_k) + \frac{1}{2}\|\tilde{x}_0 - x^*\|_2^2 = \frac{1}{2}R^2.$$

We now prove bounds for all sequences by induction. Base for $\tilde{x}_0$ is obvious, then steps:

$$\|x_k - x^*\|_2 = \|\frac{A_{k-1}}{A_k}(x_{k-1} - x^*) + \frac{\alpha_k}{A_k}(z_k - x^*)\|_2 \leqslant$$
$$\leqslant \frac{A_{k-1}}{A_k}\|x_{k-1} - x^*\|_2 + \frac{\alpha_k}{A_k}\|z_k - x^*\|_2 \leqslant R.$$

Similarly for $\tilde{x}_k$, using:

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

Theorem 4 (convergence in function) *Both inequalities take place with* $\mu \neq 0$

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

$$f(x_N) - f(x^*) \leqslant LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu_2}{L}}N\right) + \left(1 + \sqrt{\frac{L}{\mu_2}}\right)\delta_2 + \left(1 + \sqrt{\frac{L}{\mu_2}}\right)\delta_3.$$

Proof Using, all of the above is easy to show what is required, the proof of both convergence is the same with the replacement of theorem 1 by theorem 2 and replacement claim 2 by claim 3, therefore, we present only the proof of the first inequality.

$$A_N f(x_N) \leqslant \psi_N(z_N) + \delta_2\sum_{j=0}^{N}A_j + 2\tilde{R}\delta_1 A_N \leqslant \frac{1}{2}\|x^* - \tilde{x}_0\|_2^2 +$$
$$+ \delta_2\sum_{j=0}^{N}A_j + 2\tilde{R}\delta_1 A_N + \sum_{j=0}^{N}\alpha_k(f(\tilde{x}_j) + \langle\tilde{\nabla}f(\tilde{x}_j), x^* - \tilde{x}_j\rangle + \frac{\mu_1}{2}\|x^* - x_i\|_2^2) \leqslant$$
$$\leqslant \delta_2\sum_{j=0}^{N}A_j + 2\tilde{R}\delta_1 A_N + \sum_{j=0}^{N}\alpha_k(\tilde{R}\delta_1 + f(x^*)) + \frac{1}{2}R^2 =$$
$$= \delta_2\sum_{j=0}^{N}A_j + 3\tilde{R}\delta_1 A_N + A_N f(x^*) + \frac{1}{2}R^2 \Longleftrightarrow$$
$$\Longleftrightarrow f(x_N) - f(x^*) \leqslant LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{L}}N\right) + \left(1 + \sqrt{\frac{L}{\mu_1}}\right)\delta_2 + 3\tilde{R}\delta_1.$$

Remark 3

If $\mu = 0$ we can get analogue of the first convergence, repeating the proof using claims 6, 7

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

Remark 4

Suppose $\mu = 0$, then consider the auxiliary problem:

$$f^\mu(x) = f(x) + \frac{\mu}{2}\|x - \tilde{x}_0\|_2^2 \to \min_{x \in Q},$$
$$\nabla f^\mu(x) = \nabla f(x) + \mu(x - \tilde{x}_0),$$
$$\tilde{\nabla} f^\mu(x) = \tilde{\nabla} f(x) + \mu(x - \tilde{x}_0),$$
$$\|\tilde{\nabla} f^\mu(x) - \nabla f^\mu(x)\|_2 = \|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leqslant \delta.$$

The resulting function will satisfy the condition that the gradient is Lipschitz, that is $\forall x, y \in Q$:

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

Remeber, that $\mu \leqslant L$. That is, we can let $L^\mu = 4L_f = 2L$. The resulting function will already be strongly convex, which means that the second model is applicable to it $\tau = 2$. Using theorem 4 we can get the following inequality:

$$x_\mu^* = \operatorname*{argmin}_{x \in Q} f^\mu(x),$$
$$R_\mu = \|x_\mu^* - \tilde{x}_0\|_2,$$
$$f^\mu(x_k) - f^\mu(x_\mu^*) \leqslant \frac{L^\mu R_\mu^2}{2\lambda^k_{\frac{\mu}{2}, 2L^\mu}} + \left(1 + \sqrt{\frac{4L}{\mu}}\right)(\delta_2 + \delta_3) \Rightarrow$$
$$f^\mu(x_k) - f^\mu(x_\mu^*) \leqslant 2LR_\mu^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{4L}}k\right) + \left(1 + \sqrt{\frac{4L}{\mu}}\right)\left(\frac{1}{2L} + \frac{1}{\mu}\right)\delta^2,$$
$$f^\mu(x_\mu^*) \leqslant f(x^*) + \frac{\mu}{2}R^2.$$

Then we can get convergence rate for not regularized function:

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

Using strong convexity of the function f^μ we get:

$$f(x^*) + \frac{\mu}{2}R_\mu^2 \leqslant f(x_\mu^*) + \frac{\mu}{2}R_\mu^2 = f^\mu(x_\mu^*) \leqslant f^\mu(x^*) = f(x^*) + \frac{\mu}{2}R^2 \Rightarrow$$
$$R_\mu \leqslant R.$$

Finally we get convergence:

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

We choose value for parameter μ in the remark 9.

Remark 5

If we consider the problem in the first model $\tau = 1$, the case $\mu = 0$ and assume that $\|x^*\|_2 \leqslant R_*$. Then we choose a starting point for the STM algorithm in a ball of radius R_*, specifically put $\tilde{x}_0 = x_{start} = 0$.

$$R = \|x^* - \tilde{x}_0\|_2 \leqslant R_*.$$

Let us formulate a stopping rule for the this model ($\forall \zeta > 0$).

$$f(x_k) - f(x^*) \leqslant k\delta_2 + R_*\delta_1 + \delta_1 \sum_{j=1}^{k} \frac{\alpha_j}{A_k}\|\tilde{x}_j - z_{j-1}\|_2 + \zeta.$$

Lemma 3 (Bound for $\tilde{R}$) *Before the stopping criterion is satisfied, the following inequality holds:*

$$\tilde{R} \leqslant R.$$

Proof Note, that from $\|z_k - x^*\|_2 \leqslant R$ we get $\|x_k - x^*\|_2 \leqslant R$, $\|\tilde{x}_k - x^*\|_2 \leqslant R$ similarly to theorem 3. But it's worth noting that to estimate $\|\tilde{x}_k - x^*\|_2$, only inequalities are required for all $j \leqslant k - 1$.

$$\|\tilde{x}_k - x^*\|_2 = \|\frac{A_{k-1}}{A_k}(x_{k-1} - x^*) + \frac{\alpha_k}{A_k}(z_{k-1} - x^*)\|_2 \leqslant \\ \leqslant \frac{A_{k-1}}{A_k}\|x_{k-1} - x^*\|_2 + \frac{\alpha_k}{A_k}\|z_{k-1} - x^*\|_2 \leqslant R.$$

An analysis of the proof of theorem 1 gives a stronger convergence:

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

Then, using the convexity of the function ψ_k we get:

$$A_k f(x_k) + \frac{1}{2}\|z_k - x^*\|_2^2 \leqslant \frac{1}{2}\|z_k - x^*\|_2 + \psi_k(z_k) + \delta_2 \sum_{j=0}^{k} A_j +$$

$$+ \delta_1 \sum_{j=1}^{k} \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 \leqslant \psi_k(x^*) + \delta_2 \sum_{j=0}^{k} A_j +$$

$$+ \delta_1 \sum_{j=1}^{k} \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 \leqslant \frac{1}{2}R^2 + A_k f(x^*) + \delta_2 \sum_{j=0}^{k} A_j +$$

$$+ \delta_1 \sum_{j=1}^{k} \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 + \delta_1 \sum_{j=0}^{k} \alpha_j \|z_k - x^*\|_2 \Rightarrow \frac{1}{2}(R^2 - \|z_k - x^*\|_2) \geqslant$$

$$\geqslant A_k \left((f(x_k) - f(x^*)) - \left(k\delta_2 + \delta_1 \sum_{j=1}^{k} \frac{\alpha_j}{A_k} \|\tilde{x}_j - z_{j-1}\|_2 + R_*\delta_1 + \zeta \right) \right) \geqslant 0.$$

Therefore, when the stopping criterion is met, we will receive the estimate:

$$f(x_k) - f(x^*) \leqslant k\delta_2 + \delta_1 R_* + \delta_1 \sum_{j=0}^{k} \alpha_j \|\tilde{x}_j - z_{j-1}\|_2 + \zeta.$$

From remark 3 we get an estimate of the number of iterations:

$$N_{stop} \geqslant 2\sqrt{\frac{LR^2}{\zeta}}.$$

$$f(x_N) - f(x^*) \leqslant \frac{4LR^2}{N^2} + N\delta_2 + R_*\delta_1 + \delta_1 \sum_{j=1}^{N} \frac{\alpha_j}{A_N} \|\tilde{x}_j - z_{j-1}\|_2 \leqslant$$

$$\leqslant N\delta_2 + R_*\delta_1 + \delta_1 \sum_{j=1}^{N} \|\tilde{x}_j - z_{j-1}\|_2 + \zeta \Rightarrow \frac{4LR^2}{N^2} \leqslant \zeta \Leftrightarrow N^2 \geqslant \frac{4LR^2}{\zeta}.$$

Summing up, we obtain the following theorem:

Theorem 5 *For model $\tau = 1$ with $\mu = 0$, using stopping rule:*

$$f(x_N) - f(x^*) \leqslant N\delta_2 + R_*\delta_1 + \delta_1 \sum_{j=1}^{N} \frac{\alpha_j}{A_N} \|\tilde{x}_j - z_{j-1}\|_2 + \zeta.$$

We can guarantee, that:

$$\tilde{R} \leqslant R.$$

And the criterion is reached after:

$$N_{stop} = \left[2\sqrt{\frac{LR^2}{\zeta}} \right] + 1.$$

6.2 **Relative noise.**

Recall the relative noise model:

$$(\forall x \in Q)\ \|\widetilde{\nabla} f(x) - \nabla f(x)\| \leqslant \alpha \|\nabla f(x)\|_2$$

We will show how such noise affects the convergence of the accelerated method. We introduce another algorithm to show convergence fast methods in presence of relative noise, which was simplified from [49].

Algorithm 2 Fast gradient method $(L, \mu, \tau, x_{start}), \quad Q \subseteq \mathbb{R}^n$

Input: Starting point x_{start}, number of steps N
Set $y_0 = u_0 = x_0 = x_{start}$,
Set $A_0 = \frac{1}{L}$, $\alpha_0 = A_0$.
for $k = 1 \dots N$ **do**
$\alpha_k = \frac{1+\mu_\tau A_{k-1}}{2L} + \sqrt{\frac{(1+\mu_\tau A_{k-1})^2}{4L^2} + \frac{1+\mu_\tau A_{k-1}}{L}}$,
$A_k = A_{k-1} + \alpha_k$,
$y_k = \frac{A_{k-1}x_{k-1} + \alpha_k u_{k-1}}{A_k}$,
$\phi_k(x) = \alpha_k \langle \tilde{\nabla} f(y_k), x - y_k \rangle + \frac{1+\mu A_{k-1}}{2}\|u_{k-1} - x\|_2^2 + \frac{\mu\alpha_k}{2}\|y_k - x\|_2^2$,
$u_k = \operatorname{argmin}_{u \in Q} \phi_k(u)$,
$x_k = \frac{A_{k-1}x_{k-1} + \alpha_k u_k}{A_k}$.
end for
Output: x_N.

If $Q = \mathbb{R}^n$ we simplify algorithm:

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

According to the analysis in [49] we get the following convergence

$$f(x_N) - f(x^*) \leqslant \frac{R^2}{2A_N} + \sum_{k=1}^{N} \frac{3\alpha^2 A_k \|\nabla f(y_k)\|_2^2}{2\mu A_N},$$

$$\|u_N - x^*\|_2^2 \leqslant \frac{1}{\mu}\left[\frac{R^2}{A_N} + \sum_{k=1}^{N} \frac{3\alpha^2 A_k \|\nabla f(y_k)\|_2^2}{2\mu A_N}\right].$$

Using convexity of f and definition of sequence y_k we get:

$$f(y_{k+1}) - f(x^*) \leqslant \frac{\alpha_{k+1}}{A_{k+1}}\left[f(u_k) - f(x^*)\right] + \frac{A_k}{A_{k+1}}\left[f(x_k) - f(x^*)\right].$$

Then we will assume unconditional optimization, that is

$$\nabla f(x^*) = 0,$$

$$(\forall x \in Q)\ f(x) - f(x^*) \leqslant \frac{L}{4}\|x - x^*\|_2^2.$$

Then using convergence for u_k:

$$f(u_k) - f(x^*) \leqslant \frac{L}{4}\|u_k - x^*\|_2^2 \leqslant \frac{L}{4\mu}\left[\frac{R^2}{A_N} + \sum_{k=1}^{N}\frac{3\alpha^2 A_k\|\nabla f(y_k)\|_2^2}{2\mu A_N}\right].$$

Recall that we denote:

$$L = 2L_f.$$

Where L_f – Lipschitz constant of ∇f. From definition of sequence α_k:

$$\alpha_k = \frac{1+\mu_\tau A_{k-1}}{2L} + \sqrt{\frac{(1+\mu_\tau A_{k-1})^2}{4L^2} + \frac{1+\mu_\tau A_{k-1}}{L}},$$
$$\frac{L}{2\mu}\alpha_{k+1} \leqslant 4A_k.$$

we obtain:

$$f(y_{N+1}) - f(x^*) \leqslant 5\frac{A_N}{A_{N+1}}\left[\frac{R^2}{2A_N} + \sum_{k=1}^{N}\frac{3\alpha^2 A_k\|\nabla f(y_k)\|_2^2}{2\mu A_N}\right],$$
$$f(y_{N+1}) - f(x^*) \leqslant \frac{5R^2}{2A_{N+1}} + \sum_{k=1}^{N}\frac{15\alpha^2 A_k\|\nabla f(y_k)\|_2^2}{2\mu A_{N+1}}.$$

Using inequality:

$$\|\nabla f(x)\|_2^2 \leqslant L\left(f(x) - f(x^*)\right).$$

We rewrite convergence bound:

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

We define following values:

$$\lambda = \frac{5R^2}{2},$$
$$\theta = \frac{15L\alpha^2}{2\mu},$$
$$\Delta_k = f(y_k) - f(x^*).$$

Finally we obtain:

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

We add one element in sum to simplify proof. In these designations by induction we can obtain:

Claim 8

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

Proof Base, $k = 1$ is obvious. Induction step:

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

That is we can formulate the following inequality:

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

Using corollary 1 we can estimate:

$$A_k \geqslant \left(1 + \sqrt{\frac{\mu}{2L}} \right)^k A_0. \tag{26}$$

We will choose an alpha such that:

$$\frac{1+\theta}{1 + \sqrt{\frac{\mu}{2L}}} \leqslant \frac{1}{1 + \frac{1}{2\sqrt{2}} \sqrt{\frac{\mu}{L}}}.$$

Using (26) and definition of θ we obtain, that if we choose α from:

$$\begin{aligned}
\alpha &\leqslant \frac{1}{7} \left(\frac{\mu}{L} \right)^{\frac{3}{4}} \\
\alpha &= \Theta \left(\left(\frac{\mu}{L} \right)^{\frac{3}{4}} \right)
\end{aligned} \tag{27}$$

From simple inequality:

$$1 + \frac{1}{2\sqrt{2}} \sqrt{\frac{\mu}{L}} > \exp \left(\frac{1}{4\sqrt{2}} \sqrt{\frac{\mu}{L}} \right).$$

We get the following theorem:

Theorem 6 *If in the model described in* (3) *in the strongly convex case we can chose* α *according to* (27) *we obtain:*

$$f(y_k) - f(x^*) \leqslant \left(\frac{5LR^2}{2} + \frac{15L\alpha^2}{2\mu} \left(f(y_0) - f(x^*) \right) \right) \exp \left(-\frac{1}{5\sqrt{2}} \sqrt{\frac{\mu}{L}} \right).$$

Corollary 2 *Under the conditions of the theorem, we obtain convergence in the argument:*

$$\|x_k - x^*\|_2^2 \leqslant R^2 \left(\frac{5L}{\mu} + \frac{15L^2\alpha^2}{4\mu^2}\right) \exp\left(-\frac{1}{4\sqrt{2}}\sqrt{\frac{\mu}{L}}\right).$$

Proof This is a direct consequence of the inequalities:

$$f(x_k) - f(x^*) \leqslant \frac{L}{4}\|x_k - x^*\|_2^2,$$
$$f(x_k) - f(x^*) \geqslant \frac{\mu}{2}\|x_k - x^*\|_2^2.$$

7 Conclusions and observations

Remark 6

Using Theorem 3 and assume, that Q – compact set we can we can denote R as $\operatorname{diam}(Q)$ instead of $\|x_0 - x^*\|_2$, then we can also bound $\tilde{R} \leqslant R$ and this will simplify bounds in theorem 4.

Remark 7

With the same assumption $\mu \neq 0$ we obtain a comparison of the two convergences in the Theorem 4. Recall that:

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

So if

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

Then the accumulation of noise in the model corresponding to $\tau = 2$, that described in (3) is less than in model $\tau = 1$, described in (2).

Remark 8

If we use model $\tau = 2$, described in theorem 4 one can set the desired accuracy of the solution.

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

Then we get from theorem 4 that:

$$f(x_N) - f(x^*) \leqslant LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2L}}N\right) + \left(1 + \sqrt{\frac{L}{\mu_2}}\right)\delta_2 + \left(1 + \sqrt{\frac{L}{\mu_2}}\right)\delta_3,$$

$$f(x_N) - f(x^*) \leqslant LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2L}}N\right) + \left(\frac{L+\mu}{\sqrt{\mu^3 L}}(\sqrt{2}+1)\right)\delta^2.$$

That is we can get estimates for δ value and number of steps N:

$$\left(\frac{L+\mu}{\sqrt{\mu^3 L}}(\sqrt{2}+1)\right)\delta^2 \leqslant \frac{\varepsilon}{2},$$
$$\delta \leqslant \sqrt{\varepsilon}\sqrt{\frac{\sqrt{2}+1}{2}}\sqrt{\frac{L+\mu}{\sqrt{\mu^3 L}}},$$
$$\delta = O\left(\sqrt{\varepsilon}\frac{(L+\mu)^{\frac{1}{2}}}{(\mu^3 L)^{\frac{1}{4}}}\right);$$
$$LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2L}}N\right) \leqslant \frac{\varepsilon}{2},$$
$$N \geqslant 2\sqrt{\frac{2L}{\mu}}\left(\ln 2LR^2 + \ln \varepsilon^{-1}\right),$$
$$N = O\left(\sqrt{\frac{L}{\mu}}\ln\frac{LR^2}{\varepsilon}\right).$$

Remark 9

Using remark 4 and previous remark 8, we can found similar bounds. Remind that:

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

However we should value of the parameter μ. We will let:

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

Using inequality:

$$\delta^2\left(1+\sqrt{\frac{2L+4}{\mu}}\right)\left(\frac{\mu+L}{\mu L}\right) \leqslant \frac{\varepsilon}{3}.$$

And the selected value of the parameter mu we get required value of error δ:

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

Similarly, get an estimate of the number of steps:

$$LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2(L+2)}}N\right) \leqslant \frac{\varepsilon}{3},$$
$$N \geqslant \sqrt{12L+24}R \ln 2LR^2 + 2\sqrt{2L+4}\frac{1}{\sqrt{\varepsilon}} \ln \frac{1}{\varepsilon},$$
$$N = O\left(\sqrt{\frac{LR^2}{\varepsilon}} \ln \frac{LR^2}{\varepsilon}\right).$$

Remark 10

Using remark 5 and theorem 5 we can apply it to problem:

$$Ax = b,$$
$$A \in \mathrm{GL}_n(\mathbb{R}).$$

Solving such a problem is equivalent to solving the convex optimization problem:

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 \to \min,$$
$$\nabla f(x) = A^T (Ax - b).$$

We will assume similarly the estimate of the norm x^*:

$$\|x^*\|_2 \leqslant R_*.$$

Let the original problem be solved with an ε_1 accuracy in the sense:

$$\|Ax - b\|_2 \leqslant \varepsilon_1,$$
$$f(x) - f(x^*) = \frac{1}{2}\|Ax - b\|_2^2 \leqslant \varepsilon,$$
$$\varepsilon = \frac{1}{2}\varepsilon_1^2.$$

When the algorithm stops, we get the convergence:

$$f(x_{N_{stop}}) - f(x^*) \leqslant N\delta_2 + 3\delta_1 R_*,$$
$$N_{stop} = \left[2\sqrt{\frac{LR^2}{\zeta}}\right] + 1.$$

Then we choose δ, ζ from the following conditions:

$$\begin{cases} \zeta \leqslant \frac{\varepsilon}{3}, \\ \delta \leqslant \left(\frac{L^{\frac{1}{4}}}{6\sqrt{3R}}\right)\varepsilon^{\frac{3}{4}}, \\ \delta \leqslant \varepsilon\frac{1}{9R_*}. \end{cases}$$

For example, we can let:

$$\delta = \mathrm{C}_{R,R_*,L}\varepsilon,$$
$$\mathrm{C}_{R,R_*,L} = \min\left\{\frac{L^{\frac{1}{4}}}{6\sqrt{3R}}, \frac{1}{9R_*}\right\}.$$

Then the number of steps required is expressed as:

$$N_\varepsilon = \left[2\sqrt{\frac{3LR^2}{\varepsilon}}\right] + 1.$$

Accordingly, the estimate required for solving the problem of linear equations:

$$N_{\varepsilon_1} = \left[2\frac{\sqrt{3LR^2}}{\varepsilon_1}\right] + 1.$$

Remark 11

The work considered a model of additive noise in equation (20), similar to [42], that is we can consider that:

$$\tilde{\nabla} f(x) = \nabla f(x) + r_x,$$
$$\|r_x\|_2 \leqslant \delta.$$

Similarly to this work, a stopping criterion was proposed for the STM algorithm, as was proposed for gradient descent.

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k).$$

Note that in the same noise model, the convergence estimate in both considered cases will be:

$$j_N = \underset{1\leqslant k\leqslant N}{\operatorname{argmin}} f\left(x_k\right),$$
$$y_N = x_{j_N},$$
$$f(y_N) - f(x^*) = O\left(\frac{LR^2}{N} + \frac{\delta^2}{L} + \tilde{R}\delta\right),$$
$$f(y_N) - f(x^*) = O\left(LR^2\exp\left(-\frac{\mu}{L}N\right) + \frac{\delta^2}{L} + \tilde{R}\delta\right).$$
$$f(y_N) - f(x^*) = O\left(LR^2\exp\left(-\frac{\mu}{2L}\right) + \frac{\delta^2}{L} + \frac{\delta^2}{\mu}\right),$$
$$\tilde{R} = \max_{k\leqslant N}\|x_k - x^*\|_2.$$

Despite the fact that in the work [17], a slightly different model was considered, namely (δ, L) and (δ, L, μ) oracle (equation 3.1 Definition 1 in [17]) similar orders of convergence were obtained, that is theorem 4 and relevant remark

3. Namely, function satisfies the (δ, L, μ) model at point $x \in Q$ means, that exists functions $f_\delta(x)$ and $\psi_\delta(x, y)$, such that:

$$\forall y \in Q$$
$$\frac{\mu}{2}\|x-y\|_2^2 \leqslant f(x) - f_\delta(y) - \psi_\delta(x,y) \leqslant \frac{L}{2}\|x-y\|_2^2 + \delta.$$

Similarly to papers [49], [17], the results also hold in the case of an unbounded set Q(result in [49] is on the page 26, obtained for fast adaptive gradient method page 13). Stopping criteria are also formulated, which give an estimate on $\tilde{R}$ for a non-compact Q, remind that:

$$\tilde{R} = \max_{0 \leqslant k \leqslant N}\{\|x_k - x^*\|_2, \|\tilde{x}_k - x^*\|_2, \|z_k - x^*\|_2\}.$$

We also note that a similar models of (δ, Δ, L) and (δ, Δ, L, μ) oracle was considered in the work [48]. Moreover, the function satisfies (δ, Δ, L, μ)–model

$$f(y) \leqslant f_\delta + \psi(y,x) + \Delta\|x-y\| + \delta + LV(y,x),$$
$$f_\delta + \psi(x^*, x) + \mu V(y,x) \leqslant f(x^*),$$
$$f(x) - \delta \leqslant f_\delta(x) \leqslant f(x),$$
$$\psi(x,x) = 0.$$

Here $V(x, y)$ – Bregman divergence. At the same time, an adaptive analogue of STM was considered. As well as similar estimations for a δ and number of steps N, following [17] (page 24, remarks 11 – 14), namely there are remarks 8, 9, 10. Also considered an example of using regularization to obtain convergence in the model $\tau = 1$, for the case $\mu = 0$.

Remark 12

Similarly, accelerated methods in the Euclidean prox-structure in the presence of relative inaccuracy were considered in the work [25]. For the triple momentum method, the following bound was obtained,

$$\chi = \frac{L}{\mu},$$
$$\alpha < \frac{\sqrt{\chi}+1}{4\chi - 3\sqrt{\chi}+1},$$
$$\alpha = O\left(\sqrt{\frac{\mu}{L}}\right)$$

However, as experiments show, STM is more stable to the relative noise model, that is, calculating the dependence of the largest alpha α^* for the given parameters of the problem μ, L, we get bigger upper bound. More detailed information in Section 8.

8 Numerical experiments

For testing STM for degenerate problems, the function described in [38] on page 69, that is:

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

These two plots reflect the convergence of the method at the first 50 000 and 10 000 iterations, respectively, at different δ.

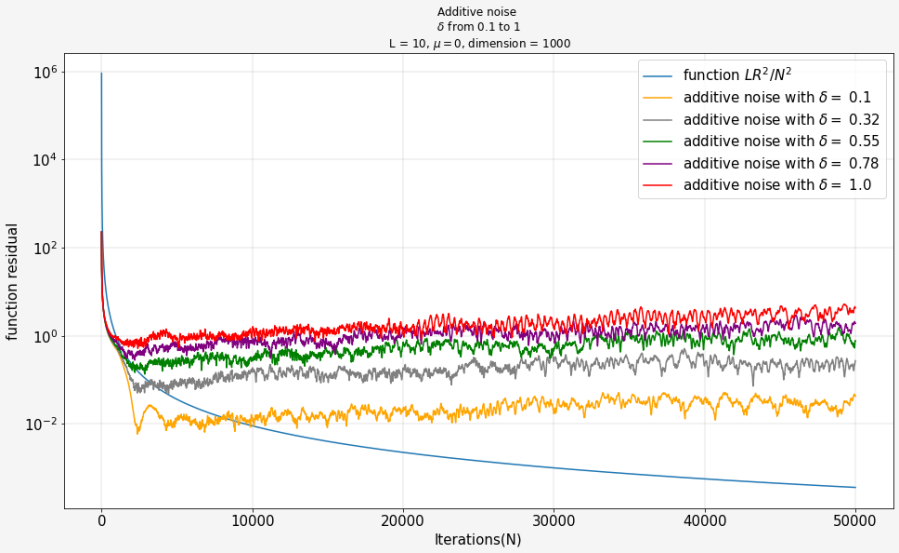

Fig. 2 First test – first 50 000 steps.

Fig. 3 First test – first 10 000 steps.

Let's also consider a drawing with two types of noise.

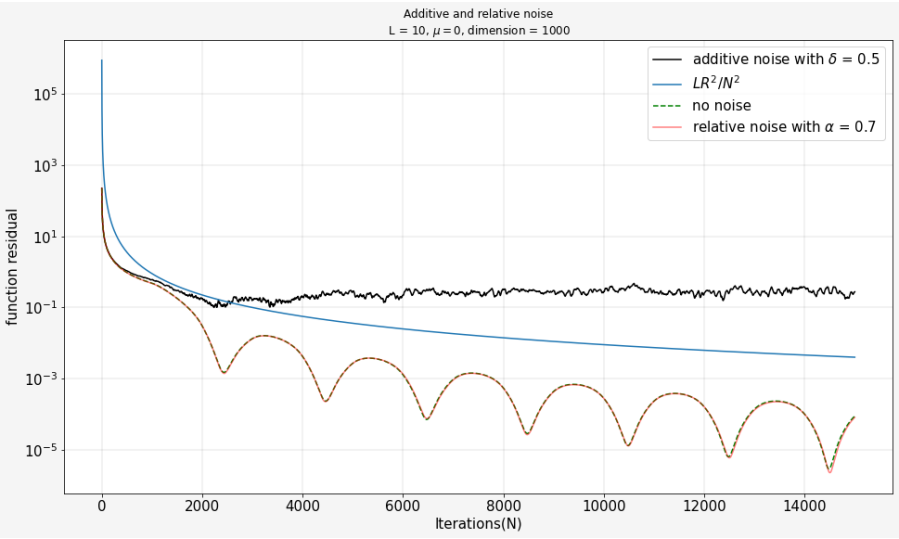

Fig. 4 Second test – relative and additive types of noises comparison.

To compare the convergence of a degenerate problem with different α parameters in the case of relative noise, consider the following graph.

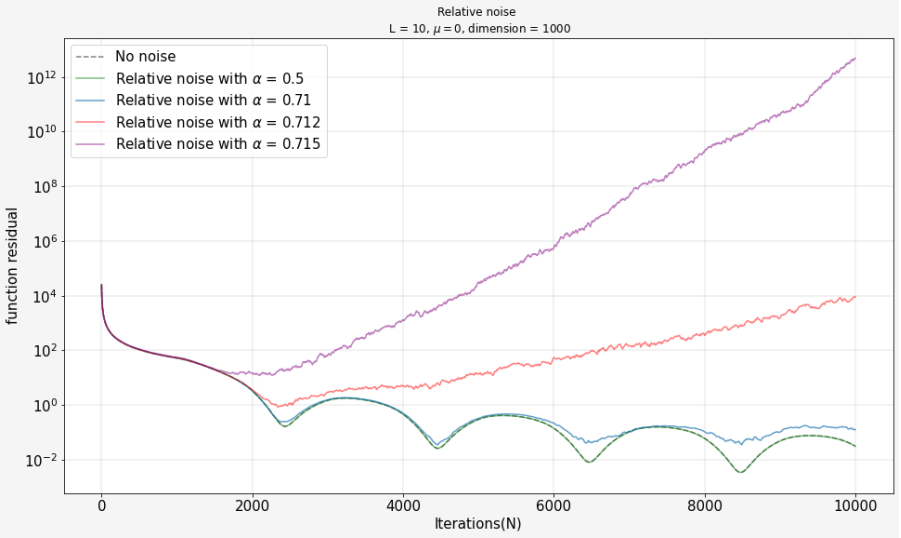

Fig. 5 Third test – relative noise with different values of α for $\mu = 0$.

The last figure shows that for $\alpha \leqslant 0.71$ the convergence of the method does not deteriorate, but we can assume the existence of such a threshold value $\alpha^* \approx 0.71$, that at large of α values the method diverges.

Also for testing on strongly convex functions, an analogue of the finite-dimensional Nesterov function was used from [38] on page 78, that is:

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

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

$$\nabla f(x) = \left(\frac{\mu\,(\chi - 1)}{4}A + \mu E\right)x - \frac{\mu\,(\chi - 1)}{4}e_1,$$

$$e_1 = (1,\, 0,\, \ldots\,, 0\,)^T,$$

where E – identity operator, A is the matrix defined as:

$$\begin{pmatrix} 2 & -1 & 0 & \ldots & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & -1 & 2 & -1 \\ 0 & \ldots & \ldots & 0 & -1 & 2 \end{pmatrix}.$$

Then minimum f, x^*, can be found from systems of linear equations. Let us consider the graphs of the residuals for different parameters of the delta additive noise.

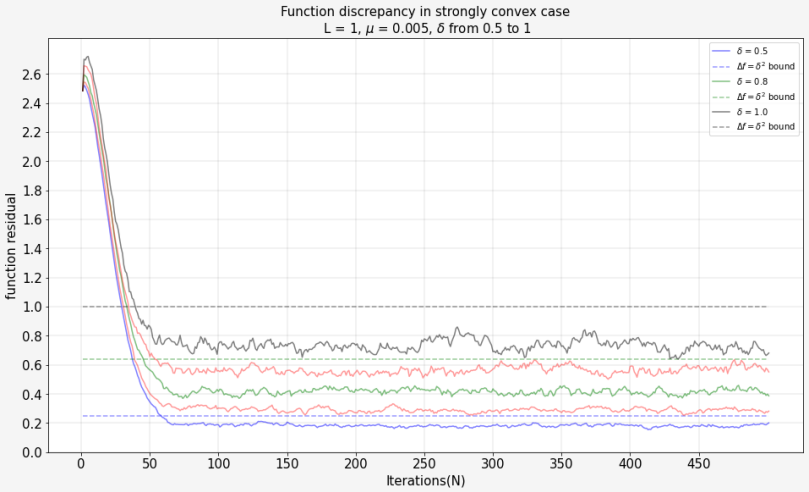

Fig. 6 Fourth test – $\delta \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

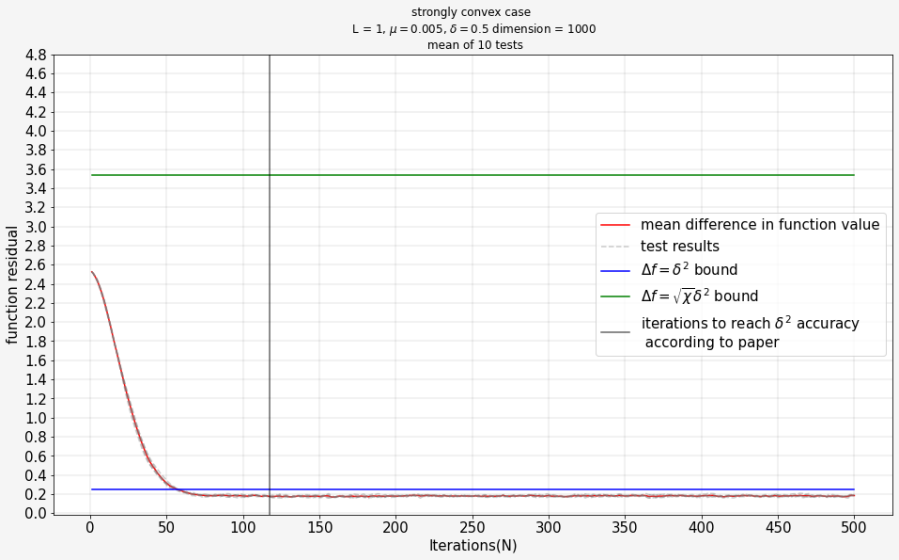

Fig. 7 Fifth test – mean of 30 tests, level of approximation and required number of steps.

The last plot confirms theorem 4 and remark 8. Similarly to the degenerate case, consider the behavior of the method for different parameters α.

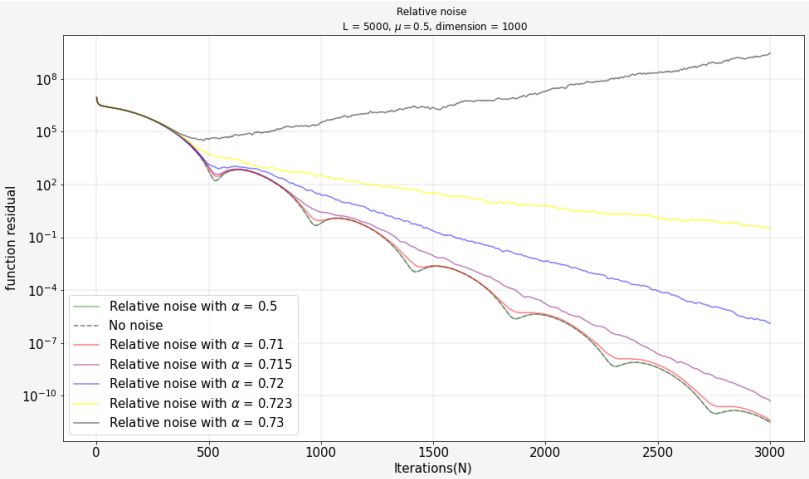

Fig. 8 Sixth test – relative noise with different values of α for $\mu > 0$.

Note that in the strongly convex case, we obtain a property similar to the degenerate case: for α values less than a certain threshold value α^*.

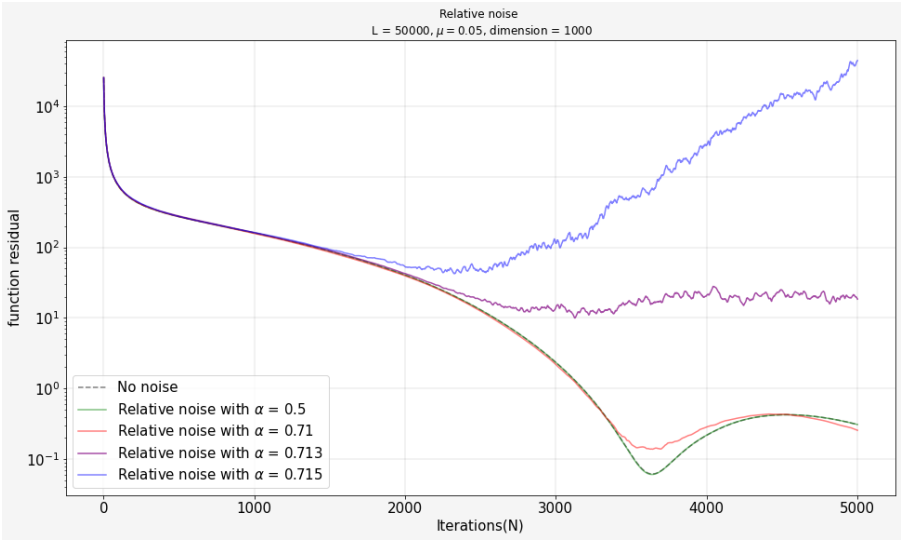

Fig. 9 Seventh test – relative noise with different values of α for other L and μ.

Comparing STM and triple momentum method we get the following plots:

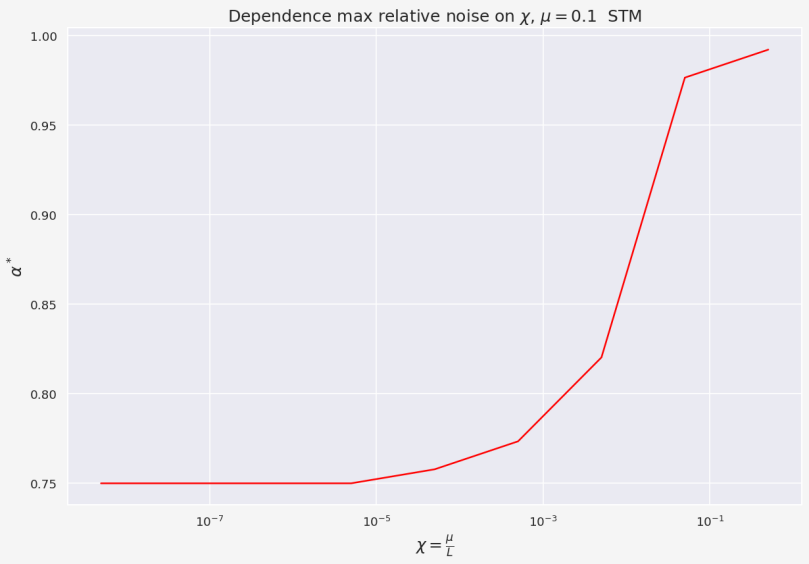

Fig. 10 Eigth test – threshold α^* for different L and $\mu = 0.1$, for STM algo

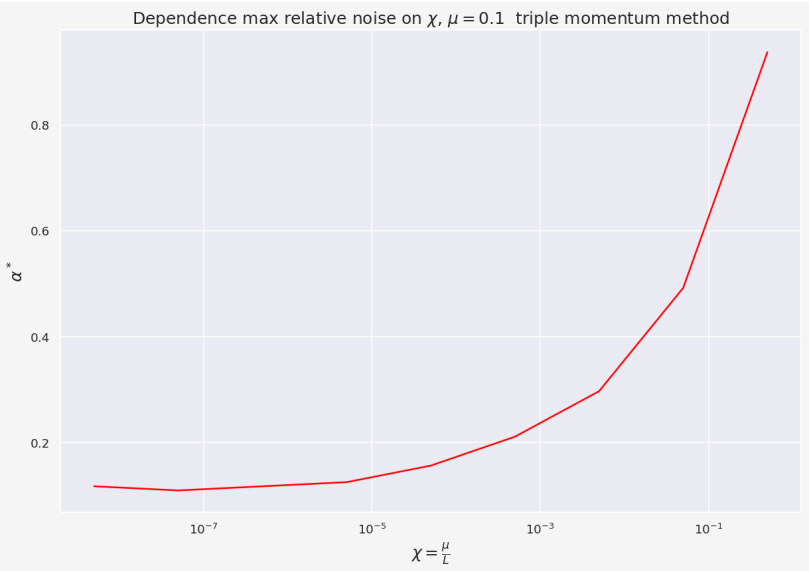

Fig. 11 Nineth test – threshold α^* for different L and $\mu = 0.1$, for tripple momentum algo

9 Acknowledgment

The authors are grateful to Eduard Gorbunov for useful discussions.

References

1. Ajalloeian, A., Stich, S.U.: Analysis of sgd with biased gradient estimators. arXiv preprint arXiv:2008.00051 (2020)
2. Akhavan, A., Pontil, M., Tsybakov, A.B.: Exploiting higher order smoothness in derivative-free optimization and continuous bandits. arXiv preprint arXiv:2006.07862 (2020)
3. Bach, F., Perchet, V.: Highly-smooth zero-th order online optimization. In: V. Feldman, A. Rakhlin, O. Shamir (eds.) 29th Annual Conference on Learning Theory, *Proceedings of Machine Learning Research*, vol. 49, pp. 257–283. PMLR, Columbia University, New York, New York, USA (2016). URL `http://proceedings.mlr.press/v49/bach16.html`
4. Beck, A.: First-order methods in optimization. SIAM (2017)
5. Belloni, A., Liang, T., Narayanan, H., Rakhlin, A.: Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In: P. Grünwald, E. Hazan, S. Kale (eds.) Proceedings of The 28th Conference on Learning Theory, *Proceedings of Machine Learning Research*, vol. 40, pp. 240–265. PMLR, Paris, France (2015). URL `http://proceedings.mlr.press/v40/Belloni15.html`
6. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization (Lecture Notes). Personal web-page of A. Nemirovski (2015)
7. Berahas, A.S., Cao, L., Choromanski, K., Scheinberg, K.: A theoretical and empirical comparison of gradient approximations in derivative-free optimization. arXiv preprint arXiv:1905.01332 (2019)
8. Beznosikov, A., Sadiev, A., Gasnikov, A.: Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem. In: International Conference on Mathematical Optimization Theory and Operations Research, pp. 105–119. Springer (2020)
9. Bubeck, S.: Convex optimization: Algorithms and complexity. arXiv preprint arXiv:1405.4980 (2014)
10. Cohen, M.B., Diakonikolas, J., Orecchia, L.: On acceleration with noise-corrupted gradients. arXiv preprint arXiv:1805.12591 (2018)
11. Conn, A., Scheinberg, K., Vicente, L.: Introduction to Derivative-Free Optimization. Society for Industrial and Applied Mathematics (2009). DOI 10.1137/1.9780898718768. URL `http://epubs.siam.org/doi/abs/10.1137/1.9780898718768`
12. d'Aspremont, A.: Smooth optimization with approximate gradient. SIAM Journal on Optimization **19**(3), 1171–1183 (2008)
13. d'Aspremont, A., Scieur, D., Taylor, A.: Acceleration methods. arXiv preprint arXiv:2001.09545 (2021)
14. Devolder, O.: Stochastic first order methods in smooth convex optimization. CORE Discussion Paper 2011/70 (2011)
15. Devolder, O.: Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. Ph.D. thesis, ICTEAM and CORE, Université Catholique de Louvain (2013)
16. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Mathematical Programming **146**(1), 37–75 (2014). DOI 10.1007/s10107-013-0677-5. URL `http://dx.doi.org/10.1007/s10107-013-0677-5`
17. Devolder, O., Glineur, F., Nesterov, Y., et al.: First-order methods with inexact oracle: the strongly convex case. CORE Discussion Papers **2013016**, 47 (2013)
18. Drusvyatskiy, D., Xiao, L.: Stochastic optimization with decision-dependent distributions. arXiv preprint arXiv:2011.11173 (2020)
19. Dvinskikh, D., Gasnikov, A.: Decentralized and parallelized primal and dual accelerated methods for stochastic convex programming problems. Journal of Inverse and Ill-posed Problems (2021)

20. Dvinskikh, D.M., Turin, A.I., Gasnikov, A.V., Omelchenko, S.S.: Accelerated and non accelerated stochastic gradient descent in model generality. Matematicheskie Zametki **108**(4), 515–528 (2020)
21. Dvurechensky, P.: Numerical methods in large-scale optimization: inexact oracle and primal-dual analysis. HSE. Habilitation (2020)
22. Dvurechensky, P., Gasnikov, A.: Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. Journal of Optimization Theory and Applications **171**(1), 121–145 (2016). DOI 10.1007/s10957-016-0999-6. URL `http://dx.doi.org/10.1007/s10957-016-0999-6`
23. Dvurechensky, P., Staudigl, M., Shtern, S.: First-order methods for convex optimization. arXiv preprint arXiv:2101.00935 (2021)
24. Evtushenko, Y.G.: Optimization and fast automatic differentiation. Computing Center of RAS, Moscow (2013)
25. Gannot, O.: A frequency-domain analysis of inexact gradient methods. Mathematical Programming pp. 1–42 (2021)
26. Gasnikov, A.: Universal gradient descent. arXiv preprint arXiv:1711.00394 (2017)
27. Gasnikov, A., Kabanikhin, S., Mohammed, A., Shishlenin, M.: Convex optimization in hilbert space with applications to inverse problems. arXiv preprint arXiv:1703.00267 (2017)
28. Gasnikov, A.V., Gasnikova, E.V., Nesterov, Y.E., Chernov, A.V.: Efficient numerical methods for entropy-linear programming problems. Computational Mathematics and Mathematical Physics **56**(4), 514–524 (2016). DOI 10.1134/S0965542516040084. URL `http://dx.doi.org/10.1134/S0965542516040084`
29. Gasnikov, A.V., Nesterov, Y.E.: Universal method for stochastic composite optimization problems. Computational Mathematics and Mathematical Physics **58**(1), 48–64 (2018)
30. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT press Cambridge (2016)
31. Gorbunov, E., Dvinskikh, D., Gasnikov, A.: Optimal decentralized distributed algorithms for stochastic convex optimization. arXiv preprint arXiv:1911.07363 (2019)
32. Kabanikhin, S.I.: Inverse and ill-posed problems: theory and applications, vol. 55. Walter De Gruyter (2011)
33. Kamzolov, D., Dvurechensky, P., Gasnikov, A.V.: Universal intermediate gradient method for convex problems with inexact oracle. Optimization Methods and Software pp. 1–28 (2020)
34. Kotsalis, G., Lan, G., Li, T.: Simple and optimal methods for stochastic variational inequalities, ii: Markovian noise and policy evaluation in reinforcement learning. arXiv preprint arXiv:2011.08434 (2020)
35. Lan, G.: First-order and Stochastic Optimization Methods for Machine Learning. Springer (2020)
36. Nemirovski, A.S.: Regularizing properties of the conjugate gradient method for ill-posed problems. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki **26**(3), 332–347 (1986)
37. Nemirovsky, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. J. Wiley & Sons, New York (1983)
38. Nesterov, Y.: Lectures on convex optimization, vol. 137. Springer (2018)
39. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. **17**(2), 527–566 (2017). DOI 10.1007/s10208-015-9296-2. URL `https://doi.org/10.1007/s10208-015-9296-2`. First appeared in 2011 as CORE discussion paper 2011/16
40. Novitskii, V., Gasnikov, A.: Improved exploiting higher order smoothness in derivative-free optimization and continuous bandit. arXiv preprint arXiv:2101.03821 (2021)
41. Pedregosa, F., Scieur, D.: Average-case acceleration through spectral density estimation. arXiv preprint arXiv:2002.04756 (2020)
42. Poljak, B.: Iterative algorithms for singular minimization problems. In: Nonlinear Programming 4, pp. 147–166. Elsevier (1981)
43. Polyak, B.: Introduction to Optimization. New York, Optimization Software (1987)
44. Polyak, B.T., Tsybakov, A.B.: Optimal order of accuracy of search algorithms in stochastic optimization. Problemy Peredachi Informatsii **26**(2), 45–53 (1990)

45. Risteski, A., Li, Y.: Algorithms and matching lower bounds for approximately-convex optimization. Advances in Neural Information Processing Systems **29**, 4745–4753 (2016)
46. Rockafellar, R.T.: Convex analysis, vol. 36. Princeton university press (1970)
47. Scieur, D., Pedregosa, F.: Universal asymptotic optimality of polyak momentum. In: International Conference on Machine Learning, pp. 8565–8572. PMLR (2020)
48. Stonyakin, F.: Adaptive methods for variational inequalities, minimization problems and functional with generalized growth condition. MIPT. Habilitation (2020)
49. Stonyakin, F., Tyurin, A., Gasnikov, A., Dvurechensky, P., Agafonov, A., Dvinskikh, D., Pasechnyuk, D., Artamonov, S., Piskunova, V.: Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model. arXiv preprint arXiv:2001.09013 (2020)
50. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Mathematical Programming **161**(1-2), 307–345 (2017)
51. Tyurin, A.: Development of a method for solving structural optimization problems. HSE. PhD Thesis (2020)
52. Vasilyev, F.: Optimization Methods. Moscow, Russia: FP (2002)