Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Stopping rules for accelerated gradient methods with additive noise in gradient

Artem Vasin[1], Alexander Gasnikov[1,2,3], Vladimir Spokoiny[2,3]

submitted: February 11, 2021

[1] Moscow Institute of Physics and Technology
Institutskiy Pereulok, 9
Dolgoprudny, Moscow Region
141701 Russian Federation
E-Mail: vasin.aa@phystech.edu

[2] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: vladimir.spokoiny@wias-berlin.de
alexander.gasnikov@wias-berlin.de

[3] Institute for Information Transmission Problems of RAS
Bolshoy Karetny per. 19, build.1
127051 Moscow
Russian Federation
E-Mail: gasnikov@yandex.ru

No. 2812
Berlin 2021

2020 *Mathematics Subject Classification.* 90C30, 90C25, 68Q25.

Key words and phrases. Accelerated methods, inexact gradient, stopping rule, inverse 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.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Stopping rules for accelerated gradient methods with additive noise in gradient

Artem Vasin, Alexander Gasnikov, Vladimir Spokoiny

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.

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, 25, 4, 26, 37, 34, 50, 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 [49, 26].

In many applications, especially for gradient-free methods (when estimating the gradient by finite differences [11, 44, 7]) optimization problems in infinite dimensional spaces (such examples arise when solving inverse problems [31, 27]) instead of an access to $\nabla f(x)$ we have an access to its inexact approximation $\widetilde{\nabla} f(x)$.

The two most popular conception of inexactness of gradient in practice are [42]: for all $x \in Q$

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

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

DOI 10.20347/WIAS.PREPRINT.2812 Berlin 2021

For the first conception (2) several results about the accumulation of error can be found in [42, 12, 10, 1], but all these results are still far from to be optimistic in general. The reason was described in [41]. 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 $0 \geq \mu = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n = L$, $L \geq 2\mu$. 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 $\partial f(x)/\partial x^1 = \mu x^1 - \delta$. For simple gradient dynamic

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

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[1]

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

So we have a problem with (5), since μ can be to small ($\mu \ll \varepsilon$ – 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, [41, 42, 35]): to propose a stopping rule for the considered algorithm or to use regularization $\mu \sim \varepsilon$ [26]. Roughly speaking, for non accelerated algorithms in [41, 42] 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 [37, 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 $(1 - \alpha)^{-1}$-times [42]. 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 $\alpha \ll \left(\frac{\mu}{L}\right)^{3/4}$ in μ-strongly convex case and (on k-th iteration) $\alpha_k \ll \left(\frac{1}{k}\right)^{3/2}$ 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 $\alpha \ll \left(\frac{\mu}{L}\right)^{5/4}$) in the case $\mu \sim \varepsilon$ were recently obtained by using another techniques in Stochastic Optimization with decision dependent distribution [18] and

[1] This bound corresponds to the worst-case philosophy concerning the choice of considered example for considered class of methods [36, 37, 9, 26]. We expect more interesting results here by considering average-case complexity [46, 40] (spectrum $\{\lambda_i\}$ average).

Policy Evaluation in Reinforcement Learning via reduction to stochastic Variational Inequality with Markovian noise [33]. In [33, 18] it was assumed that

$$k\nabla f(x) - \nabla f(x)k_2 \le Bkx - x_*k_2, \quad \alpha \in [0,1). \tag{6}$$

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

$$k\nabla f(x)k_2 = k\nabla f(x) - \nabla f(x_*)k_2 \le Lkx - x_*k_2.$$

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 $\nabla_\delta f(x)$, satisfying for all $x, y \in Q$

$$\begin{aligned} f(x) + h\nabla_\delta f(x), y - xi + \frac{\mu}{2}ky - xk_2^2 - \delta \le f(y) \\ \le f(x) + h\nabla_\delta f(x), y - xi + \frac{L}{2}ky - xk_2^2 + \delta. \end{aligned} \tag{7}$$

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

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

and for accelerated ones [16, 20]

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

where $R = kx_{start} - x_*k_2$ – 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, 32] for "indermediate" situations between accelerated and non-accelerated methods.

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

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

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

$$h\nabla f(x) - \nabla f(x), y - xi \le \frac{1}{2L}k\nabla f(x) - \nabla f(x)k_2^2 + \frac{L}{2}ky - xk_2^2,$$

$$h\nabla f(x) - \nabla f(x), y - xi \ge \frac{1}{\mu}k\nabla f(x) - \nabla f(x)k_2^2 - \frac{\mu}{4}ky - xk_2^2.$$

So, when $\mu > 0$ for non-accelerated methods this result is almost the same as we've obtained by considering example (4). To reach $f(x_k) - f(x_*) = \varepsilon$ when[2] $\mu \gtrsim \varepsilon$ 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, 47, 48] the conception (7) was further developed

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

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

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

and for accelerated ones [16, 20]

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

where R is the maximal distance between generated points and the solution.

Thus from (12), (13) we may conclude that if R is bounded,[3] 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 R is bounded". As we may see from example (4) in general degenerate regime only such bound

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

takes place [26]. This dependence spoils the result. The growth of R we observe in different experiments. In the paper below we investigate this problem. In particular, we propose an alternative approach to regularization[4] that is based on "early stopping"[5] of considered iterative procedure by developing proper stopping rule.

[2] 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 [41]. 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 [25, 26].

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

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

[5] This terminology is popular also in Machine Learning community, where "early stopping" is used also as alternative to regularization to prevent overfitting [29].

Now we explain how to reduce relative inexactness (3) to (7) and to apply (9) when μ ε. 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 $\log(LR^2/\varepsilon)$, where ε – accuracy in function value)

$$f(x_k) - f(x_*) \overset{(9),(10)}{\cdot} \frac{\varepsilon}{2} + \sqrt{\frac{L}{\mu}}\frac{\delta_{(2)}^2}{\mu} \quad , \quad \sqrt{\frac{L}{\mu}}\frac{\delta_{(2)}^2}{\mu}$$
$$\overset{(3),(7)}{\cdot} \sqrt{\frac{L}{\mu}}\frac{\alpha^2 \max_{t=1,\dots,k} \|\nabla f(x_t)\|_2^2}{\mu} \le \sqrt{\frac{L}{\mu}}\frac{2L\alpha^2 \max_{t=1,\dots,k}(f(x_k) - f(x_*))}{\mu}$$
$$\cdot \sqrt{\frac{L}{\mu}}\frac{4L\alpha^2 (f(x_0) - f(x_*))}{\mu}. \tag{14}$$

To guarantee that (restart condition)

$$f(x_k) - f(x_*) \le \frac{1}{2}(f(x_0) - f(x_*))$$

we should have $\alpha \cdot \left(\frac{1}{L}\right)^{5=4}$. Then we restart the method. After $\log(f/\varepsilon)$ 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 $\alpha_k \cdot \left(\frac{1}{k}\right)^{5=2}$. More accurate analysis in the subsequent part of the paper allows to improve these bounds:

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

Below we'll concentrate only on accelerated method and choose the method with one projection (Similar Triangles Method (STM)), see [28, 10, 30, 47, 23] and reference there in. We decided to choose this method because: 1) it's primal-dual [28]; 2) has a nice theory of how to bound $\tilde{R}$ in no noise regime [28, 37] ($\tilde{R} \le R$) and noise one [30]; 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.

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[6] $f(x)$ with accuracy δ_f [11]:

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

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)

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

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 $\mathrm{Poly}(n, \frac{1}{\varepsilon})$ [44] (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 [43, 3]. The interest to this alternative has grown last time [2, 39].

(Gaussian Smoothed Gradients)

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

where $e \sim N(0, I_n)$ 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 [38, 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!

[6] 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 R^n with preserving of convexity and Lipschitz continuity [44]: $f_{new}(x) := f\left(\mathrm{proj}_Q(x)\right) + \alpha \min_{y \in Q} \|x - y\|_2$.

(Sphere Smoothed Gradients)

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

where e is random vector with uniform distribution in a unit sphere (with center at 0) in 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 $\delta \quad (n\delta_f)^{\frac{p}{p+1}}$. Hence,

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

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, 44]: For Sphere smoothed gradient in conception (7) we have

$$\delta \ ' \ 2L_0 h + \frac{p\overline{n}\delta_f R}{h}, \tag{16}$$

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 $\varepsilon \quad \delta \quad n^{1=4}\delta_f^{1=2}$. Hence,

$$\delta_f \quad \frac{\varepsilon^2}{\overline{p}\overline{n}}.$$

The rest part ($\delta_f \quad \frac{''}{n}$) 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 [51]. Such optimization problems arise, in particular, in inverse problems theory [31].

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

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

hence

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

Therefore

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

The same result could be obtained by considering Lagrange functional

$$L(q,u;\psi) = J(q,u(q)) + h\psi, G(q,u)i$$

with

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

and

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

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 $G(q,u) = G \cdot (q,u) = 0$ 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) [31]. This is an inverse problem. We can reduce this problem to optimization one [31]:

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

We can solve (17) numerically. This problem is convex quadratic optimization problem. We can directly apply Lagrange multipliers principle to (17), see [51]:

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

To obtain 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

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

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 [45]

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

from the Demyanov–Danskin's formula [45][7]

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

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

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

[7]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.

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

$$\min_{u2Q}\max_{2(D)} L(q,u;\psi)$$

saddle-point problem. Since $\delta\ L(q,u;\psi) = 0$ entails $G\cdot(q,u) = 0$ that is form (P) if we add $u\ 2\ Q$ and $\delta_u L_(q,u;\psi) = 0$, when $u\ 2\ Q$ entails (D) as we've shown above.

Note also that

$$L_q(q,u(q);\psi(q))(y) = \lambda(y),\quad y\ 2\ (0,1)\,.$$

Hence, due to (19)

$$r\ \mathtt{J}(q)(y) = \psi_x(1,y),\quad y\ 2\ (0,1)$$

So we reduce $r\ \mathtt{J}(q)(y)$ calculation to the solution of two correct initial-boundary value problem for elliptic equation in a square (P) and (D) [31].

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

$$A:\quad q(y) := u(1,y)\ \overline{?}\ \ u(0,y).$$

Here $u(x,y)$ is the solution of problem (P). It was shown in [31] that

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

Conjugate operator is [31]

$$A^*:\quad p(y) := \psi_x(0,y)\ \overline{?}\ \ \psi_x(1,y),\qquad A^*: L_2(0,1)\ \ !\ \ L_2(0,1).$$

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

$$\mathtt{J}(q)(y) = kAq - bk_2^2,$$

we can write

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

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* $p(y) = 2\,(u(0,y) - b(y))$.

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

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

The described above technique can be applied to many different inverse problems [31] and optimal control problems [51]. 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 R^n:

$$\min_{x2Q} f(x).$$

Assume that

$$k\tilde{r} f(x) - r f(x)k_2 \ 6 \ \delta, \tag{20}$$

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

$$R = kx_{start} - x^* k_2$$

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:

$$8x, y \ 2 \ Q, \ k r f(x) - r f(y)k_2 \ 6 \ L_f kx - yk_2. \tag{21}$$

This implies inequality:

$$8x, y \ 2 \ Q, \ f(y) \ 6 \ f(x) + h r f(x), y - xi + \frac{L_f}{2} kx - yk_2^2. \tag{22}$$

We will use following lemma:

Lemma 4.1 (Fenchel inequality)**.** *Let* $(E, h\cdot, \cdot i)$ *– euclidean space, then* $8\lambda \ 2 \ \mathrm{R}_+, 8u, v \ 2 \ E$ *the inequality holds:*

$$hu, vi \ 6 \ \frac{1}{2\lambda} kuk_E^2 + \frac{\lambda}{2} kvk_E^2.$$

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

Claim 1. $8x, y \ 2 \ Q$, *the following estimate holds:*

$$f(y) \ 6 \ f(x) + h\tilde{r} f(x), y - xi + \frac{L}{2} kx - yk_2^2 + \delta_2,$$

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

Proof. The proof follows from

$$\begin{aligned} f(y) \ 6 \ f(x) + h r f(x), y - xi + \frac{L_f}{2} kx - yk_2^2 \ 6 \\ 6 \ f(x) + h\tilde{r} f(x), y - xi + \frac{1}{2L_f} k r f(x) - \tilde{r} f(x)k_2^2 + \frac{L_f}{2} kx - yk_2^2 + \frac{L_f}{2} kx - yk_2^2 \ 6 \\ 6 \ f(x) + h\tilde{r} f(x), y - xi + \frac{L}{2} kx - yk_2^2 + \delta_2. \end{aligned}$$

□

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:

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

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

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

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

where $\delta_1 = \delta$.

Proof. Using Cauchy inequality and (23) we obtain:

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

□

Claim 3. *$\forall x, y \in Q$, if in* (23) *$\mu \geq 0$, 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. 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:

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

□

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 τ, 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 $STM(L, \mu, \tau, x_{start}), \quad Q \quad \mathrm{R}^n$

Input: Starting point x_{start}, number of steps N

Output: x_N

Set $x_0 = x_{start}$

Set $A_0 = \frac{1}{L}$, $\alpha_0 = \frac{1}{L}$,

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

Set $z_0 = \operatorname{argmin}_{y \in Q} \psi_0(y)$,

Set $x_0 = z_0$.

for $k = 1, 2 \ldots N$ **do**

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

  $A_k = A_{k-1} + \alpha_k$,

  $x_k = \frac{A_{k-1}x_{k-1} + \tau \alpha_k z_{k-1}}{A_k}$,

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

  $z_k = \operatorname{argmin}_{y \in Q} \psi_k(y)$,

  $x_k = \frac{A_{k-1}x_{k-1} + \tau \alpha_k z_k}{A_k}$.

end for

return 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 $Q = \mathrm{R}^n$, we can simplify the step of the algorithm by replacing it with:

$$z_k = z_{k-1} - \frac{\alpha_k}{1 + A_k \mu}\left(\nabla f(x_k) + \mu\ (z_{k-1} - x_k)\right).$$

We define constant:

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

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

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

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, x_k, z_k. Also very important are the estimates for the sequence A_k.

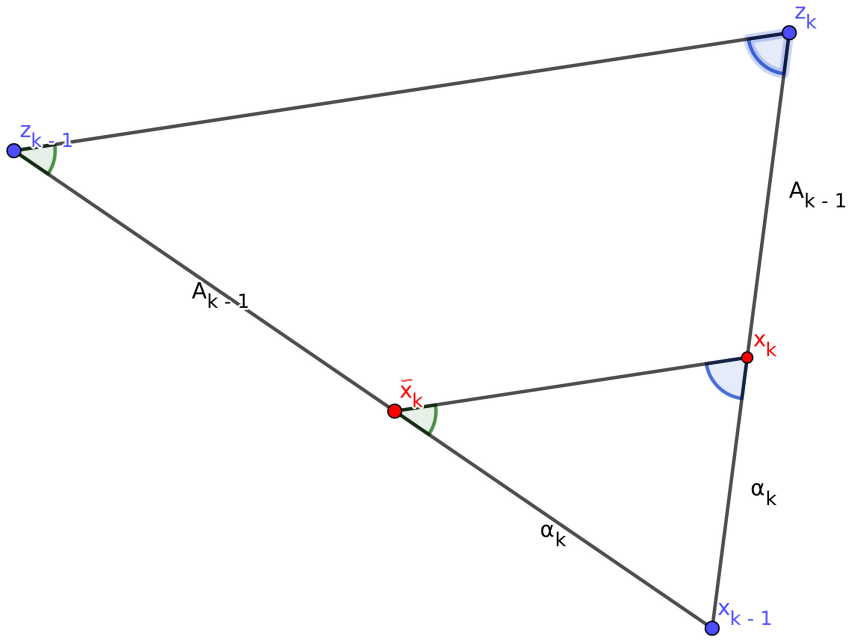

Figure 1: Geometry of Similar Triangles method [28]

Claim 4. *If* $\mu > 0$ *and* $8k \geq 2$ $\mathbb{N}$ *the following inequality holds:*

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

where

$$\theta_{;L} = \frac{\mu}{L},\ \lambda_{;L} = \left(1 + \frac{1}{2}\theta_{;L} + \frac{1}{2}\sqrt{\theta^2_{;L} + 4\theta_{;L}}\right).$$

Proof. Using the definition of the sequences A_k and solving the quadratic equation, we obtain:

$$A_k(1 + \mu A_{k-1}) = L\alpha_k^2 = L(A_k - A_{k-1})^2 = LA_k^2 - 2LA_kA_{k-1} + LA_{k-1},$$

$$, A_k^2 - A_k(1 + A_{k-1}(2 + \theta_{;L})) + A_{k-1}^2 = 0, D = (1 + 2A_{k-1} + \theta_{;L})^2 - 4A_k^2$$

$$A_{k;apex} = \frac{1}{2} + \left(1 + \frac{1}{2}\theta_{;L}\right)A_k \ ) \ A_k = \frac{1}{2}\left((1 + (2 + \theta_{;L}))A_{k-1} + \sqrt{D}\right)$$

$$\sqrt{D} = \sqrt{1 + (2\theta_{;L} + 4)A_{k-1} + \left(\theta^2_{;L} + 4\theta_{;L}\right)A_{k-1}^2} > A_{k-1}\sqrt{\theta^2_{;L} + 4\theta_{;L}} \ )$$

$$) \ A_k > \frac{1}{2}\left(2 + \theta_{;L} + \sqrt{\theta^2_{;L} + 4\theta_{;L}}\right)A_{k-1} =$$

$$= \left(1 + \frac{1}{2}\theta_{;L} + \frac{1}{2}\sqrt{\theta^2_{;L} + 4\theta_{;L}}\right)A_{k-1} \ )$$

$$) \ A_k > \frac{1}{L}\lambda^k_{;L},\ \lambda_{;L} = \left(1 + \frac{1}{2}\theta_{;L} + \frac{1}{2}\sqrt{\theta^2_{;L} + 4\theta_{;L}}\right).$$

□

Corollary 5.1.

$$\lambda_{;L} > \left(1 + \frac{1}{2}\sqrt{\theta_{;L}}\right)^2 = \left(1 + \sqrt{\theta_{;L}} + \frac{1}{4}\theta_{;L}\right),$$
$$\left(1 + \frac{1}{2}\sqrt{\theta_{;L}}\right)^2 > e^{\frac{1}{2}\,{}_{;L}}.$$

Claim 5. *If* μ 6 0 $8k$ 2 N *the following inequality holds:*

$$\frac{\sum_{j=0}^{k} A_j}{A_k} \;6\; 1 + \sqrt{\frac{L}{\mu}}.$$

Proof. According to the previous designations:

$$\lambda_{;L} = \left(1 + \frac{1}{2}\theta_{;L} + \frac{1}{2}\sqrt{\theta^2{}_{;L} + 4\theta_{;L}}\right), \theta_{;L} = \frac{\mu}{L}.$$

Using previous claim we can reduce this amount exponentially.

$$\frac{\sum_{j=0}^{k} A_j}{A_k} \;6\; \sum_{j=0}^{k} \lambda^{-j}{}_{;L} = \frac{\lambda^{k+1}{}_{;L} - 1}{\lambda^{k+1}{}_{;L} - \lambda^{k}{}_{;L}} \;6\; \frac{\lambda_{;L}}{\lambda_{;L} - 1} \;6\; 1 + \sqrt{\frac{L}{\mu}}.$$

□

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

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

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

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

Then by induction it is easy to get that:

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

□

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

$$\frac{\sum_{j=0}^{k} A_j}{A_k} \;6\; k.$$

Proof. The proof follows from the simple calculations:

$$\frac{\sum_{j=0}^{k} A_j}{A_k} \leq \frac{\sum_{j=0}^{k} a_j^2}{a_k^2} \leq \frac{k\alpha_{k-1}^2}{\alpha_k^2} = \frac{k\alpha_{k-1}^2}{(\frac{1}{2L} + \sqrt{\frac{1}{4L^2} + \alpha_{k-1}^2})^2} \leq k.$$

□

Lemma 5.2. *$8k > 1$ the following inequality holds:*

$$\psi_k(z_k) > \psi_{k-1}(z_{k-1}) + \frac{1+\mu A_{k-1}}{2}\|z_k - z_{k-1}\|_2^2 + \\ + \alpha_k \left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{\mu}{2}\|z_k - x_k\|_2^2 \right).$$

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

$$\langle \nabla \psi_{k-1}(z_{k-1}), z_k - z_{k-1} \rangle > 0, \quad \nabla \psi_{k-1}(z_{k-1}) = (z_{k-1} - x_0) + \\ + \sum_{j=0}^{k-1} \alpha_j \left(\nabla f(x_j) + \mu (z_{k-1} - x_j) \right))$$

$$) \quad \psi_k(z_k) = \psi_{k-1}(z_k) + \alpha_k \left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{\mu}{2}\|z_k - x_k\|_2^2 \right) =$$

$$= \frac{1}{2}\|z_k - x_0\|_2^2 + \sum_{j=0}^{k-1} \alpha_j \left(f(x_j) + \langle \nabla f(x_j), z_k - x_j \rangle + \frac{\mu}{2}\|z_k - x_j\|_2^2 \right) + \\ + \alpha_k \left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle \right) + \frac{\mu}{2}\|z_k - x_k\|_2^2.$$

From (25) and the above we obtain:

$$\psi_k(z_k) > \frac{1}{2}\|z_{k-1} - x_0\|_2^2 + \langle z_{k-1} - x_0, z_k - z_{k-1} \rangle + \frac{1}{2}\|z_{k-1} - z_k\|_2^2 + \\ + \sum_{j=0}^{k-1} \alpha_j \left(f(x_j) + \langle \nabla f(x_j), z_k - x_j \rangle + \frac{\mu}{2}\|z_k - x_j\|_2^2 \right) + \\ + \alpha_k \left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{\mu}{2}\|z_k - x_k\|_2^2 \right) = \\ = \sum_{j=0}^{k-1} \alpha_j \left(\langle \nabla f(x_j) + \mu (z_{k-1} - x_j), z_{k-1} - z_k \rangle \right) + \\ + \sum_{j=0}^{k-1} \alpha_j \left(f(x_j) + \langle \nabla f(x_j), z_k - x_j \rangle + \frac{\mu}{2}\|z_k - x_j\|_2^2 \right) + \\ + \alpha_k \left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{\mu}{2}\|z_k - x_k\|_2^2 \right) + \frac{1}{2}\|z_{k-1} - x_0\|_2^2 + \frac{1}{2}\|z_{k-1} - z_k\|_2^2.$$

Using the linearity of the dot product, we split the sum by two and apply to:

$$\mu \langle z_{k-1} - x_j, z_{k-1} - z_k \rangle.$$

Equality from (25), and finally get:

$$\psi_k(z_k) > \frac{1}{2}\|z_{k-1} - x_0\|_2^2 + \frac{1+\mu A_{k-1}}{2}\|z_{k-1} - z_k\|_2^2 + \\ + \sum_{j=0}^{k-1} \alpha_j \left(f(x_j) + \langle \nabla f(x_j), z_{k-1} - x_j \rangle + \frac{\mu}{2}\|z_{k-1} - x_j\|_2^2 \right) + \\ + \alpha_k \left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{\mu}{2}\|z_k - x_k\|_2^2 \right) = \\ = \psi_{k-1}(z_{k-1}) + \frac{1+\mu A_{k-1}}{2}\|z_k - z_{k-1}\|_2^2 + \\ + \alpha_k \left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{\mu}{2}\|z_k - x_k\|_2^2 \right).$$

□

Remark 1.

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

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

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 6.1. *$\forall k \in \mathbb{N}$ the following inequality holds:*

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

Proof. Base, $k = 0$:

$$f(x_0) \leq f(x_0) + \langle \nabla f(x_0), x_0 - x_0 \rangle + \frac{L}{2}\|x_0 - x_0\|_2^2 + \delta_2 \leq \\ \leq L\psi_0(z_0) - \frac{L\mu}{2}\|z_0 - x_0\|_2^2 + \delta_2 \leq L\psi_0(z_0) + \delta_2.$$

Induction step:

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

Using equations (25) we obtain:

$$A_k f(x_k) - A_{k-1}\delta_1 k x_{k-1} - x_k k_2 \le$$
$$\le A_{k-1}\left(f(x_k) + h\nabla f(x_k), x_{k-1} - x_k i\right) + \alpha_k\left(f(x_k) + h\nabla f(x_k), z_k - x_k i\right) +$$
$$+ \frac{(1+\mu_1 A_{k-1})}{2} k z_k - z_{k-1} k_2^2 + A_k\delta_2 - A_{k-1}\delta_1 k x_{k-1} - x_k k_2 \le$$
$$\le A_{k-1} f(x_{k-1}) + \alpha_k (f(x_k) h\nabla f(x_k), z_k - x_k i +$$
$$+ \frac{1+\mu_1 A_{k-1}}{2} k z_k - z_{k-1} k_2^2 + A_k\delta_2.$$

Using the induction hypothesis, we obtain:

$$A_k f(x_k) - A_{k-1}\delta_1 k x_{k-1} - x_k k_2 \le \psi_{k-1}(z_{k-1}) + \delta_2 \sum_{j=0}^{k-1} A_j + 2R\delta_1 A_{k-1} +$$
$$+ \frac{1+\mu_1 A_{k-1}}{2} k z_k - z_{k-1} k_2^2 + \alpha_k\left(f(x_k) + h\nabla f(x_k), z_k - x_k i\right) + A_k\delta_2.$$

Using lemma 5.2 we can get:

$$A_k f(x_k) \le A_{k-1}\delta_1 k x_{k-1} - x_k k_2 + \psi_k(z_k) + \delta_2 \sum_{j=0}^{k} A_j + 2R\delta_1 A_{k-1} =$$
$$= \psi_k(z_k) + \delta_2 \sum_{j=0}^{k} A_j + 2R\delta_1 A_{k-1} + \alpha_k k x_k - z_{k-1} k_2 \le$$
$$\le \psi_k(z_k) + \delta_2 \sum_{j=0}^{k} A_j + 2R\delta_1 A_{k-1} + \alpha_k \left(k z_{k-1} - x^* k_2 + k x_k - x^* k_2\right)\delta_1 \le$$
$$\le \psi_k(z_k) + \delta_2 \sum_{j=0}^{k} A_j + 2R\delta_1 A_{k-1} + 2\alpha_k R \ )$$
$$) \ A_k f(x_k) \le \psi(z_k) + \delta_2 \sum_{j=0}^{k} A_j + 2R\delta_1 A_k.$$

□

Remark 2.

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

Theorem 6.2. *If* $\mu \ge 0$ $8k \in \mathbb{N}$ *the following inequality holds:*

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

The proof repeats verbatim theorem 6.1, except for claim 2, replaced by claim 3.

Theorem 6.3. *If* $\delta_1 = \delta_2 = \delta_3 = 0$ *then* $R = R$.

Proof. Using theorem 6.1 we get $A_k f(x_k) \leq \psi_k(z_k)$ then:

$$\frac{1}{2}\|z_k - x^*\|_2^2 = \frac{1}{2}\|z_k - x^*\|_2^2 + A_k f(x_k) - A_k f(x_k) \leq$$
$$\leq \psi_k(z_k) + \frac{1}{2}\|z_k - x^*\|_2^2 - A_k f(x_k) \leq \sum_{j=0}^{k} \alpha_j (f(x_j) + \langle \nabla f(x_k), x^* - x_k \rangle +$$
$$+ \frac{\mu}{2}\|x^* - x_k\|_2^2) + \frac{1}{2}\|x^* - x_0\|_2^2 \leq A_k f(x_k) - A_k f(x_k) + \frac{1}{2}\|x_0 - x^*\|_2^2 = \frac{1}{2}R^2.$$

We now prove bound for the sequence x_k, similarly for x_k. We prove by induction, so assume fairness for $k-1$, base is obvious.

$$\|x_k - x^*\|_2^2 = \|\frac{A_{k-1}}{A_k}(x_{k-1} - x^*) + \frac{\alpha_k}{A_k}(z_k - x^*)\|_2^2 \leq$$
$$\leq \frac{A_{k-1}}{A_k}\|x_{k-1} - x^*\|_2^2 + \frac{\alpha_k}{A_k}\|z_k - x^*\|_2^2 = R^2.$$

□

Theorem 6.4 (convergence in function)**.** *Both inequalities take place with* $\mu \geq 0$

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

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 6.1 by theorem 6.2 and replacement claim 2 by claim 3, therefore, we present only the proof of the first inequality.

$$A_N f(x_N) \leq \psi_N(z_N) + \delta_2 \sum_{j=0}^{N} A_j + 2\bar{R}\delta_1 A_N \leq \frac{1}{2}\|x^* - x_0\|_2^2 +$$
$$+ \delta_2 \sum_{j=0}^{N} A_j + 2\bar{R}\delta_1 A_N + \sum_{j=0}^{N} \alpha_k (f(x_j) + \langle \nabla f(x_j), x^* - x_j \rangle + \frac{\mu_1}{2}\|x^* - x_i\|_2^2) \leq$$
$$\leq \delta_2 \sum_{j=0}^{N} A_j + 2\bar{R}\delta_1 A_N + \sum_{j=0}^{N} \alpha_k (\bar{R}\delta_1 + f(x^*)) + \frac{1}{2}R^2 =$$
$$= \delta_2 \sum_{j=0}^{N} A_j + 3\bar{R}\delta_1 A_N + A_N f(x^*) + \frac{1}{2}R^2 \quad (*)$$
$$(*) \quad f(x_N) - f(x^*) \leq LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{L}}N\right) + \left(1 + \sqrt{\frac{L}{\mu_1}}\right)\delta_2 + 3\bar{R}\delta_1.$$

□

Remark 3.

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

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

Remark 4.

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

$$f_\tau(x) = f(x) + \frac{\mu}{2}\|x - x_0\|_2^2 \to \min_{x \in Q},$$
$$r f_\tau(x) = r f(x) + \mu(x - x_0),$$
$$\tilde{r} f_\tau(x) = \tilde{r} f(x) + \mu(x - x_0),$$
$$\|\tilde{r} f_\tau(x) - r f_\tau(x)\|_2 = \|\tilde{r} f(x) - r f(x)\|_2 \leq \delta.$$

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

$$\begin{aligned}\|r f_\tau(x) - r f_\tau(y)\|_2^2 &= \|(r f(x) - r f(y)) + \mu(x - y)\|_2 \leq \\ &\leq \|(r f(x) - r f(y)\|_2 + \mu\|x - y\|_2 \leq \\ &\leq L_f\|x - y\|_2 + \mu\|x - y\|_2 \leq (L_f + \mu)\|x - y\|_2.\end{aligned}$$

We will assume, that $\mu < 1$. That is, we can let $L = 2(L_f + 1) = L + 2 > L_f$. The resulting function will already be strongly convex, which means that the second model is applicable to it $\tau = 2$. Using theorem 6.4 we can get the following inequality:

$$x^* = \operatorname*{argmin}_{x \in Q} f_\tau(x),$$
$$R = \|x^* - x_0\|_2,$$
$$f_\tau(x_k) - f_\tau(x^*) \leq \frac{L R^2}{2\lambda^k_{\overline{2}:2L}} + \left(1 + \sqrt{\frac{2L+4}{\mu}}\right)(\delta_2 + \delta_3))$$
$$f_\tau(x_k) - f_\tau(x^*) \leq LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2(L+2)}}k\right) + \left(1 + \sqrt{\frac{2L+4}{\mu}}\right)\left(\frac{1}{L} + \frac{1}{\mu}\right)\delta^2,$$
$$f_\tau(x^*) \leq f(x^*) + \frac{\mu}{2}R^2.$$

Then we can get convergence rate for not regularized function:

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

Using strong convexity of the function f we get:

$$f(x^*) + \frac{\mu}{2}R^2 \leq f(x^*) + \frac{\mu}{2}R^2 = f_\tau(x^*) \leq f_\tau(x^*) = f(x^*) + \frac{\mu}{2}R^2)$$
$$R \leq R.$$

Finally we get convergence:

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

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 $\|x^*\|_2 \leq R_*$. Then we choose a starting point for the STM algorithm in a ball of radius R_*, specifically put $x_{start} = 0$.

$$R = \|x^* - x_0\|_2 \leq R_*.$$

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

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

Lemma 6.5 (Bound for $\bar{R}$)**.** *Before the stopping criterion is satisfied, the following inequality holds:*

$$\bar{R} \leq R.$$

Proof. Note, that from $\|z_k - x^*\|_2 \leq R$ we get $\|x_k - x^*\|_2 \leq R, \|x_k - x^*\|_2 \leq R$ similarly to theorem 6.3. But it's worth noting that to estimate $\|x_k - x^*\|_2$, only inequalities are required for all $j \leq k-1$.

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

An analysis of the proof of theorem 6.2 gives a stronger convergence:

$$A_k f(x_k) \leq \psi_k(z_k) + \delta_2 \sum_{j=0}^{k} A_j + \delta_1 \sum_{j=1}^{k} \alpha_j \|x_j - z_{j-1}\|_2^2.$$

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

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

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

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

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

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

□

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

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

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

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

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

$$\le N\delta_2 + R_*\delta_1 + \delta_1 \sum_{j=1}^{N} \|x_j - z_{j-1}\|_2 + \zeta \ ) \ \frac{4LR^2}{N^2} \le \zeta \ , \ N^2 > \frac{4LR^2}{\zeta}.$$

Summing up, we obtain the following theorem:

Theorem 6.6. *For model $\tau = 1$ with $\mu = 0$, using stopping rule:*

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

We can guarantee, that:

$$\tilde{R} \le R.$$

And the criterion is reached after:

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

6.2 Relative noise.

Recall that we denote:

$$L = 2L_f.$$

Where L_f – Lipschitz constant of $r f$. From theorem 6.2 and similar to theorem 6.4 reasoning we get:

$$f(x_k) - f(x^*) \leq \frac{R^2}{A_k} + \frac{3}{2}\sum_{j=0}^{k-1}\frac{A_j\alpha^2 k r f(x_k)k_2^2}{A_k\mu} + \frac{3\alpha^2 k r f(x_k)k_2^2}{2\mu},$$

$$_k = f(x_k) - f(x^*),$$

$$_k \leq \frac{R^2}{A_k} + \frac{3}{2}\sum_{j=0}^{k-1}\frac{A_j\alpha^2 k r f(x_j)k_2^2}{A_k\mu} + \frac{3\alpha^2}{2\mu}k r f(x_k)k_2^2.$$

We define:

$$\theta = \frac{3L\alpha^2}{2\mu(1 - \frac{3L}{2})},$$

$$\lambda = \frac{R^2}{1 - \frac{3L}{2}}.$$

From inequality:

$$k r f(x_k)k_2^2 \leq L\left(f(x_k) - f(x^*)\right).$$

We obtain:

$$_k \leq \frac{\lambda}{A_k} + \theta\sum_{j=0}^{k-1}\frac{A_j}{A_k} \quad _j.$$

In these designations by induction we can obtain:

Claim 8.

$$_k \leq \frac{(1+\theta)^{k-1}}{A_k}\lambda + \theta\frac{A_0(1+\theta)^{k-1}}{A_k} \quad _0.$$

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

$$_k \leq \frac{\lambda}{A_k} + \theta\sum_{j=0}^{k-1}\frac{A_j}{A_k} \quad _j \leq$$

$$\leq \frac{\lambda}{A_k} + \sum_{j=0}^{k-1}\left(\frac{A_j}{A_k}\frac{(1+\theta)^{j-1}}{A_k}\lambda + \theta\frac{A_0(1+\theta)^{j-1}}{A_k} \quad _0\right) + \frac{A_0}{A_k} \quad _0 \leq$$

$$\leq \frac{\lambda}{A_k} + \sum_{j=0}^{k-2}\left(\frac{\lambda(1+\theta)^j}{A_k} + \theta\frac{A_0(1+\theta)^j}{A_k} \quad _0\right) + \frac{A_0}{A_k} \quad _0 =$$

$$= \frac{(1+\theta)^{k-1}}{A_k}\lambda + \theta\frac{A_0(1+\theta)^{k-1}}{A_k} \quad _0.$$

□

That is we can formulate the following inequality:

$$f(x_k) - f(x^*) \leq \frac{\lambda(1+\theta)^k}{A_k} + \theta\frac{A_0(1+\theta)^k}{A_k}\left(f(x_0) - f(x^*)\right).$$

Using corollary 5.1 we can estimate:

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

We will choose an alpha such that:

$$\frac{1+\theta}{A_k} \leq \frac{1}{1 + \frac{1}{3}\sqrt{\frac{1}{L}}}.$$

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

$$\begin{gathered}\alpha \leq \sqrt{\frac{1}{\frac{1}{2}\left(\frac{L}{\mu}\right)^{\frac{3}{2}} + \frac{2L}{\mu}}} \\ \alpha = O\left(\left(\frac{\mu}{L}\right)^{\frac{3}{4}}\right)\end{gathered} \tag{27}$$

From simple inequality:

$$1 + \frac{1}{3}\sqrt{\frac{\mu}{L}} > \exp\left(\frac{1}{6}\sqrt{\frac{\mu}{L}}\right).$$

We get the following theorem:

Theorem 6.7. *If in the model described in* (3) *in the strongly convex case we can chose* α *according to* (27) *we obtain:*

$$f(x_k) - f(x^*) \leq \left(\frac{LR^2}{1-\alpha^2} + \frac{3L\alpha^2}{2\mu(1-\alpha^2)}\left(f(x_0) - f(x^*)\right)\right)\exp\left(-\frac{1}{6}\sqrt{\frac{\mu}{L}}\right).$$

Corollary 6.8. *Under the conditions of the theorem, we obtain convergence in the argument:*

$$\|x_k - x^*\|_2^2 \leq R^2\left(\frac{2L}{\mu(1-\alpha^2)} + \frac{3L^2\alpha^2}{4\mu^2(1-\alpha^2)}\right)\exp\left(-\frac{1}{6}\sqrt{\frac{\mu}{L}}\right).$$

Proof. This is a direct consequence of the inequalities:

$$\begin{aligned}f(x_k) - f(x^*) &\leq \frac{L}{4}\|x_k - x^*\|_2^2 \\ f(x_k) - f(x^*), &> \frac{\mu}{2}\|x_k - x^*\|_2^2.\end{aligned}$$

□

7 Conclusions and observations

Remark 6.

Using Theorem 6.3 and assume, that Q – compact set we can we can denote R as $\mathrm{diam}(Q)$ instead of $\|x_0 - x^*\|_2$, then we can also bound $\bar{R} \leq R$ and this will simplify bounds in theorem 6.4.

Remark 7.

With the same assumption $\mu \geq 0$ we obtain a comparison of the two convergences in the Theorem 6.4. Recall that:

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

So if

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

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 6.4 one can set the desired accuracy of the solution.

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

Then we get from theorem 6.4 that:

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

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

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

$$\left(\frac{L+\mu}{\sqrt{\mu^3 L}}(\sqrt{2}+1)\right)\delta^2 \leq \frac{\varepsilon}{2},$$

$$\delta \leq \frac{\varepsilon}{\sqrt{2}}\sqrt{\frac{\sqrt{2}+1}{2}}\sqrt{\frac{L+\mu}{\sqrt{\mu^3 L}}},$$

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

$$LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2L}}N\right) \leq \frac{\varepsilon}{2},$$

$$N > 2\sqrt{\frac{2L}{\mu}}\left(\ln 2LR^2 + \ln \varepsilon^{-1}\right),$$

$$N = O\left(\sqrt{\frac{L}{\mu}}\ln\frac{LR^2}{\varepsilon}\right).$$

Remark 9.

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

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

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

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

Using inequality:

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

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

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

Similarly, get an estimate of the number of steps:

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

Remark 10.

Using remark 5 and theorem 6.6 we can apply it to problem:

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

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^*:

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

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

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

When the algorithm stops, we get the convergence:

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

Then we choose δ, ζ from the following conditions:

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

For example, we can let:

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

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 [41], that is we can consider that:

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

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:

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

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 6.4 and relevant remark 3. Namely, function satisfies the (δ, L, μ) model at point $x \in Q$ means, that exists functions $f(x)$ and $\psi(x, y)$, such that:

$$
\begin{gathered}
\forall y \in Q \\
\frac{\mu}{2}\|x - y\|_2^2 \leq f(x) - f(y) - \psi(x, y) \leq \frac{L}{2}\|x - y\|_2^2 + \delta.
\end{gathered}
$$

Similarly to papers [47], [17], the results also hold in the case of an unbounded set Q(result in [47] is on the page 26, obtained for fast adaptive gradient method page 13). Stopping criteria are also formulated, which give an estimate on $\bar{R}$ for a non-compact Q, remind that:

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

We also note that a similar models of $(\delta, \quad, L)$ and $(\delta, \quad, L, \mu)$ oracle was considered in the work [48]. Moreover, the function satisfies $(\delta, \quad, L, \mu)$–model

$$
\begin{gathered}
f(y) \leq f + \psi(y, x) + \quad \|x - y\| + \delta + LV(y, x), \\
f + \psi(x^*, x) + \mu V(y, x) \leq f(x^*), \\
f(x) - \delta \leq f(x) \leq f(x), \\
\psi(x, x) = 0.
\end{gathered}
$$

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 $\mu = 0$.

8 Numerical experiments

For testing STM for degenerate problems, the function described in [37] on page 69, that is:

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

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

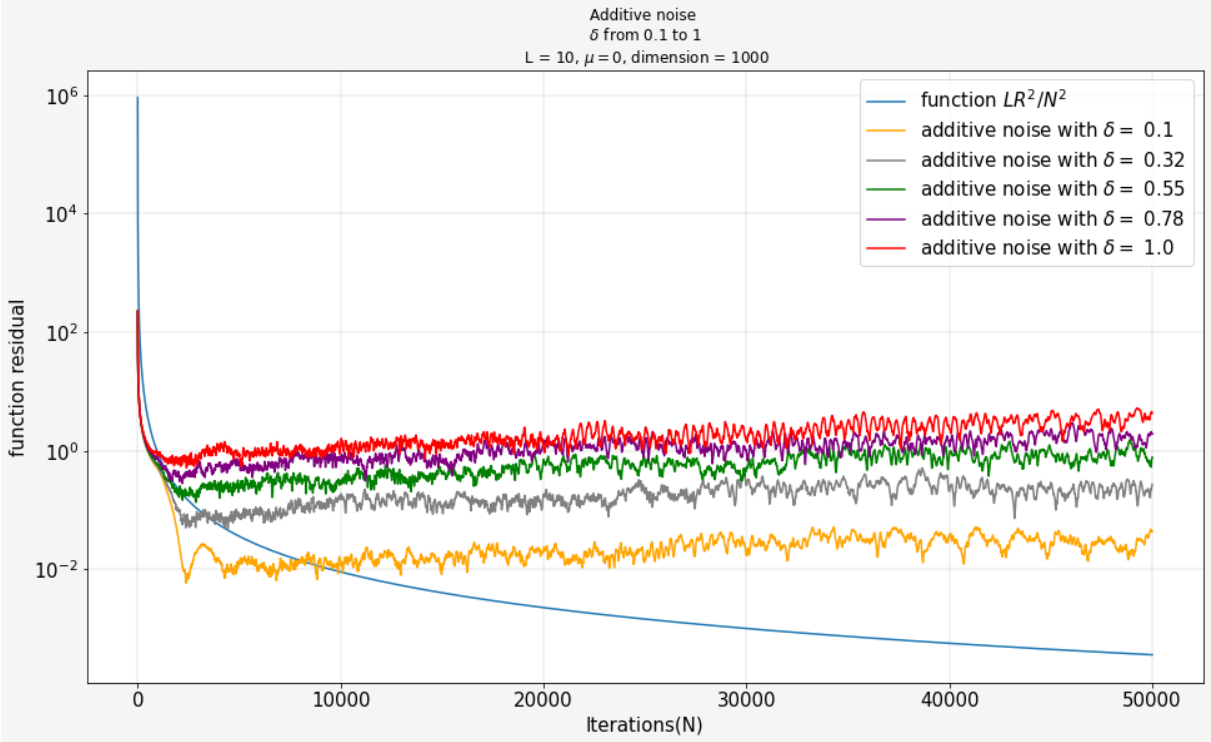

Figure 2: First test – first 50 000 steps.

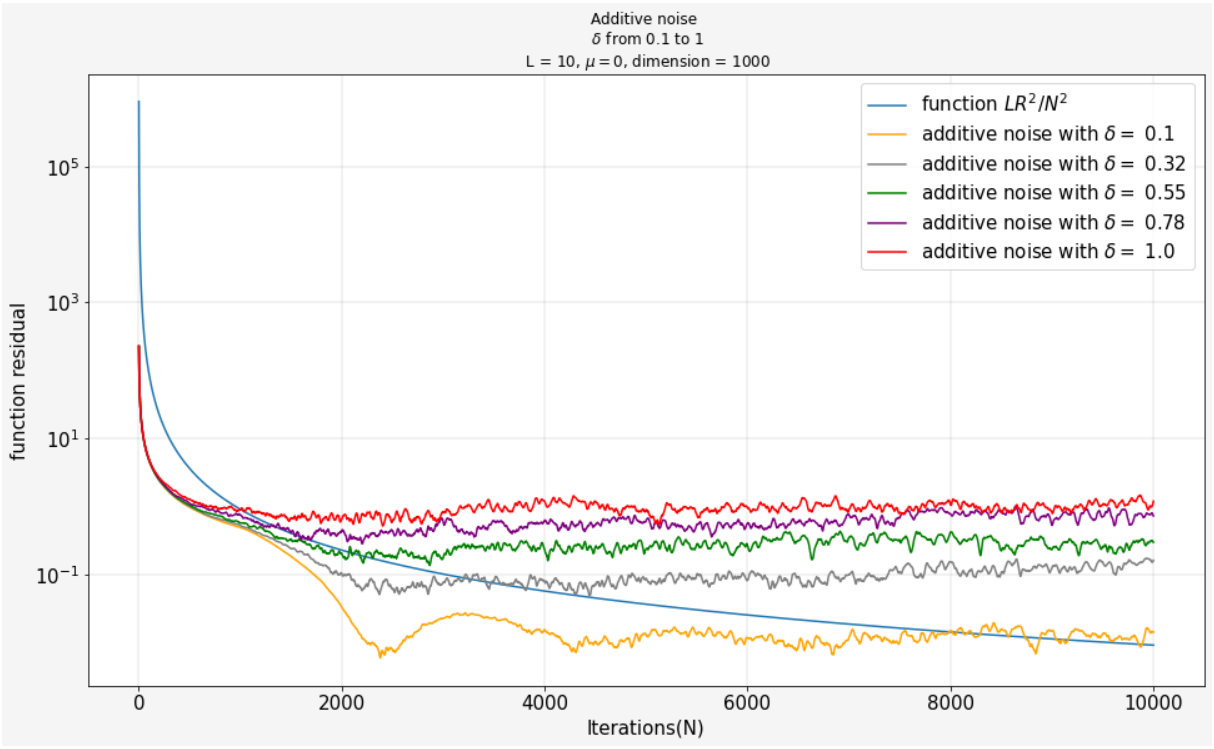

Figure 3: First test – first 10 000 steps.

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

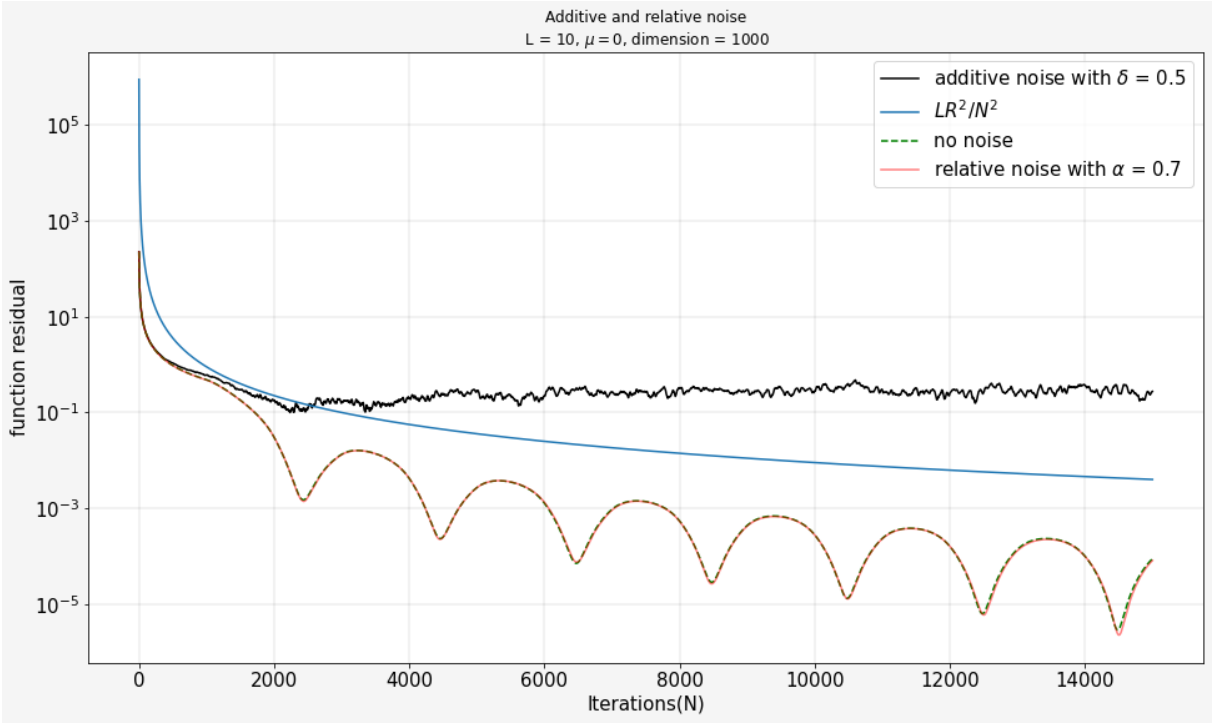

Figure 4: Second test – relative and additive types of noises comparison.

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

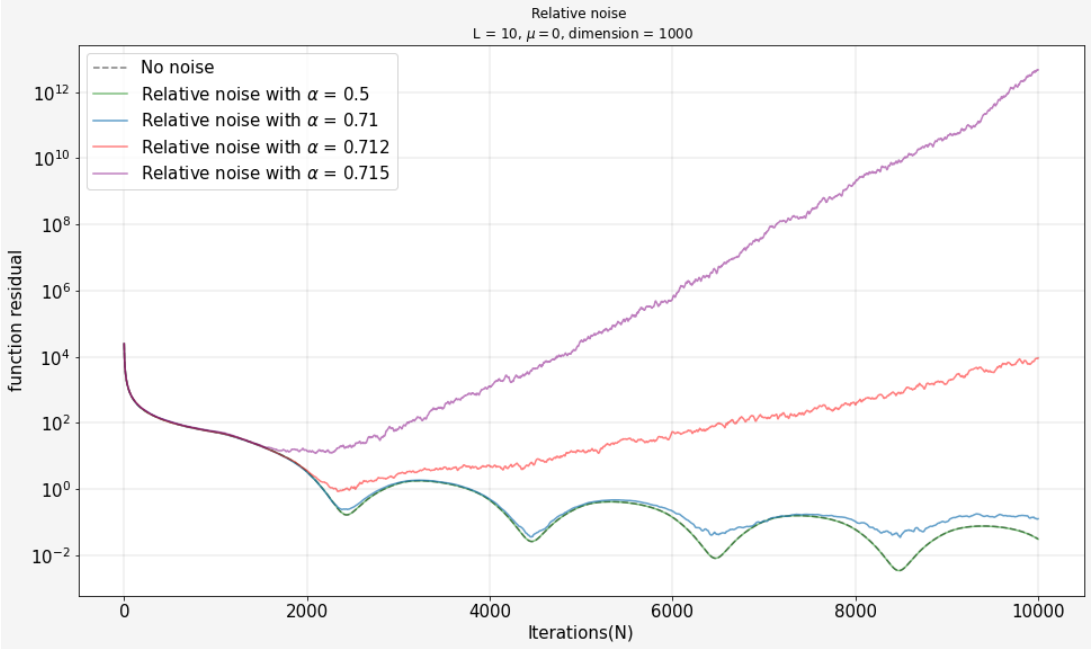

Figure 5: Third test – relative noise with different values of α for $\mu = 0$.

The last figure shows that for $\alpha \leq 0.71$ 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 [37] 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.

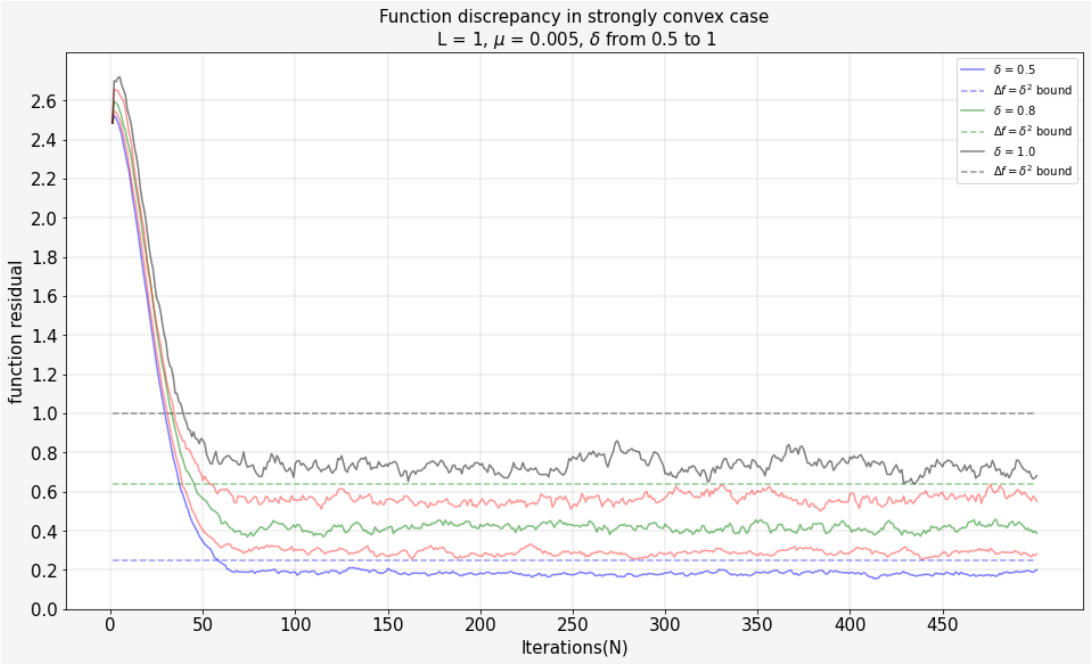

Figure 6: Fourth test – $\delta \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

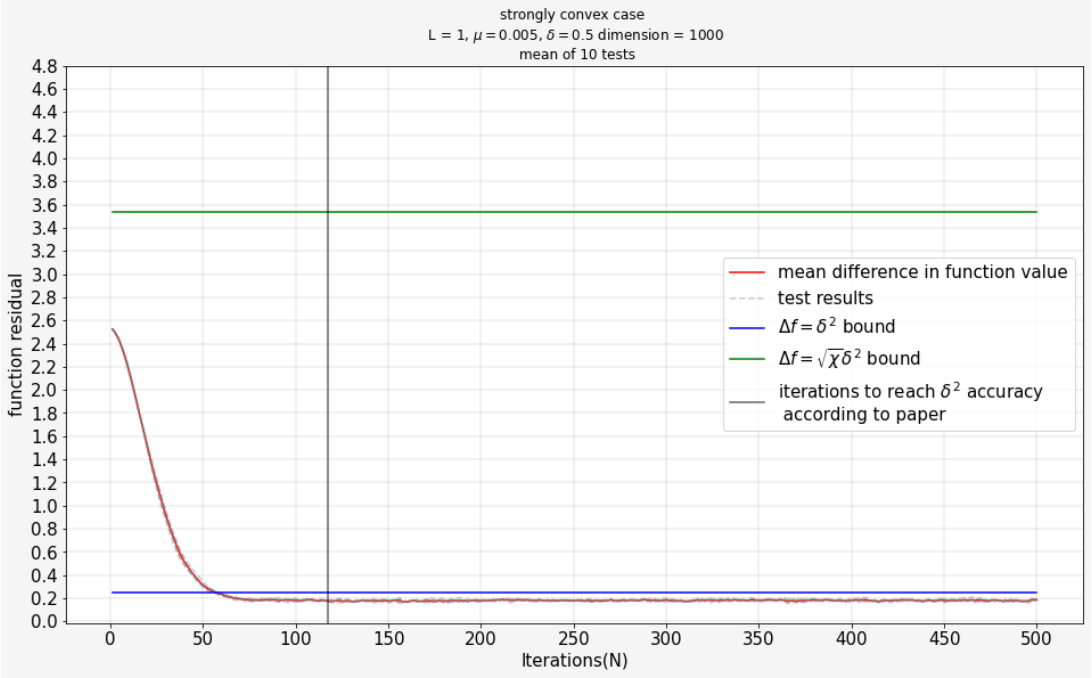

Figure 7: Fifth test – mean of 30 tests, level of approximation and required number of steps.

The last plot confirms theorem 6.4 and remark 8. Similarly to the degenerate case, consider the behavior of the method for different parameters α.

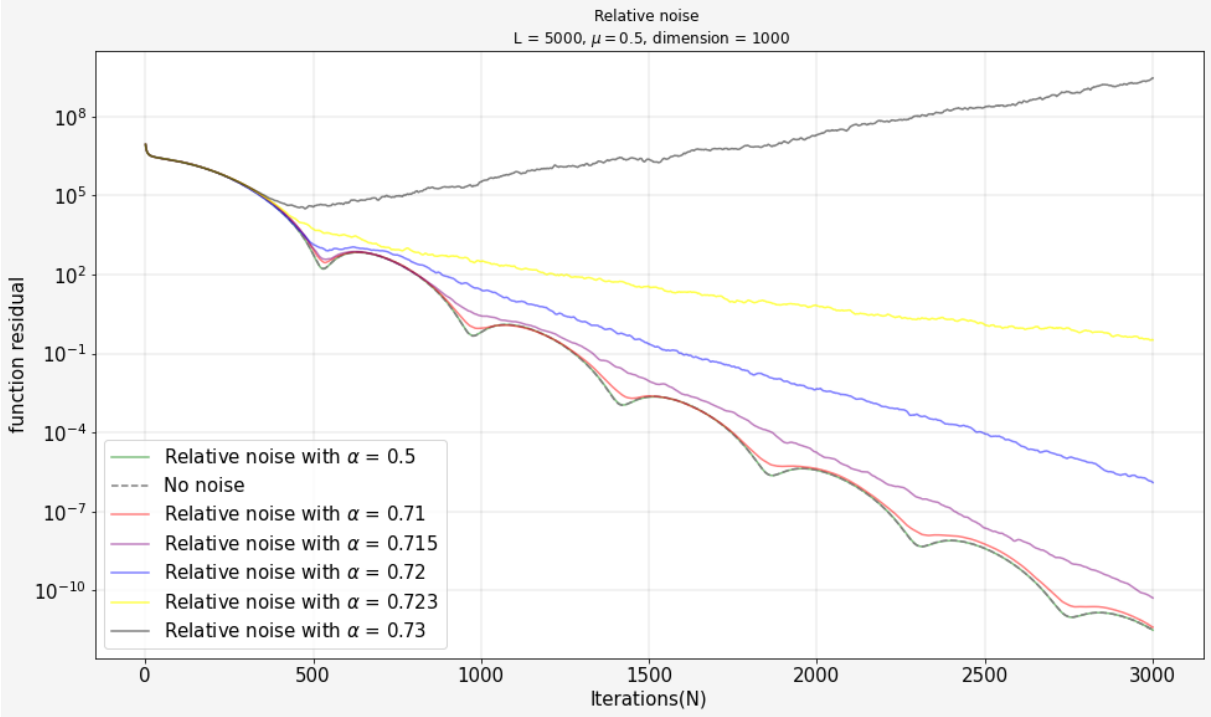

Figure 8: Sixth test – relative noise with different values of α for $\mu > 0$.

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^*, from the figure we can assume a value of 0.71, the convergence of the method does not deteriorate, and for large α values, the method diverges.

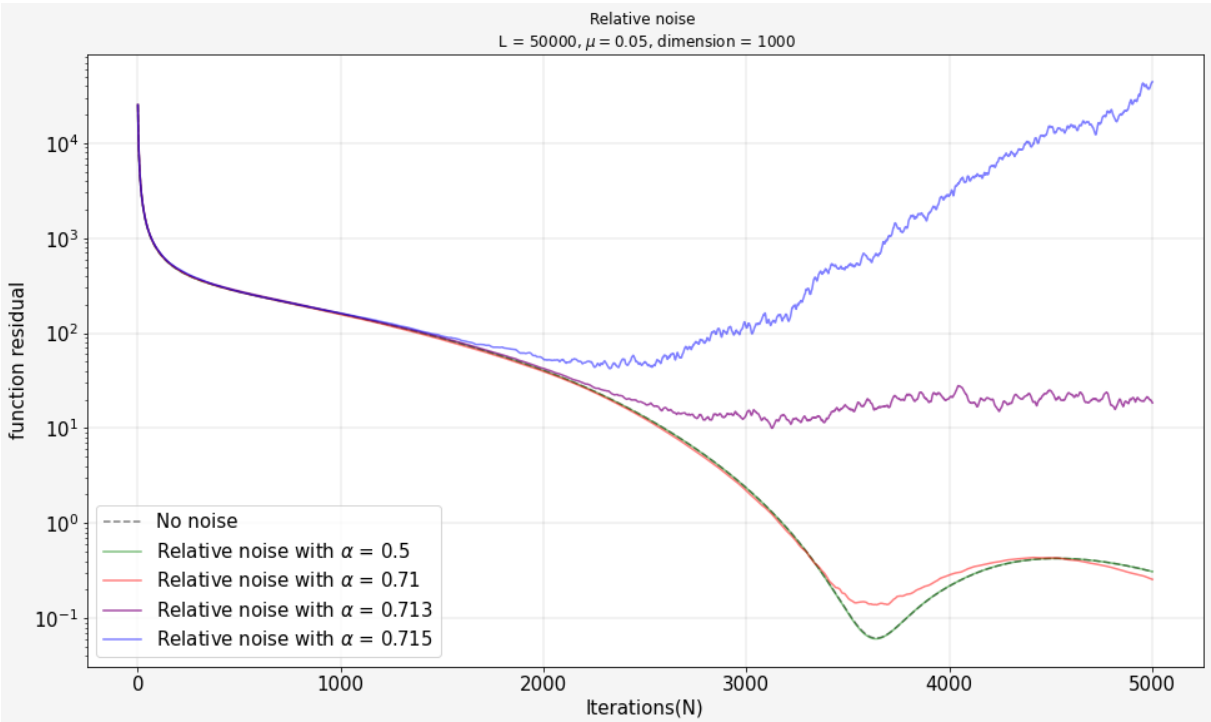

Figure 9: Seventh test – relative noise with different values of α for other L and μ.

Figure 9 shows, that The figure shows that changing the parameters L and μ, the value of the assumed threshold α^* does not change much. We also note that such threshold values turned out to be approximately equal for the degenerate and strongly convex problem.

9 Acknowledgment

The authors are grateful to Eduard Gorbunov for useful discussions.

References

[1] Ahmad Ajalloeian and Sebastian U Stich. Analysis of sgd with biased gradient estimators. *arXiv preprint arXiv:2008.00051*, 2020.

[2] Arya Akhavan, Massimiliano Pontil, and Alexandre B Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. *arXiv preprint arXiv:2006.07862*, 2020.

[3] Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 257–283, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[4] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[5] Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 240–265, Paris, France, 03–06 Jul 2015. PMLR.

[6] Aaron Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization (Lecture Notes)*. Personal web-page of A. Nemirovski, 2015.

[7] Albert S Berahas, Liyuan Cao, Krzysztof Choromanski, and Katya Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *arXiv preprint arXiv:1905.01332*, 2019.

[8] Aleksandr Beznosikov, Abdurakhmon Sadiev, and Alexander Gasnikov. Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 105–119. Springer, 2020.

[9] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

[10] Michael B Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. *arXiv preprint arXiv:1805.12591*, 2018.

[11] A. Conn, K. Scheinberg, and L. Vicente. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics, 2009.

[12] Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

[13] Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *arXiv preprint arXiv:2001.09545*, 2021.

[14] Olivier Devolder. Stochastic first order methods in smooth convex optimization. *CORE Discussion Paper 2011/70*, 2011.

[15] Olivier Devolder. *Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization*. PhD thesis, ICTEAM and CORE, Université Catholique de Louvain, 2013.

[16] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

[17] Olivier Devolder, François Glineur, Yurii Nesterov, et al. First-order methods with inexact oracle: the strongly convex case. *CORE Discussion Papers*, 2013016:47, 2013.

[18] Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *arXiv preprint arXiv:2011.11173*, 2020.

[19] Darina Dvinskikh and Alexander Gasnikov. Decentralized and parallelized primal and dual accelerated methods for stochastic convex programming problems. *Journal of Inverse and Ill-posed Problems*, 2021.

[20] Darina M Dvinskikh, Aleksandr Igorevich Turin, Alexander Vladimirovich Gasnikov, and Sergey Sergeevich Omelchenko. Accelerated and non accelerated stochastic gradient descent in model generality. *Matematicheskie Zametki*, 108(4):515–528, 2020.

[21] Pavel Dvurechensky. Numerical methods in large-scale optimization: inexact oracle and primal-dual analysis. *HSE. Habilitation*, 2020.

[22] Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1):121–145, 2016.

[23] Pavel Dvurechensky, Mathias Staudigl, and Shimrit Shtern. First-order methods for convex optimization. *arXiv preprint arXiv:2101.00935*, 2021.

[24] Yu G Evtushenko. Optimization and fast automatic differentiation. *Computing Center of RAS, Moscow*, 2013.

[25] A. V. Gasnikov, E. V. Gasnikova, Yu. E. Nesterov, and A. V. Chernov. Efficient numerical methods for entropy-linear programming problems. *Computational Mathematics and Mathematical Physics*, 56(4):514–524, 2016.

[26] Alexander Gasnikov. Universal gradient descent. *arXiv preprint arXiv:1711.00394*, 2017.

[27] Alexander Gasnikov, Sergey Kabanikhin, Ahmed Mohammed, and Maxim Shishlenin. Convex optimization in hilbert space with applications to inverse problems. *arXiv preprint arXiv:1703.00267*, 2017.

[28] Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018.

[29] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[30] Eduard Gorbunov, Darina Dvinskikh, and Alexander Gasnikov. Optimal decentralized distributed algorithms for stochastic convex optimization. *arXiv preprint arXiv:1911.07363*, 2019.

[31] Sergey I Kabanikhin. *Inverse and ill-posed problems: theory and applications*, volume 55. Walter De Gruyter, 2011.

[32] Dmitry Kamzolov, Pavel Dvurechensky, and Alexander V Gasnikov. Universal intermediate gradient method for convex problems with inexact oracle. *Optimization Methods and Software*, pages 1–28, 2020.

[33] Georgios Kotsalis, Guanghui Lan, and Tianjiao Li. Simple and optimal methods for stochastic variational inequalities, ii: Markovian noise and policy evaluation in reinforcement learning. *arXiv preprint arXiv:2011.08434*, 2020.

[34] Guanghui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

[35] Arkadi S Nemirovski. Regularizing properties of the conjugate gradient method for ill-posed problems. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 26(3):332–347, 1986.

[36] A.S. Nemirovsky and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. J. Wiley & Sons, New York, 1983.

[37] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[38] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, April 2017. First appeared in 2011 as CORE discussion paper 2011/16.

[39] Vasilii Novitskii and Alexander Gasnikov. Improved exploiting higher order smoothness in derivative-free optimization and continuous bandit. *arXiv preprint arXiv:2101.03821*, 2021.

[40] Fabian Pedregosa and Damien Scieur. Average-case acceleration through spectral density estimation. *arXiv preprint arXiv:2002.04756*, 2020.

[41] BT Poljak. Iterative algorithms for singular minimization problems. In *Nonlinear Programming 4*, pages 147–166. Elsevier, 1981.

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

[43] Boris Teodorovich Polyak and Aleksandr Borisovich Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, 26(2):45–53, 1990.

[44] Andrej Risteski and Yuanzhi Li. Algorithms and matching lower bounds for approximately-convex optimization. *Advances in Neural Information Processing Systems*, 29:4745–4753, 2016.

[45] R Tyrrell Rockafellar. *Convex analysis*, volume 36. Princeton university press, 1970.

[46] Damien Scieur and Fabian Pedregosa. Universal asymptotic optimality of polyak momentum. In *International Conference on Machine Learning*, pages 8565–8572. PMLR, 2020.

[47] Fedor Stonyakin, Alexander Tyurin, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, Darina Dvinskikh, Dmitry Pasechnyuk, Sergei Artamonov, and Victorya Piskunova. Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model. *arXiv preprint arXiv:2001.09013*, 2020.

[48] Fedyor Stonyakin. Adaptive methods for variational inequalities, minimization problems and functional with generalized growth condition. *MIPT. Habilitation*, 2020.

[49] Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017.

[50] Alexander Tyurin. Development of a method for solving structural optimization problems. *HSE. PhD Thesis*, 2020.

[51] F. Vasilyev. *Optimization Methods*. Moscow, Russia: FP, 2002.