Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Convex optimization with inexact gradients in Hilbert space and applications to elliptic inverse problems

Vladislav Matyukhin[1], Sergey Kabanikhin[2,3,4], Maxim Shishlenin[2,3,4],

Nikita Novikov[2,3], Artem Vasin[1], Alexander Gasnikov[1,5,6]

submitted: February 22, 2021

[1] Moscow Institute of Physics and Technology
Institutskiy Pereulok, 9
Dolgoprudny, Moscow Region
141701 Russian Federation
E-Mail: vladmatyukh@gmail.com

[2] Novosibirsk State University
Pirogova street, 1
Novosibirsk
630090 Russian Federation
E-Mail: m.shishlenin@g.nsu.ru

[3] Institute of computational mathematics and mathematical geophysics
Acad. Lavrentieva avenue, 6
630090 Novosibirsk Russian Federation
E-Mail: novikov-1989@yandex.ru

[4] Sobolev Institute of Mathematics
Acad. Koptyug avenue, 4
630090 Novosibirsk Russian Federation
E-Mail: kabanikh@math.nsc.ru

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

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

No. 2815
Berlin 2021

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

Key words and phrases. Convex optimization, inexact oracle, inverse and ill-posedproblem, gradient method.

The research of V.V. Matyukhin and A.V. Gasnikov in Sections 1,2,3,4,5 was supported by Russian Science Foundation (project No. 21-71-30005). The research of S.I. Kabanikhin, M.A. Shishlenin and N.S. Novikov in the last section was supported by RFBR 19-01-00694 and by the comprehensive program of fundamental scientific researches of the SB RAS II.1, project No. 0314-2018-0009. 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/

Convex optimization with inexact gradients in Hilbert space and applications to elliptic inverse problems

Vladislav Matyukhin, Sergey Kabanikhin, Maxim Shishlenin, Nikita Novikov, Artem Vasin, Alexander Gasnikov

Abstract

In this paper we propose the gradient descent type methods to solve convex optimization problems in Hilbert space. We apply it to solve ill-posed Cauchy problem for Poisson equation and make a comparative analysis with Landweber iteration and steepest descent method. The theoretical novelty of the paper consists in the developing of new stopping rule for accelerated gradient methods with inexact gradient (additive noise). Note that up to the moment of stopping the method "doesn't feel the noise". But after this moment the noise start to accumulate and the quality of the solution becomes worse for further iterations.

1 Introduction

In this paper we propose the gradient descent type methods to solve convex optimization problems in Hilbert space. We apply it to solve ill-posed Cauchy problem for Poisson equation and make a comparative analysis with Landweber iteration and steepest descent method. The theoretical novelty of the paper consists in the developing of proper stopping rule for accelerated gradient methods with inexact gradient.

Following to the works [1, 32, 2] we develop new approaches to solve convex optimization problems in Hilbert space [29, 28]. The main difference from the existing approaches is that we don't approximate infinite-dimensional problem by the finite one (see [1, 2]). We try to solve the problem in Hilbert space (infinite-dimensional). But we try to do it with the conception of the inexact oracle. That is we use an approximation of the problem only when we calculate gradient (Frechet derivative) of the functional. This generates inexactness in gradient calculations. We try to combine known results in this area and to understand the best way to solve convex optimization problems in Hilbert space with application to ill-posed and inverse problems [32].

It's important to note, that in the paper we consider only gradient type procedures without 1D-line search. So it means that very popular in practice methods, like steepest descent and conjugate gradient [30, 25] and their nonlinear analogues [21], do not take into account. The reason is that we try to develop an approach that justified theoretically. For all of these methods, there exist some troubles with error accumulation [27]. In the worth case, algorithms may diverge. In this paper we describe how to control this divergence and stop in time for gradient-type methods without 1D-line search. Fortunately, there exist alternative procedures to 1D-line search (Armijo, Wolf, Nesterov rules [30, 31, 25]) that perform the same function as 1D-line search. We will use the Nesterov's rule [31, 25, 22]. It allows us to choose an adaptive stepsize policy.

The important part of the paper is an adaptation of the modern results developed for convergence of gradient type methods with inexact gradient for a specific class of inverse ill-posed convex problems

DOI 10.20347/WIAS.PREPRINT.2815

in Hilbert space. We will use the conception of inexactness developed for about 7 years ago by Yu. Nesterov, G. Glineur and O. Devolder [9] and demonstrate how to reduce an additive noise conception to considered one. The basic algorithms are gradient descent [30, 25], fast (accelerated) gradient method in variant of Similar Triangles Method (STM) [4] and its combinations [9, 20]. For these algorithms the theory of gradient error accumulation is well developed [9, 25, 26, 24, 20, 23]. Basically, the theoretical foundation of the facts we use in this paper can be found in the paper [20] and recent arXive preprint [23].

The structure of the paper is as follows. In section 2 we described primal approaches (we solve exactly the problem we have) based on contemporary versions of fast gradient descent methods and its adaptive variants.

In section 3 we described dual approaches (we solve a dual problem) based on the same methods. We try to describe all the methods with the exact estimations of their convergence. But every time we have in mind concrete applications. Since that we include in the description of algorithms such details that allow methods to be more practical.

Section 4 contains new result about proper stopping rule for STM when the noise in gradient is additive. This result can be briefly formulated as follows. Having δ-inexact gradient (inexactness is additive), gradient descent and accelerated gradient descent (we consider STM) converge almost like their noise-free analogues up to an accuracy in function $\sim \delta R$, where R – corresponds to the size of the solution. After that we should stop the algorithm, since an error can be further accumulated and caused divergence of the method [27]. This result seems to be rather unexpected for accelerated algorithms, due to pessimistic results, mentioned in section 3, about accumulation of the error in another conception of noise.

The rest part of the paper devoted to applications of the described results to elliptic ill-posed inverse problems. This part experimentally confirm conclusions that have been done in the previous sections.

2 Primal approaches

Assume that $q \in H$, where H is a Hilbert space with scalar product denoted by $\langle\ ,\ \rangle$ (H isn't necessarily finite). Let's introduce convex functional $J(q)$. In this paper we investigate the following optimization problem

$$J(q) \to \min_q . \tag{1}$$

Let's introduce starting point y^0 and

$$R = \left\|y^0 - q_*\right\|_2 ,$$

where q_* is such a solution of (1) that gives R the smallest value. We assume that at least one solution exists [1].

Assume that $J(q)$ has Lipchitz Frechet derivative

$$\|\nabla J(q_2) - \nabla J(q_1)\|_2 \leq L \|q_2 - q_1\|_2 , \tag{2}$$

where $\|q\|_2^2 = \|q\|_H^2 = \langle q, q\rangle_H$. In (2) we also use that due to the Riesz representation theorem [3], one may considered $\nabla J(q)$ to be the element of $H^* = H$.

Example 1. Assume that linear operator $A:\ H_1 \to H_2$, $b \in H_2$. Let's consider the following convex optimization problem [1, 32]:

$$J(q) = \frac{1}{2}\|Aq - f\|_{H_2}^2 \to \min_q .$$

Note that

$$\nabla J\left(q\right) = A^{*}\left(Aq - f\right).$$

Formula (2) is equivalent to

$$\left\langle Aq, Aq\right\rangle_{H_2} = \left\|Aq\right\|_{H_2}^{2} \le L\left\|q\right\|_{H_1}^{2} = L\left\langle q, q\right\rangle_{H_1},$$

i.e. $L = \|A\|_{H_1 \to H_2}^{2}$. ■

Now following to [4, 5, 6] (most of the ideas below goes back to the pioneer's works of B.T. Polyak, A.S. Nemirovski, Yu.E. Nesterov) we describe optimal (up to absolute constant factor or logarithmic factor in strongly convex case) numerical methods [7, 28, 43] (in terms of the number of ideal calculations of $\nabla J\left(q\right)$ and $J\left(q\right)$) for solving the problem (1). The rates of convergence that obtained in theorems 1, 2 can be reached (in case of example 1) also by conjugate-gradient methods [1, 32], but we lead these estimates under more general conditions.

Algorithm 1 Similar Triangular Method STM $\left(y^{0}, L\right)$

Input: $A_0 = \alpha_0 = 1/L, \quad k = 0; \quad q^{0} = u^{0} = y^{0} - \alpha_0 \nabla J\left(y^{0}\right).$

1: **Put**

$$\alpha_{k+1} = \frac{1}{2L} + \sqrt{\frac{1}{4L^2} + \frac{A_k}{L}}, \quad A_{k+1} = A_k + \alpha_{k+1},$$

$$y^{k+1} = \frac{\alpha_{k+1}u^{k} + A_k q^{k}}{A_{k+1}},$$

$$u^{k+1} = u^{k} - \alpha_{k+1}\nabla J\left(y^{k+1}\right),$$

$$q^{k+1} = \frac{\alpha_{k+1}u^{k+1} + A_k q^{k}}{A_{k+1}}.$$

2: If stopping rule doesn't satisfy, put $k := k+1$ and **go to** 1.

If $J\left(q_*\right) = 0$ (see example 1) then stopping rule has the form $J\left(q^{k}\right) \le \varepsilon$.

Theorem 2.1. *(see [4]) Assume that (2) holds true. Then for STM* $\left(y^{0}, L\right)$*:*

$$J\left(q^{N}\right) - J\left(q_*\right) \le \frac{4LR^2}{N^2}.$$

Sometimes it's hardly possible to estimate L that are used in STM. Moreover even when we can estimate L we have to used the worth one (the largest one). Is it possible to change the worth case L to the average one (among all the iterations)? The answer is YES [4] (see ASTM below).

Theorem 2.2. *(see [4]) Assume that (2) holds true. Then for ASTM* $\left(y^{0}\right)$*:*

$$J\left(q^{N}\right) - J\left(q_*\right) \le \frac{8LR^2}{N^2}.$$

The average number of calculations of $J\left(q\right)$ per iteration roughly equals 4 and the average number of calculations Frechet derivative $\nabla J\left(q\right)$ per iteration roughly equals 2.

Algorithm 2 Adaptive Similar Triangular Method ASTM (y^0)

Input: $A_0 = \alpha_0 = 1/L_0^0 = 1, \quad k = 0, \quad j_0 = 0; \quad q^0 := u^0 := y^0 - \alpha_0 \nabla J(y^0)$.

1: **while** $J(q^0) > J(y^0) + \langle \nabla J(y^0), q^0 - y^0 \rangle + \frac{L_0^{j_0}}{2} \|q^0 - y^0\|_2^2$ **do**

2: $j_0 := j_0 + 1; L_0^{j_0} := 2^{j_0} L_0^0; (A_0 :=)\, \alpha_0 := \frac{1}{L_0^{j_0}}, \quad q^0 := u^0 := y^0 - \alpha_0 \nabla J(y^0)$.

3: **end while**

4: **Put** $L_{k+1}^0 = L_k^{j_k}/2$, $j_{k+1} = 0$.

$$\alpha_{k+1} := \frac{1}{2L_{k+1}^0} + \sqrt{\frac{1}{4\left(L_{k+1}^0\right)^2} + \frac{A_k}{L_{k+1}^0}}, \quad A_{k+1} := A_k + \alpha_{k+1},$$

$$y^{k+1} = \frac{\alpha_{k+1} u^k + A_k q^k}{A_{k+1}},$$

$$u^{k+1} = u^k - \alpha_{k+1} \nabla J\left(y^{k+1}\right),$$

$$q^{k+1} = \frac{\alpha_{k+1} u^{k+1} + A_k q^k}{A_{k+1}}.$$

5: **while** $J\left(y^{k+1}\right) + \left\langle \nabla J\left(y^{k+1}\right), q^{k+1} - y^{k+1} \right\rangle + \frac{L_{k+1}^{j_{k+1}}}{2} \left\|q^{k+1} - y^{k+1}\right\|_2^2 < J\left(q^{k+1}\right)$ **do**

6:

$$j_{k+1} := j_{k+1} + 1; L_{k+1}^{j_{k+1}} = 2^{j_{k+1}} L_{k+1}^0;$$

$$\alpha_{k+1} := \frac{1}{2L_{k+1}^{j_{k+1}}} + \sqrt{\frac{1}{4\left(L_{k+1}^{j_{k+1}}\right)^2} + \frac{A_k}{L_{k+1}^{j_{k+1}}}}, \quad A_{k+1} := A_k + \alpha_{k+1};$$

$$y^{k+1} := \frac{\alpha_{k+1} u^k + A_k q^k}{A_{k+1}}, \quad u^{k+1} := u^k - \alpha_{k+1} \nabla J\left(y^{k+1}\right),$$

$$q^{k+1} := \frac{\alpha_{k+1} u^{k+1} + A_k q^k}{A_{k+1}}.$$

7: **end while**

8: If stopping rule doesn't satisfy, put $k := k + 1$ and **go to** 4.

2.1 Restart technique

Assume that (2) holds true, $J\left(q^{*}\right)=0$ and [1]

$$\langle \tilde{q}, J''\left(q\right)\tilde{q}\rangle \geq \mu\left\langle \tilde{q}, \tilde{q}\right\rangle \tag{3}$$

for all $\tilde{q}, q \in H$ $(\mu>0)$.

For example 1 (3) can be simplified: for all $q \in H$

$$\langle Aq, Aq\rangle_{H_2}=\|Aq\|_{H_2}^{2} \geq \mu\|q\|_{H_1}^{2}=\mu\langle q, q\rangle_{H_1} \quad (\mu>0).$$

In this case we may restart the method we choose each time when the value of the goal functional becomes less than one half from starting value at this restart. Since restart criteria is verifiable in practice this construction can be easily implemented in practice. For example, if we choose (A)STM, then by restart construction one can obtain the method that required

$$N=\mathrm{O}\left(\sqrt{\frac{L}{\mu}}\log_2\left(\frac{\mu R^2}{\varepsilon}\right)\right).$$

calculations of $\nabla J(q)$ (and $J(q)$ in case of ASTM) to generate q^N that guarantee

$$J(q^N)-J(q_*)=J(q^N)\leq\varepsilon$$

Another approach assumes that we have verifiable stopping criteria $J(\bar{q}_k^{N_k})-J(q_*)\leq\varepsilon$ and doesn't assume that $J(q_*)=0$. For example, since

$$J(\bar{q}_k^{N_k})-J(q_*)\leq\frac{1}{2\mu}\left\|\nabla J(\bar{q}_k^{N_k})\right\|_2^2$$

we may have such a criteria if we know lower bound on μ. Suppose now that we can estimate μ from above: $\mu\leq\mu_0\ll L$. Let's consider RSTM (y^0, L, μ_0). We introduce the following restart version of this algorithm (here $\text{STM}_q^{N_k}(y_0^k, L)=q^{N_k}$ and $\text{STM}_u^{N_k}(y_0^k, L)=u^{N_k}$)

Algorithm 3 RSTM (y^0, L, μ_0)

1: $k=0;\quad y_0^0=y^0;$

2: **repeat**

$$\bar{q}_k^{N_k}=\frac{1}{2}\text{STM}_q^{N_k}\left(y_k^0, L\right)+\frac{1}{2}\text{STM}_u^{N_k}\left(y_k^0, L\right),\ \text{where } N_k=2\sqrt{L/\mu_0};$$

3: **until** $J(\bar{q}_k^{N_k})-J(q_*)\leq\varepsilon$

One can show [8] that after $k=\mathrm{O}\left(\sqrt{\mu_0/\mu}\log_2\left(L\left(J\left(y^0\right)-J(q_*)\right)/(\mu\varepsilon)\right)\right)$ restarts RSTM (y^0, L, μ_0) will stop and the total number of calculations of $\nabla J\left(q\right)$ (and $J\left(q\right)$ in case ASTM) can be estimated as follows

$$\mathrm{O}\left(\sqrt{\frac{L}{\mu}}\log_2\left(\frac{L\left(J\left(y^0\right)-J(q_*)\right)}{\mu\varepsilon}\right)\right).$$

Since (3) is true and $J\left(q^{*}\right)=0$ one has

$$\frac{\mu}{2}\left\|\bar{q}_k^{N_k}-q_*\right\|_2^2\leq J\left(\bar{q}_k^{N_k}\right)-J\left(q_*\right)\leq\varepsilon.$$

Therefore,

$$\left\|\bar{q}_k^{N_k} - q_*\right\|_2 \le \sqrt{\frac{2\varepsilon}{\mu}}.$$

Hence after ($\varepsilon := \mu\varepsilon^2/2$)

$$\mathrm{O}\left(\sqrt{\frac{L}{\mu}}\log_2\left(\frac{2LJ\left(y^0\right)}{\mu^2\varepsilon^2}\right)\right)$$

calculations of $\nabla J\left(q\right)$ one can obtain

$$\left\|\bar{q}_k^{N_k} - q_*\right\|_2 \le \varepsilon.$$

Hence after ($\varepsilon := \mu\varepsilon^2/2$)

$$\mathrm{O}\left(\sqrt{\frac{L}{\mu}}\log_2\left(\frac{2L\left(J\left(y^0\right) - J(q_*)\right)}{\mu^2\varepsilon^2}\right)\right)$$

calculations of $\nabla J\left(q\right)$ one can obtain

$$\left\|\bar{q}_k^{N_k} - q_*\right\|_2 \le \varepsilon.$$

For the adaptive case in RASTM (y^0, μ_0) we have to replace $N_k = 2\sqrt{L/\mu_0}$ (since we don't know L) by N_k – the smallest natural number that satisfies $A_{N_k} \ge 4/\mu_0$ [20]. All the estimates hold true up to constant factors.

2.2 Gradient descent

In the literature one can typically meet non-accelerated simple gradient descent method GD $(q^0 = y^0, L)$ [1, 32, 9]

$$q^{k+1} = q^k - \frac{1}{L}\nabla J\left(q^k\right) \tag{4}$$

or $\overline{\text{GD}}\,(y^0, L)$, $q^0 = 0$

$$\begin{cases} y^{k+1} = y^k - \frac{1}{L}\nabla J\left(y^k\right), \\ q^{k+1} = \frac{k}{k+1}q^k + \frac{1}{k+1}y^{k+1}. \end{cases} \tag{5}$$

In the case (2), (3) method (4) requires $\mathrm{O}\left(\left(L/\mu\right)\log_2\left(2LR^2/(\mu\varepsilon^2)\right)\right)$ calculations of $\nabla J\left(q\right)$ for $\left\|q^N - q_*\right\|_2 \le \varepsilon$. In the case (2) method (5) requires $\mathrm{O}\left(LR^2/\varepsilon\right)$ calculations of $\nabla J\left(q\right)$ for $J\left(q^N\right) - J\left(q_*\right) \le \varepsilon$. Note that for the (A)STM these quantities are smaller

$$\mathrm{O}\left(\sqrt{\frac{L}{\mu}}\log_2\left(\frac{2LR^2}{\mu\varepsilon^2}\right)\right), \quad \mathrm{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right). \tag{6}$$

In reality in (6) for ASTM (analogously for RASTM) one can insert the average L among all the iterations. This could be much smaller than the worth one.

One can easily propose adaptive version of GD and $\overline{\text{GD}}$. For (5) let's introduce AGD $(q^0 = y^0)$.

Let's introduce $\overline{\text{AGD}}\,(y^0)$.

Algorithm 4 AGD $\left(q^0 = y^0\right)$

Input: $L_0^0 = 1, \quad k = 0, \quad j_0 = 0; \quad q^1 = q^0 - \frac{1}{L_0^0}\nabla J\left(q^0\right).$

1: **while** $J\left(q^1\right) > J\left(q^0\right) + \left\langle \nabla J\left(q^0\right), q^1 - q^0\right\rangle + \frac{L_0^{j_0}}{2}\left\|q^1 - q^0\right\|_2^2$ **do**

2:
$$j_0 := j_0 + 1; L_0^{j_0} := 2^{j_0} L_0^0;$$
$$q^1 = q^0 - \frac{1}{L_0^{j_0}}\nabla J\left(q^0\right).$$

3: **end while**

4: **Put**
$$L_{k+1}^0 = L_k^{j_k}/2, j_{k+1} = 0.$$
$$q^{k+1} = q^k - \frac{1}{L_{k+1}^0}\nabla J\left(q^k\right).$$

5: **while** $J\left(q^k\right) + \left\langle \nabla J\left(q^k\right), q^{k+1} - q^k\right\rangle + \frac{L_{k+1}^{j_{k+1}}}{2}\left\|q^{k+1} - q^k\right\|_2^2 < J\left(q^{k+1}\right)$ **do**

6:
$$j_{k+1} := j_{k+1} + 1; L_{k+1}^{j_{k+1}} = 2^{j_{k+1}} L_{k+1}^0; q^{k+1} := q^k - \frac{1}{L_{k+1}^{j_{k+1}}}\nabla J\left(q^k\right).$$

7: **end while**

8: If stopping rule doesn't satisfy, put $k := k + 1$ and **go to** 4.

Algorithm 5 $\overline{\text{AGD}}$ $\left(y^0\right)$

Input: $L_0^0 = 1, \quad k = 0, \quad j_0 = 0; \quad q^0 = 0; \quad y^1 = y^0 - \frac{1}{L_0^0}\nabla J\left(y^0\right).$

1: **while** $J\left(y^1\right) > J\left(y^0\right) + \left\langle \nabla J\left(y^0\right), y^1 - y^0\right\rangle + \frac{L_0^{j_0}}{2}\left\|y^1 - y^0\right\|_2^2$ **do**

2:
$$j_0 := j_0 + 1; L_0^{j_0} := 2^{j_0} L_0^0;$$
$$y^1 = y^0 - \frac{1}{L_0^{j_0}}\nabla J\left(y^0\right).$$

3: **end while**

4: **Put**
$$L_{k+1}^0 = L_k^{j_k}/2, j_{k+1} = 0.$$
$$y^{k+1} = y^k - \frac{1}{L_{k+1}^0}\nabla J\left(y^k\right),$$
$$q^{k+1} = \frac{k}{k+1}q^k + \frac{1}{k+1}y^{k+1}.$$

5: **while** $J\left(y^k\right) + \left\langle \nabla J\left(y^k\right), y^{k+1} - y^k\right\rangle + \frac{L_{k+1}^{j_{k+1}}}{2}\left\|y^{k+1} - y^k\right\|_2^2 < J\left(y^{k+1}\right)$ **do**

6:
$$j_{k+1} := j_{k+1} + 1; L_{k+1}^{j_{k+1}} = 2^{j_{k+1}} L_{k+1}^0;$$
$$y^{k+1} := y^k - \frac{1}{L_{k+1}^{j_{k+1}}}\nabla J\left(y^k\right), \quad q^{k+1} = \frac{k}{k+1}q^k + \frac{1}{k+1}y^{k+1}.$$

7: **end while**

8: If stopping rule doesn't satisfy, put $k := k + 1$ and **go to** 4.

The same ("one can insert the average L among all the iterations in estimations $\mathrm{O}\left(\left(L/\mu\right)\log_2\left(LR^2/\varepsilon\right)\right)$, $\mathrm{O}\left(LR^2/\varepsilon\right)$" and "the average number of calculations of $J\left(q\right)$ and $\nabla J\left(q\right)$ per one iteration roughly equals 2") one can say about the rates of convergence for AGD $\left(q^0 = y^0\right)$ and $\overline{\text{AGD}}\left(y^0\right)$.

Note, that if we change in A(GD), $\overline{\text{AGD}}$ $y^{k+1} = y^k - \frac{1}{L}\nabla J\left(y^k\right)$ by $y^{k+1} = y^k - \alpha_k \nabla J\left(y^k\right)$, where $\alpha_k = \arg\min\limits_{\alpha\geq 0} J\left(y^k - \alpha\nabla J\left(y^k\right)\right)$, then the rates of convergence (in general) don't change up to a constant factors.

2.3 Devolder–Glineur–Nesterov conception of inexact oracle (gradient)

From section 2.2 one may conclude that (A)STM is better than (A)GD. In terms of the number of iterations (calculations of $\nabla J\left(q\right)$ $\left(J\left(q\right)\right)$) this is indeed so. But the right criterion is the total number of arithmetic operations (a.o.). Unfortunately, (A)STM is more sensitive to the error in calculation of $\nabla J\left(q\right)$ $\left(J\left(q\right)\right)$ than (A)GD. Now we plane to say in more details about this issue, following to [9, 10]. First of all, let's denote that if $J\left(q\right)$ is convex (μ-strongly convex, $\mu \geq 0$ – see (3)) and (2) holds true then for all q_1, q_2

$$0 \leq \left(\frac{\mu}{2}\left\|q_2 - q_1\right\|_2^2 \leq\right) J\left(q_2\right) - J\left(q_1\right) - \left\langle \nabla J\left(q_1\right), q_2 - q_1\right\rangle \leq \frac{L}{2}\left\|q_2 - q_1\right\|_2^2.$$

Assume that for (A)STM at each point y^{k+1} we can observe only such approximate values of $J^\delta\left(y^{k+1}\right)$, $\nabla J^\delta\left(y^{k+1}\right)$ that (in ASTM instead of while{} condition in line 5 one should use the right part of the inequality below with $J\left(q^{k+1}\right) \to J^\delta\left(q^{k+1}\right)$, $\delta \to 2\delta$)

$$0 \leq \left(\frac{\mu}{2}\left\|q^{k+1} - y^{k+1}\right\|_2^2 \leq\right) J\left(q^{k+1}\right) - J^\delta\left(y^{k+1}\right) - \left\langle \nabla J^\delta\left(y^{k+1}\right), q^{k+1} - y^{k+1}\right\rangle \leq$$

$$\frac{L}{2}\left\|q^{k+1} - y^{k+1}\right\|_2^2 + \delta$$

then (A)STM (restart version in μ-strongly convex case) converges as (constants in $\mathrm{O}\left(\right)$ is smaller then 5)

$$J\left(q^N\right) - J\left(q_*\right) = \mathrm{O}\left(\min\left\{\frac{4LR^2}{N^2}, LR^2\exp\left(-\frac{N}{2}\sqrt{\frac{\mu}{2L}}\right)\right\}\right) + \mathrm{O}\left(N\delta\right).$$

Assume that for (A)GD at each point y^k we can observe only such approximate values of $J^\delta\left(y^k\right)$, $\nabla J^\delta\left(y^k\right)$ that (in AGD instead of while{} condition in line 5 one should use the right part of the inequality below with $J\left(y^{k+1}\right) \to J^\delta\left(y^{k+1}\right)$, $\delta \to 2\delta$)

$$0 \leq \left(\frac{\mu}{2}\left\|y^{k+1} - y^k\right\|_2^2 \leq\right) J\left(y^{k+1}\right) - J^\delta\left(y^k\right) - \left\langle \nabla J^\delta\left(y^k\right), y^{k+1} - y^k\right\rangle \leq$$

$$\frac{L}{2}\left\|y^{k+1} - y^k\right\|_2^2 + \delta$$

then (A)GD converges as (constants in $\mathrm{O}\left(\right)$ is smaller then 5)

$$J\left(q^N\right) - J\left(q_*\right) = \mathrm{O}\left(\min\left\{\frac{LR^2}{N}, LR^2\exp\left(-N\frac{\mu}{2L}\right)\right\}\right) + \mathrm{O}\left(\delta\right).$$

Typically in applications (to inverse problems [32]) we have to calculate conjugate operator A^* for calculation of $\nabla J\left(q\right) = A^*\left(Aq - b\right)$. These lead us rather often to the initial-boundary value problem for the linear system of partial differential equations. In the most of the cases we can solve this system only numerically by choosing properly small size of the grid τ. So we have $\delta = \mathrm{O}(\tau^p)$, where $p = 1, 2, ...$ corresponds to the order of the approximation (Chapter 4, [9]).[1] But the cost of calculation approximate values, say, $J^\delta(y^k)$, $\nabla J^\delta(y^k)$ also depends on τ like $\mathrm{O}(\tau^{-r})$, where $r = 1, 2, ...$ corresponds to the dimension of the problem (we restrict ourselves here by simple explicit-type scheme). The main problem here is that we can obtain only very rough estimations of the constants in the last two expressions $\mathrm{O}()$. Since that one can propose the following practice-aimed version of the mentioned above algorithms. For the desired accuracy ε we chose $\delta \sim \varepsilon$ for (A)GD and $\delta \sim \varepsilon\sqrt{\max\{\mu R^2, \varepsilon\}}$(A)STM (restart version) and after that $\tau \simeq C_{GD}\varepsilon^{1/p}$ for (A)GD and $\tau \simeq C_{STM} \cdot \left(\varepsilon\sqrt{\max\{\mu R^2, \varepsilon\}}\right)^{1/p}$(A)STM (restart version). The constant factors C_{GD}, C_{STM} are unknown. Since that it is proper to use restart on this constants. Start with $C_{GD} = 1$ and after

$$N \simeq 4\min\left\{\frac{LR^2}{\varepsilon}, \frac{L}{\mu}\ln\left(\frac{LR^2}{\varepsilon}\right)\right\}$$

iterations verify $J^\delta(q^N) \le \varepsilon$ or $J(q^N) \le \varepsilon$ if we can calculate $J(q)$ exactly ($J(q_*) = 0$). If $J^\delta(q^N) \le \varepsilon$ put $C_{GD} := C_{GD}/3$. Analogously for C_{STM} with

$$N \simeq 4\min\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \sqrt{\frac{L}{\mu}}\ln\left(\frac{LR^2}{\varepsilon}\right)\right\}.$$

For adaptive methods we can use here $L = \max\limits_{k=0,..,N} L_k^{j_k}$.

Note, that total number of a.o. for (A)GD can be estimated as follows

$$\mathrm{O}\left(\frac{1}{\varepsilon^{r/p}}\min\left\{\frac{LR^2}{\varepsilon}, \frac{L}{\mu}\ln\left(\frac{LR^2}{\varepsilon}\right)\right\}\right)$$

and for (A)STM (restart version) as follows

$$\mathrm{O}\left(\frac{1}{\left(\varepsilon\sqrt{\max\{\mu R^2, \varepsilon\}}\right)^{r/p}}\min\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \sqrt{\frac{L}{\mu}}\ln\left(\frac{LR^2}{\varepsilon}\right)\right\}\right).$$

It is hardly possible to say what is better from these estimations. As we've already mentioned – we don't know even the right order of the constant in $\mathrm{O}()$. But, roughly, we may expect that (A)GD works better for $r > p$ and (A)STM works better for $r < p$.

Note that in example 1 inexactness in some applications can also arise due to f [32, 11]. The mentioned above theory can be used also in this case.

Numerical experiments (fulfilled by Anastasia Pereberina) show that the described above approaches sometimes work very bad in practice due to the large values of L. But in these cases we can use another approach that typically allows us to win at least one or two order in the rate of convergence (these trick was proposed to us by A. Birjukov and A. Chernov). The idea is very simple: start with fixed L (say, $L = 1$) and run (non adaptive) algorithm. Then put $L := 2L$ and run algorithm with this

[1]This is a hypothesis. The alternative hypothesis is that $O(\tau^p)$ is an additive noise in the gradient (see section 4). This hypothesis seems to be more realistic. In this case the error doesn't accumulated up to a stopping time. So, due to the results of section 4 in any situation accelerated algorithms dominate non-accelerated ones, that was experimentally observed, see section 6.

parameter (initial/starting point is the same for all the starts). Repeat these restarts $L := 2L$ until we begin to observe stabilization (algorithm with L and $L := 2L$ converges in terms of functional values to the same limit). One can show that the total number of required calculations, that should be growth due to the fact that we don't know real L, growth at most for 8 times [12]. But typically we can win much more (than we lose due to restarts) because the methods starts to converges with the value of L that is much smaller than the real one.

2.4 Concluding remarks and examples

If instead of (1) we consider regularized problem

$$J^{\mu}(q) = J(q) + \frac{\mu}{2}\|q\|_2^2 \to \min_q \tag{7}$$

with arbitrary positive $\mu \le \varepsilon/(2R^2)$ and can find such q^N that

$$J^{\mu}\left(q^N\right) - \min_q J^{\mu}(q) \le \varepsilon/2,$$

then

$$J\left(q^N\right) - \min_q J(q) \le \varepsilon.$$

But the problem (7) is μ-strongly convex.

Note that estimation for GD and AGD $\mathrm{O}\left(LR^2/\varepsilon\right)$ can be obtained (up to a logarithmic factor) from $\mathrm{O}\left((L/\mu)\ln\left(2LR^2/(\mu\varepsilon^2)\right)\right)$ under $\mu \simeq \varepsilon/(2R^2)$ (see above), analogously for STM and ASTM $\mathrm{O}\left(\sqrt{LR^2/\varepsilon}\right)$ can be obtained (up to a logarithmic factor) from $\mathrm{O}\left(\sqrt{L/\mu}\ln\left(2LR^2/(\mu\varepsilon^2)\right)\right)$.

All the results mentioned above can be generalized for the convex optimization problem in reflexive Banach space with Lipchitz continuous functional

$$J(q) \to \min_{q\in Q}.$$

Not that the set Q is assumed to be of simple structure that allows to "project" on it efficiently. The proper generalization can be found in [4, 5] (see also [9] for (A)GD and (A)$\overline{\text{GD}}$ – these methods can be applied also to non convex problems, see [13, 14]).

Example 2. (convex optimal control problems) Let's consider the following optimal control problem $(q \equiv u(\,\cdot\,))$

$$J(u(\,\cdot\,)) = \int\limits_0^T f^0(t, x(t), u(t))\,dt + \Phi(x(T)) \to \min_{u(\,\cdot\,)\in U,\; u(\,\cdot\,)\subseteq L_2^m[0,T]},$$

$$\frac{dx}{dt} = f(t, x(t), u(t)), \quad x(0) = x^0, \quad (DE)$$

where U is a convex set in $\mathbb{R}^m$ $(U \equiv Q)$, all the functions are smooth enough (Chapter 8, [1]), $f(t, x, u)$ is a linear functional of (x, u) with coefficients depend only on t and functional $f^0(t, x, u)$

is convex on (x, u). In this case $J(u(\cdot))$ is convex functional [15]. Due to § 5, Chapter 8 [1][2]

$$\nabla J(u(\cdot)) = \left.\frac{\partial H(t,x,u,\psi)}{\partial u}\right|_{x=x(t,u),\ u=u(t),\ \psi=\psi(t,u)} =$$

$$H_u(t, x(t, u(t)), u(t), \psi(t, u(t))),$$

where $H = f^0 + \langle \psi, f\rangle$, $x(t,u)$ is the solution of (DE) and $\psi(t,u)$ is the solution of the conjugate system

$$\frac{d\psi}{dt} = -\frac{\partial H(t,x,u,\psi)}{\partial x}, \quad \psi(T) = \nabla\Phi(x(T,u)). \quad (CDE)$$

Unfortunately, one can't calculate precisely gradient since one should solve two system of ordinary differential equations (DE), (CDE). But one can solve these two systems by introducing the same lattice (it is significant [2]) in t, with the size of each element τ: $t^{k+1} - t^k \equiv \tau$, for both of the systems (DE), (CDE)

$$\frac{x\left(t^{k+1}\right) - x\left(t^k\right)}{\tau} = f\left(t^k, x\left(t^k\right), u\left(t^k\right)\right), \quad x\left(t^0\right) = x(0) = x^0,$$

$$\frac{\psi(t^k) - \psi(t^{k+1})}{\tau} = \frac{\partial H}{\partial x}(t^{k+1}, x(t^{k+1}), u(t^{k+1}), \psi(t^{k+1})),$$

$$\psi(T) = \nabla\Phi(x(t^{T/h})).$$

Here we use the standard Euler's scheme [16] with the quality of approximation $\delta \simeq \tau e^{cT}$ (i.e. $\delta \sim \tau$) and the complexity of one iteration is $\sim \tau^{-1}$ (in terms of remark 3 $r = 1$ and $p = 1$). So using ASTM with proper choice of $\tau \sim \varepsilon^{3/2}$ one can find ε-solution with the total complexity $\sim \varepsilon^{-2}$. The same result (about the total complexity $\sim \varepsilon^{-2}$) is true for AGD. But proper modification of the last method works also with non convex problems (find local extreme).

Note, that if $f^0(t,x,u)$ is also linear functional of (x,u) with coefficients depend only on t then

$$\frac{\partial H(t,x,u,\psi)}{\partial x} \equiv h_0(t) + h_1(t)\psi.$$

Since that instead of Euler's scheme one can use Runge–Kutta's schemes of order $p \geq 2$ [16]. The complexity of one iteration is still $\sim \tau^{-1}$ $(r = 1)$, therefore (see section 2.3) we may expect that ASTM will work better than AGD. ■

3 Dual approaches

Now we concentrate on example 1. The described below approaches goes back to the Yu.E. Nesterov and A.S. Nemirovski (see historical notes in [17, 18]).

Assume that we have to solve the following convex optimization problem

$$g(q) \to \min_{Aq=f}, \tag{8}$$

[2]This means that for all small enough $h(\cdot) \in L_2^m[0,T]$ the following holds true
$J(u(\cdot) + h(\cdot)) - J(u(\cdot)) = \langle \nabla J(u(\cdot)), h(\cdot)\rangle_{L_2^m[0,T]} + \mathrm{O}\left(\|h(\cdot)\|^2_{L_2^m[0,T]}\right) =$
$= \int\limits_0^T H_u(t, x(t, u(t)), u(t), \psi(t, u(t)))h(t)\,dt + \mathrm{O}\left(\int\limits_0^T h(t)^2 dt\right).$

where $g(q)$ is 1-strongly convex in H_1. We build the dual problem

$$\phi(\lambda) = \max_q \{\langle \lambda, f - Aq \rangle - g(q)\} = \langle \lambda, f - Aq(\lambda) \rangle - g(q(\lambda)) \to \min_\lambda . \tag{9}$$

Note, that $\nabla \phi(\lambda) = f - Aq(\lambda)$.

Let (A)STM with $y^0 = 0$ for the problem (9) generates points $\{y^k\}_{k=0}^N$, $\{u^k\}_{k=0}^N$ and $\{\lambda^k\}_{k=0}^N$ (in (A)STM we denote the last ones by $\{q^k\}_{k=0}^N$). Put

$$q^N = \sum_{k=0}^N \frac{\alpha_k}{A_N} q(y^k).$$

Let q_* be the solution of (8) (this solution is unique due to strong convexity of $g(q)$). Then

$$g(q^N) - g(q_*) \le \phi(\lambda^N) + g(q^N).$$

The next theorem [17, 18] allows us to calculate the solution of (8) with prescribed precision.

Theorem 3.1. *Assume that we want to solve the problem* (8) *by passing to the dual problem* (9)*, according to the formulas mentioned above. Let's use (A)STM to solve* (9) *with the following stopping rule*

$$\phi(\lambda^N) + g(q^N) \le \varepsilon, \quad \|Aq^N - f\|_{H_2} \le \tilde{\varepsilon}.$$

Then (A)STM stops by making no more than

$$6 \cdot \max\left\{ \sqrt{\frac{L\breve{R}^2}{\varepsilon}}, \sqrt{\frac{L\breve{R}}{\tilde{\varepsilon}}} \right\} \tag{10}$$

iterations, where $L = \|A^*\|^2_{H_2 \to H_1} = \|A\|^2_{H_1 \to H_2}$, $\breve{R} = \|\lambda_*\|_{H_2}$, λ_* *– solution of the problem* (9) *(if the solution is not unique than we can choose such a solution* λ_* *that minimize* $\breve{R}$*).*

For ASTM the average number of calculations of $\phi(\lambda)$ *per iteration roughly equals 4 and the average number of calculations Frechet derivative* $\nabla\phi(\lambda) = f - Aq(\lambda)$ *per iteration roughly equals 2.*

Example 3. (see [17, 19]) Let us consider the following optimization problem

$$\frac{1}{2}\|q\|^2_{H_1} \to \min_{Aq=f}.$$

One can built the dual one

$$\min_{Aq=f} \frac{1}{2}\|q\|^2_{H_1} = \min_q \max_\lambda \left\{ \frac{1}{2}\|q\|^2_{H_1} + \langle f - Aq, \lambda \rangle \right\} =$$

$$= \max_\lambda \min_q \left\{ \frac{1}{2}\|q\|^2_{H_1} + \langle f - Aq, \lambda \rangle \right\} = \max_\lambda \left\{ \langle f, \lambda \rangle - \frac{1}{2}\|A^*\lambda\|^2_{H_1} \right\}. \tag{11}$$

We assume that $Aq = f$ is compatible, hence for the Fredgolm's theorem it's not possible that there exists such a λ: $A^*\lambda = 0$ and $\langle b, \lambda\rangle > 0$.[3] Hence the dual problem is solvable (but the solution isn't necessarily unique). Let's denote λ_* to be the solution of the dual problem

$$\phi(\lambda) = \frac{1}{2}\|A^*\lambda\|_{H_1}^2 - \langle f, \lambda\rangle \to \min_{\lambda}$$

with minimal H_2-norm. Let's introduce (from the optimality condition in (11) for q): $q(\lambda) = A^*\lambda$. Using (A)STM for the dual problem one can find (Theorem 3)

$$\left\|Aq^N - f\right\|_{H_1} = \mathrm{O}\left(\frac{L\breve{R}}{N^2}\right), \tag{12}$$

where $L = \|A^*\|_{H_2\to H_1}^2 = \|A\|_{H_1\to H_2}^2$ (as in example 1), $\breve{R} = \|\lambda_*\|_{H_2}$.
If one will try to solve the primal problem in example (1)

$$\frac{1}{2}\|Aq - f\|_{H_2}^2 \to \min_q$$

by (A)STM, one can obtain the following estimate

$$\left\|Aq^N - f\right\|_{H_2} = \mathrm{O}\left(\frac{\sqrt{L}R}{N}\right), \tag{13}$$

where $L = \|A\|_{H_1\to H_2}^2$, $R = \|q_*\|_{H_1}$. Estimate (13) seems worse than (12). But estimate (13) cannot be improving up to a constant factor [7]. There is no contradiction here, since in general $\breve{R}$ can be big enough, i.e. this parameter is uncontrollable. But in real applications we can hope that this (dual) approach lead us to a faster convergence rate (12). ■

Indeed, all the mentioned above methods (expect (A)GD) are primal-dual ones [17, 18, 19] (if we use their non strongly convex variants). That is for these methods analogues of theorem 3 holds true with proper modification of (10) for (A)$\overline{\text{GD}}$

$$3\cdot\max\left\{\frac{L\breve{R}^2}{\varepsilon}, \frac{L\breve{R}}{\tilde{\varepsilon}}\right\}.$$

This means that we can apply the results of section 2 for this approach (with the same sensitivity results as in section 2.3). Moreover we can also generalize the problem formulation (8) for more general class of the problems [17, 19] (compare this with section 2.4)

$$g(q) \to \min_{Aq=f,\, q\in Q}, \tag{14}$$

where q belongs to reflexive Banach space (with norm $\|\ \|$) and $g(q)$ is 1-strongly convex in norm $\|\ \|$. The dual problem has the form

[3] Indeed, if there exists such q that $Aq = f$ then for all λ: $\langle Aq, \lambda\rangle = \langle f, \lambda\rangle$. Hence, $\langle q, A^*\lambda\rangle = \langle f, \lambda\rangle$. Assume that there exists such a λ, that $A^*\lambda = 0$ and $\langle f, \lambda\rangle > 0$. If it is so we observe a contradiction:

$$0 = \langle q, A^*\lambda\rangle = \langle f, \lambda\rangle > 0.$$

$$\phi\left(\lambda\right)=\max_{q\in Q}\left\{\left\langle\lambda,f-Aq\right\rangle-g\left(q\right)\right\}=\left\langle\lambda,f-Aq\left(\lambda\right)\right\rangle-g\left(q\left(\lambda\right)\right)\to\min_{\lambda}. \quad (9)$$

Let's consider another approach to solve problem (14) [12, 17]. This approaches based on the section 2.4. We regularize the dual problem (9) (we use $q^N=q\left(\lambda^N\right)$ for the solution of (14))

$$\phi^{\mu}\left(\lambda\right)=\phi\left(\lambda\right)+\frac{\mu}{2}\left\|\lambda\right\|_{H_2}^2\to\min_{\lambda},$$

where $\mu\simeq\frac{\varepsilon}{2\breve{R}^2}$. Since we don't know $\breve{R}$ (and therefore μ) we may use restart technique on μ (see remark 1 and 3 and restarts for C_{GD}). Let q_* be the solution of (14). Since ($q\left(\lambda\right)$ is determine by (9))

$$g\left(q\left(\lambda\right)\right)+\left\langle\lambda,Aq\left(\lambda\right)-f\right\rangle\leq g\left(q_*\right)$$

we have

$$g\left(q\left(\lambda\right)\right)-g\left(q_*\right)\leq\left\|\lambda\right\|_{H_2}\left\|Aq\left(\lambda\right)-f\right\|_{H_2}.$$

The next theorem [12, 17] allows us to calculate the solution of (14) with prescribed precision.

Theorem 3.2. *Assume that we want to solve the problem* (14) *by passing to the dual problem* (9)*, according to the formulas mentioned above. Let's use (A)STM to solve* (9) *with the following stopping rule*

$$\left\|\lambda^N\right\|_{H_2}\left\|Aq\left(\lambda^N\right)-f\right\|_{H_2}\leq\varepsilon,\quad\left\|Aq\left(\lambda^N\right)-f\right\|_{H_2}\leq\tilde{\varepsilon}.$$

Then (A)STM stops by making no more than

$$N\simeq 2\sqrt{\frac{L\cdot\left(\varepsilon+2\breve{R}\tilde{\varepsilon}\right)}{\tilde{\varepsilon}^2}}\ln\left(\frac{8L\max\limits_{q_1,q_2\in Q}\left|g\left(q_2\right)-g\left(q_1\right)\right|\cdot\left(\varepsilon+2\breve{R}\tilde{\varepsilon}\right)}{\varepsilon\cdot\tilde{\varepsilon}^2}\right)$$

iterations.

For ASTM the average number of calculations of $\phi^{\mu}\left(\lambda\right)$ per iteration roughly equals 4 and the average number of calculations Frechet derivative $\nabla\phi^{\mu}\left(\lambda\right)=f-Aq\left(\lambda\right)+\mu\lambda$ per iteration roughly equals 2.

Note that one can spread most of the result of section 2 for this approach too.

Now let's describe the main motivating example for this paper.

Example 4 (inverse problem for elliptic initial-boundary value problem). Let u be the solution of the following problem (P)

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

$$u_x\left(0,y\right)=0,\quad y\in\left(0,1\right),$$

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

$$u\left(x,0\right)=u\left(x,1\right)=0,\quad x\in\left(0,1\right).$$

And corresponding dual problem (D)

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

$$\psi_x\left(0,y\right)=\lambda\left(y\right),\quad y\in\left(0,1\right),$$

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

Let's introduce operator

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

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

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

Conjugate operator

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

Here $\psi(x,y)$ is the solution of problem (D) [32]. To obtain these formulas one may use the general approach, described, for example, in § 7, Chapter 8 [1] (see also Chapter 4 [2]).

Let us formulate inverse problem [32]: find the function q by known additional information

$$u(0,y)=f(y).$$

Inverse problem is reduced to the optimization problem of the following cost functional

$$J(q)=\|Aq-f\|_{L_2(0,1)}\to\min_q,$$

$$\|q\|_H^2\to\min_{Aq=f}\quad //\quad \phi(\lambda)=\|A^*\lambda\|_H^2-\langle f,\lambda\rangle\to\min_\lambda.$$

It is obvious that $\nabla J(q)=A^*(Aq-f)$ and $\nabla J(q)$ can be found by following formula:

$$\nabla J(q)(y)=\psi_x(1,y).$$

Here $\psi(x,y)$ is the solution of (D) with $\lambda(y)=2(u(0,y)-f(y))$.

For example 3 one can obtain that $\nabla\phi(\lambda)=2(f-A(A^*\lambda))$, $\nabla\phi(\lambda)\in L_2(0,1)$ and

$$\nabla\phi(\lambda)(y)=2(f(y)-u(0,y)).$$

Note, that for this example $L=1$ [32], see (2) for definition of L. ■

4 Stopping rule for STM

Let us consider $STM(y^0,L)$ in the following conception of inexact oracle [27]: for all $q_1,q_2\in H$

$$\|\nabla J(q_1)-\tilde\nabla J(q_2)\|_2\leqslant\tilde\delta.$$

Using inequality (2), convexity of J and Fenchel inequality we can get, that: for all $q_1,q_2\in H$

$$J(q_1)\leqslant J(q_2)+\langle J(q_2),q_1-q_2\rangle+\frac{2L}{2}\|q_1-q_2\|_2^2+\frac{\tilde\delta^2}{2L},$$
$$J(q_2)+\langle\tilde\nabla J(q_2),q_1-q_2\rangle-\tilde\delta\|q_1-q_2\|_2\leqslant J(q_1).$$

If we introduce:

$$\psi_k(q) = \frac{1}{2}\|q - y^0\|_2^2 + \sum_{j=0}^{k} \left(J(y^j) + \langle \tilde{\nabla} J(y^j), q - y^j \rangle \right).$$

It can be shown, that in general degenerate situation (not strongly convex case) the following estimates hold true [44, 23]:[4]

$$\begin{gathered} A_k J(q_k) \leqslant \psi_k(u_k) + \frac{\tilde{\delta}^2}{2L} \sum_{j=0}^{k} A_j + \tilde{\delta} \sum_{j=1}^{k} \alpha_j \|y^j - u^{j-1}\|_2^2, \\ J(q^N) - J(q_*) \leqslant \frac{4LR^2}{N^2} + 3\tilde{R}\tilde{\delta} + N\frac{\tilde{\delta}^2}{2L}, \\ \tilde{R} = \max_{k \leqslant N}\{\|q_* - y^k\|_2, \|q_* - u^k\|_2, \|q_* - q^k\|_2\}. \end{gathered} \tag{15}$$

Note, that if we know, that $\|u^k - q_*\|_2 \leqslant R$ we can easily show that $\|y^k - q_*\|_2 \leqslant R$ and $\|q^k - q_*\|_2 \leqslant R$:

$$\|y^k - q_*\|_2 \leqslant \frac{A_{k-1}}{A_k}\|q^{k-1} - q_*\|_2 + \frac{\alpha_k}{A_k}\|u^{k-1} - q_*\|_2 \leqslant R.$$

Similarly for the sequence q^k. Therefore, we show how, using the stopping criterion, to obtain this inequality for the sequence u^k. If we know the value of $J(q_*)$ and such bound $R_* > 0$, that $\|q_*\|_2 \leqslant R_*$. Then by choosing $y^0 = 0$ (obviously, that in this case $R \leq R_*$), we can formulate a computable stopping criterion: for all $\zeta > 0$:

$$J(q_k) - J(q_*) \leqslant k\frac{\tilde{\delta}^2}{2L} + 3R_*\tilde{\delta} + \zeta.$$

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

$$\begin{gathered} A_k J(q^k) + \frac{1}{2}\|u^k - q_*\|_2^2 \leqslant \frac{1}{2}\|u^k - q_*\|_2 + \psi_k(u^k) + \frac{\tilde{\delta}^2}{2L}\sum_{j=0}^{k} A_j + \\ + \tilde{\delta}\sum_{j=1}^{k}\alpha_j\|\tilde{q}^j - u^{j-1}\|_2 \leqslant \psi_k(q_*) + \tilde{\delta}\sum_{j=1}^{k}\alpha_j\|\tilde{q}^j - u^{j-1}\|_2 \leqslant \frac{\tilde{\delta}^2}{2L}\sum_{j=0}^{k} A_j + \\ + \frac{1}{2}R^2 + A_k J(q_*) + \tilde{\delta}\sum_{j=1}^{k}\alpha_j\|\tilde{q}^j - u^{j-1}\|_2 + \tilde{\delta}\sum_{j=0}^{k}\alpha_j\|u^k - q_*\|_2 \leqslant \\ \leqslant \frac{1}{2}R^2 + A_k 3\tilde{\delta}R_* + \frac{\tilde{\delta}^2}{2L}\sum_{j=0}^{k} A_j + A_k J(q_*) \Rightarrow \\ \Rightarrow \frac{1}{2}(R^2 - \|u^k - q_*\|_2) \geqslant A_k\left(\left(J(q^k) - J(q_*)\right) - \left(k\frac{\tilde{\delta}^2}{2L} + 3R_*\tilde{\delta} + \zeta\right)\right) \geqslant 0. \end{gathered}$$

[4]Recall that $R = \left\|q_* - y^0\right\|_2$.

Also this criterion is achievable:

$$J(q^N) - J(q_*) \leqslant \frac{4LR^2}{N^2} + 3R_*\tilde{\delta} + N\frac{\tilde{\delta}^2}{2L},$$
$$\frac{4LR^2}{N^2} \leqslant \zeta,$$
$$N \geqslant 2\sqrt{\frac{LR^2}{\zeta}},$$
$$N = O\left(\sqrt{\frac{LR^2}{\zeta}}\right).$$

That is N iterations is enough to reach the stopping criterion. Finally we get the following theorem:

Theorem 4.1. *Assume that we solve problem* (1) *and we know value $J(q_*)$ and the bound $R_* > 0$ for $\|q_*\|_2$. Using STM(0, L) with stopping rule:*

$$J(q^k) - J(q_*) \leqslant \frac{\tilde{\delta}^2}{2L}k + 3R_*\tilde{\delta} + \zeta.$$

We get estimation:

$$\tilde{R} \leqslant R,$$
$$\tilde{R} = \max_{k \leqslant N}\{\|q_* - y^k\|_2, \|q_* - u^k\|_2, \|q_* - q^k\|_2\}.$$

And it is guaranteed, that the criteria will reached in:

$$N = O\left(\sqrt{\frac{LR^2}{\zeta}}\right).$$

Using this theorem we can get, that solving problem:

$$J(q^N) - J(q_*) \leqslant \varepsilon. \tag{16}$$

We can choose $\zeta \sim \varepsilon$ and $\tilde{\delta} \sim \frac{\varepsilon}{R_*}$ and number of iterations will be estimated as

$$N = O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right). \tag{17}$$

Similar results can be formulated for all other methods from section 2 including adaptive ones, since (15) holds true for this generality too [23]. As an example, we remark that for GD (16) takes place with the same requirement for $\tilde{\delta} \sim \frac{\varepsilon}{R_*}$, but with a worse bound on

$$N = O\left(\frac{LR^2}{\varepsilon}\right). \tag{18}$$

From this we may expect that with the same level of noise in the gradient $\tilde{\delta}$ non accelerated algorithms (STM) reach the same quality $J(q^N) - J(q_*) \simeq \tilde{\delta}R$ as non-accelerated ones (GD), but they do it faster (compare (17) and (18)). Therefore we may expect that with proper stopping rule STM must outperform GD for degenerate (non strongly convex) problems. That was confirmed in the numerical experiments described in section 6.

5 Application to the continuation problem for Laplace equation

Let us consider the following continuation problem for an elliptic equation:

$$u_{xx} + L(y)u = 0, \qquad (x, y) \in \Omega, \tag{19}$$
$$u(0, y) = f(y), \qquad y \in \mathcal{D}, \tag{20}$$
$$u_x(0, y) = 0, \qquad y \in \mathcal{D}, \tag{21}$$
$$u(x, y) = 0, \qquad x \in (0, h), \quad y \in \partial\mathcal{D} \tag{22}$$

with the matching conditions

$$f(y) = 0, \qquad y \in \partial\mathcal{D}. \tag{23}$$

Here

$$\Omega = (0, h) \times \mathcal{D}, \mathcal{D} \in \mathbb{R}^n$$

is the bounded domain with a Lipschitz boundary $\partial\mathcal{D}$,

$$L(y)u = \sum_{i,j=1}^{n} \frac{\partial}{\partial y_i} \left(a_{ij}(y) \frac{\partial u}{\partial y_j} \right) - c(y)u, \tag{24}$$

$$M_4 \sum_{j=1}^{n} \nu_j^2 \leq \sum_{i,j=1}^{n} a_{ij}(y) \nu_i \nu_j,$$

$$\forall \nu_i \in \mathbb{R}, \quad a_{ij} = a_{ji}, \quad i, j = 1, \ldots, n,$$

$$0 \leq c(y) \leq M_5,$$

$$a_{ij} \in C^1(\overline{\mathcal{D}}), \qquad c \in C(\overline{\mathcal{D}}).$$

Let us consider the ill-posed continuation problem (19)–(23) as the inverse problem of the following direct problem:

$$u_{xx} + L(y)u = 0, \qquad (x, y) \in \Omega, \tag{25}$$
$$u_x(0, y) = 0, \qquad y \in \mathcal{D}, \tag{26}$$
$$u(h, y) = q(y), \qquad y \in \mathcal{D}, \tag{27}$$
$$u(x, y) = 0, \qquad x \in (0, h), \quad y \in \partial\mathcal{D} \tag{28}$$

with the matching conditions:

$$q(y) = 0, \qquad y \in \partial\mathcal{D}. \tag{29}$$

In the direct problem (25)–(29) one has to find $u(x, y)$ in the domain Ω for the function $q(y)$ set for a part of the boundary $x = h$ of the domain Ω.

The inverse problem is to determine $q(y)$ from conditions of (25)–(29) and known additional information

$$u(0, y) = f(y), \qquad y \in \partial\mathcal{D}. \tag{30}$$

To familiarize yourself with some the results based on the theory of direct and inverse problems address to [32].

Let us apply a gradient method to solve the continuation problem for elliptic equation. For that purpose consider the adjoint problem:

$$\psi_{xx} + L(y)\psi = 0, \qquad (x, y) \in \Omega, \tag{31}$$
$$\psi_x(0, y) = \mu(y), \qquad y \in \mathcal{D}, \tag{32}$$
$$\psi(h, y) = 0, \qquad y \in \mathcal{D}, \tag{33}$$
$$\psi|_{\partial\mathcal{D}} = 0, \qquad x \in (0, h). \tag{34}$$

The problem consists in finding the function $\psi(x, y)$ using given $\mu(y)$.

We introduce an operator

$$A : q(y) \to u(0, y),$$

where $u(x, y)$ is the solution of the direct problem (25)–(29).

Therefore, the adjoint operator A^* is expressed as

$$A^* : \mu(y) \to \psi_x(h, y),$$

where $\psi(x, y)$ is a solution of the adjoint problem (31)–(34).

It follows from [32, 11] that operators A and A^* map $L_2(\mathcal{D})$ to $L_2(\mathcal{D})$ and that the solution is unique and conditional stability estimation holds true of the continuation problem (19)–(23). Therefore, the inverse problem (25)–(30) can be written in the operator form

$$Aq = f. \tag{35}$$

To find the solution (35) we apply gradient method. It should be noted that the gradient of functional $J'q$ is calculated based on the formula:

$$\big(J'q\big)(y) = \psi_x(h, y), \tag{36}$$

where $\psi(x, y)$ is a solution of the adjoint problem (31)–(34), in which

$$\mu(y) = 2\big[u(0, y) - f(y)\big].$$

Theorem 5.1. *Let the problem $Aq = f$ have the exact solution $q_T \in L_2(\mathcal{D})$. Let $\|f - f^\delta\| \le \delta$ and $\{q^n_\delta\}$ be an Landweber iteration sequence to solve the inverse problem* (25)–(30) *with the additional information $u^n_\delta(0, y) = f_\delta(y)$ Then, to solve the corresponding direct problem* (25)–(29) *the following estimation should be carried out [11]:*

$$\int\limits_{\mathcal{D}} (u^\delta_n(x, y) - u_T(x, y))^2 \, \mathrm{d}y \le M_{13}\Big(\beta(n)\delta + n^{\frac{x-h}{h}}\Big), \quad x \in (0, h). \tag{37}$$

Here $u_T \in L_2(\Omega)$ is the exact solution and

$$\beta(n) = \frac{(1 + 2\alpha\|A\|^2)^{n-1} - 1}{\|A\|}.$$

Analogous results have been obtained for steepest descent and conjugate gradients methods [32, 11].

The estimation (37) shows that the sequence $\{u^\delta_n\}$ is regularizing one where n is the regularization parameter. Actually, due to the fact that the fist member is going monotonously to infinity, while the second in the same way to zero, at $n \to \infty$, the stopping criterion for the corresponding number of iterations n_* can be selected based on the following rule. Having differentiated the right part (37) with respect to n, one finds the root n_r of the following equation:

$$\delta\frac{\ln(1 + 2\alpha\|A\|^2)}{\|A\|}(1 + 2\alpha\|A\|^2)^{n-1} + \frac{x-h}{h}n^{\frac{x-2h}{h}} = 0 \tag{38}$$

and can select the stopping number n_s to be a natural number closest to the equation root (38).

6 Numerical results

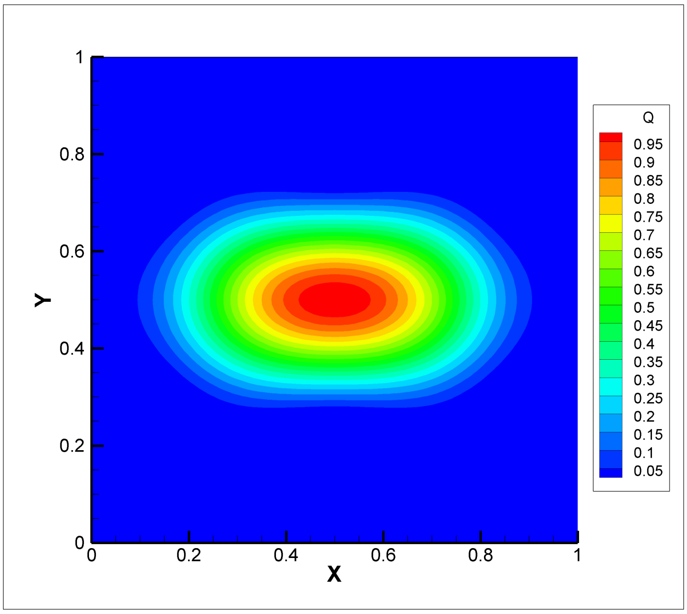


Figure 1: Test1 - True solution

We consider the following continuation problem in the domain $\Omega = \{(x, y, z) \in [0, 1]^2 \times [0, H]\}$:

$$\Delta u(x, y, z) \equiv u_{xx} + u_{yy} + u_{zz} = h(x, y, z),\ (x, y, z) \in \Omega$$
$$u|_{x=0} = u|_{x=1} = u|_{y=0} = u|_{y=1} = 0,$$
$$u_z|_{z=0} = 0, u|_{z=H} = q(x, y)$$

The problem is to determine the unknown function $q(x, y)$ by using the additional information of the function $u(x, y, z)$ on the boundary $z = 0$:

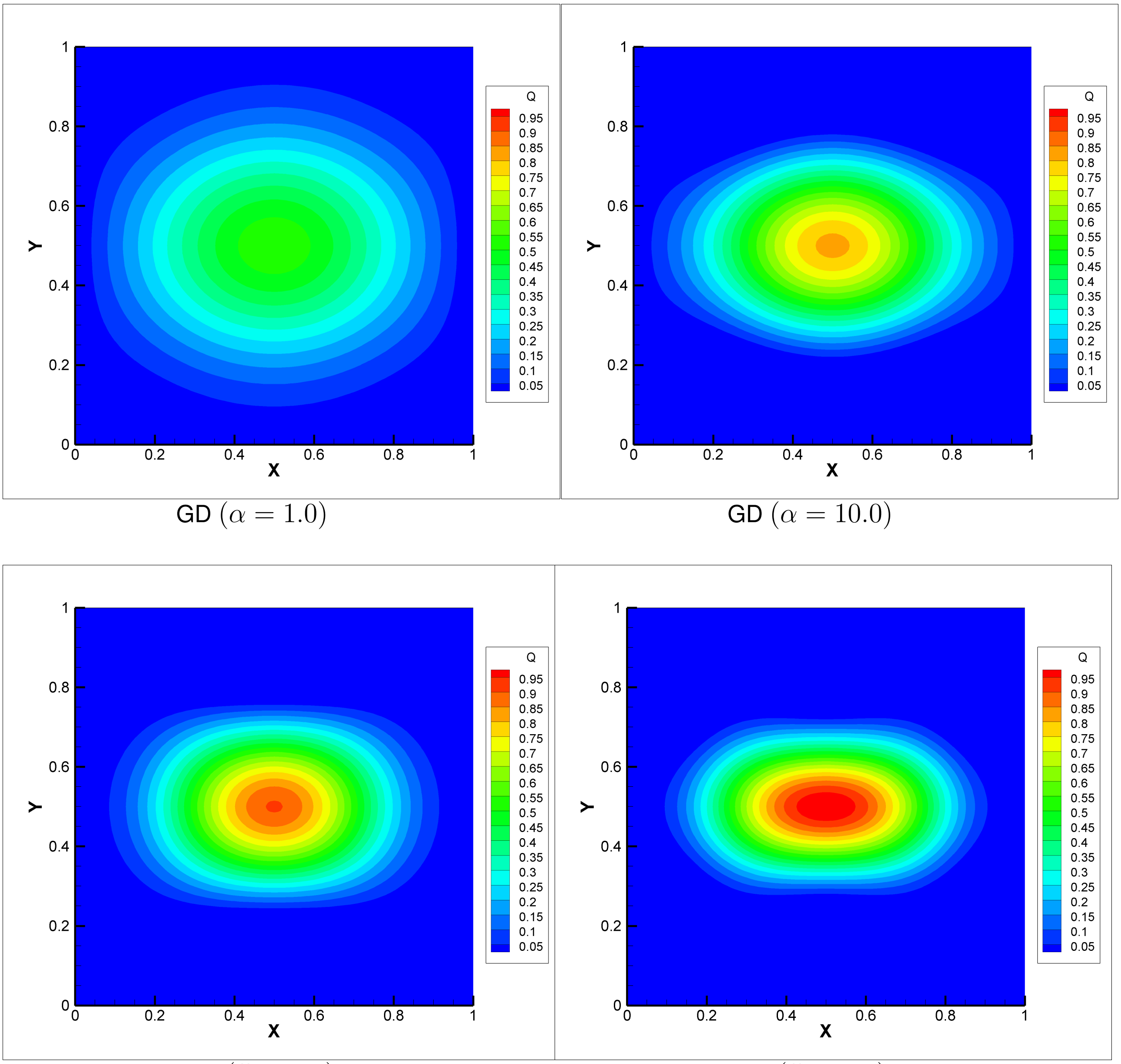
$$u|_{z=0} = f(x, y).$$

We solve the problem formulated by using different versions of STM and GD methods. The structure of the gradient of the functional was mentioned in the previous chapter and has the form (36). We choose the test solution as follows:

$$q(x, y) = \begin{cases} e^{l_1(x)+l_2(y)}, (x, y) \in [0.1, 0.9] \times [0.3, 0.7] \\ 0, \text{if else} \end{cases}$$

Here $l_1(x) = 1 + \frac{0.16}{(x-0.5)^2-0.16}, l_2(y) = 1 + \frac{0.04}{(y-0.5)^2-0.04}$. The structure of this function is presented on the figure 1. We solve the direct problem, using $q(x, y)$ as given function, to calculate true data $f(x, y)$.

For the first series of tests we use the following parameters: $H = 0.5, N_{iter} = 1000$. We consider the similar triangles method, simple gradient descent method and steepest descent method. We used initial approximation $q(x, y) = 0$ for all methods. Due to the fact, that it is hard to get the accurate estimation for the norm of the operator, we couldn't get the precise values for the parameters L, α of the STM and GD methods correspondingly during the numerical solution. Thus, we choose the parameters of STM and GD methods by trials and errors. However, in case of the homogeneous right hand side



GD $(\alpha = 1.0)$ GD $(\alpha = 10.0)$

STM $(L = 1.0)$ STM $(L = 0.1)$

Figure 2: Test 1 - solution of continuation problem by GD and STM methods

and boundary conditions we use analytic expression for the descent parameters of steepest descent method: $\alpha_n = \frac{|J'(q_n)|^2}{2|A(J'(q_n))|^2}$. The results of computations are presented on figures 2 – 4.

The similar triangles method provides the most efficient results of the considered methods. The steepest descent methods converges faster on the first iterations, but eventually the STM method provides better results in terms of both the residual and errors. The accuracy of the methods is acceptable (if suitable parameters of the methods were chosen). In order to illustrate the influence of the parameter L on the problem, we considered two different values of the parameter of STM method during this experiment.

For the second series of tests we added some non-homogeneous boundary conditions, and the right hand side of the following form:

$$h(x, y, z) = (1 - z)cos(\pi x)cos(\pi y)$$

We increased the depth to $H = 1.0$. The structure of the function $q(x, y)$ remains the same. However,

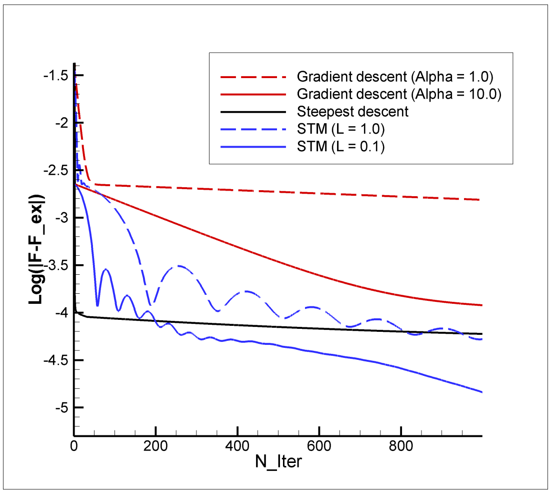


Figure 3: Test 1 - The residual functional (logarithmic scale)

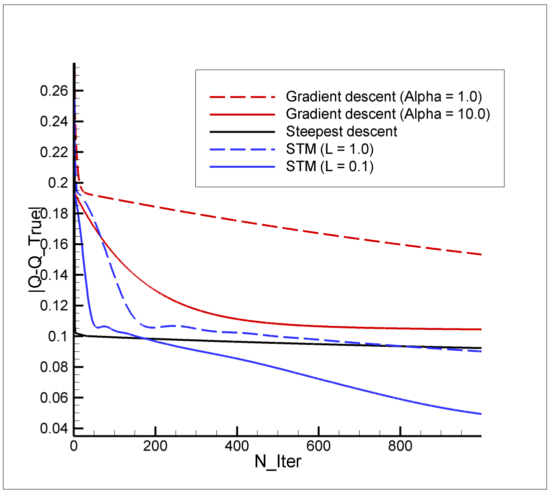


Figure 4: Test 1 - The errors of the GD, SGD, STM methods

the increased depth significantly decreases the influence of data, that we have during the experiments with synthetic data, that we balance by increasing the number of iterations to $N_{iter} = 16000$. The behavior of the functional is presented on the figures 5, 6. This allows us to provide the solution, almost identical to exact one. During the last series of test we considered the medium number of iterations

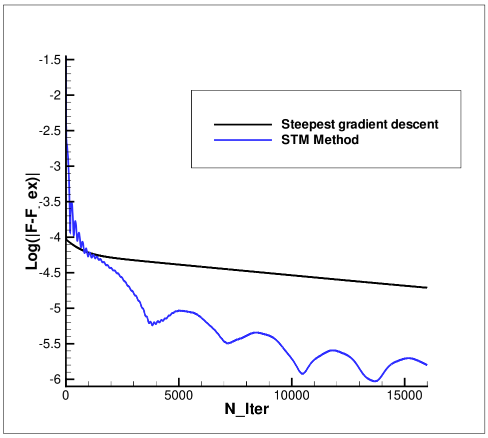


Figure 5: Test 3 (increased number of iterations) - The residual functional (logarithmic scale)

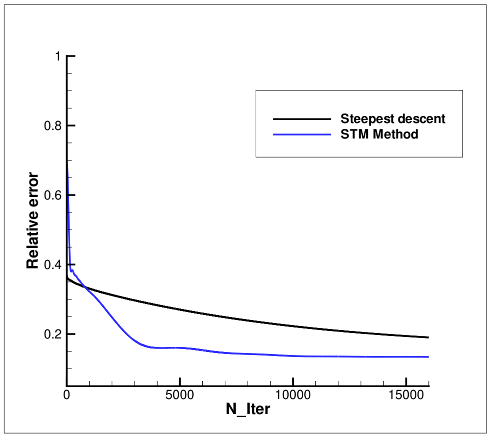
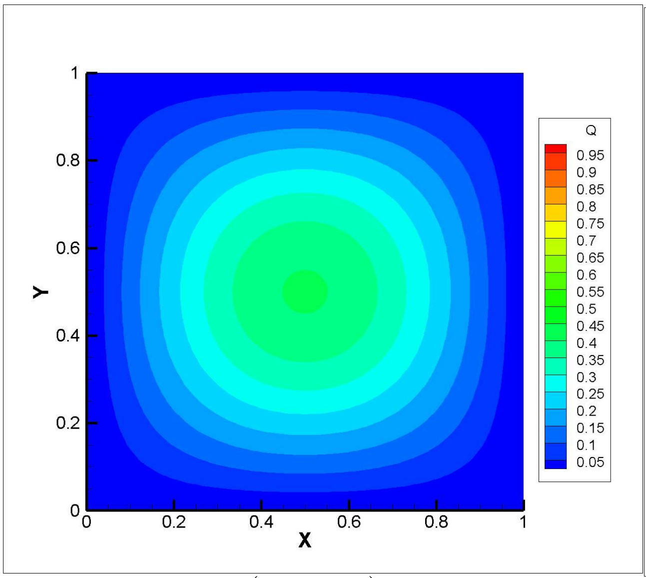
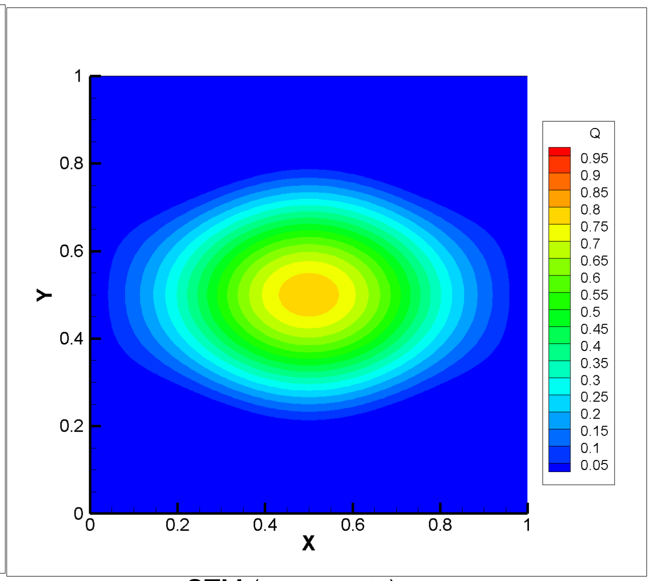


Figure 6: Test 3 (increased number of iterations) - The errors of the methods

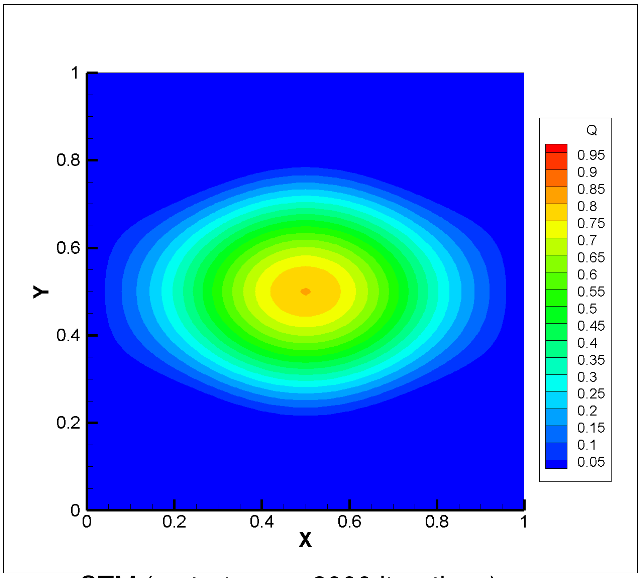
$N_{iter} = 6000$ and the depth $H = 1.25$ to study the variation of STM method with restarts. The computational results are presented on figures 7 – 9. We notice, that similar triangles method provides significantly better results, compared to simple gradient descent. The usage of restart technique allows to obtain better results in terms of residual, but the effects of restarts are much less significant in terms of errors. The cause of this difference is the ill-posedness of the problem, which becomes more noticeable with the increase of depth.



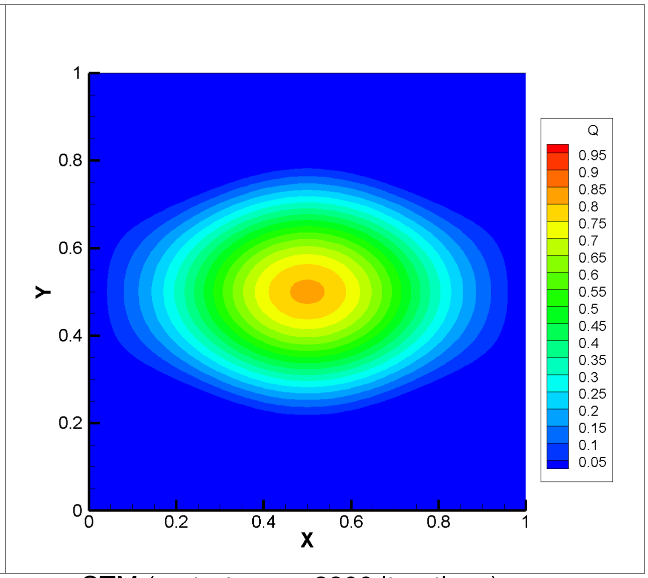
GD $(\alpha = 100)$



STM (no restarts)



STM (restart every 2000 iterations)



STM (restart every 2800 iterations)

Figure 7: Test 2 (increased depth) - solution of continuation problem by GD and STM methods.

References

[1] Vasiliev F.P., Optimization methods, MCCME, Moscow (2011) (in Russian)

[2] Evtushenko Yu.G., Optimization and fast automatic differentiation. Preprint CCAS (2013) (in Russian)

[3] Halmos P., A Hilbert space problem book, Springer (1982)

[4] Gasnikov A.V., Nesterov Yu.E., Universal fast gradient method for stochastic composite optimization problems, Comp. Math. & Math. Phys., V. 58 (2018) arXiv:1604.05275

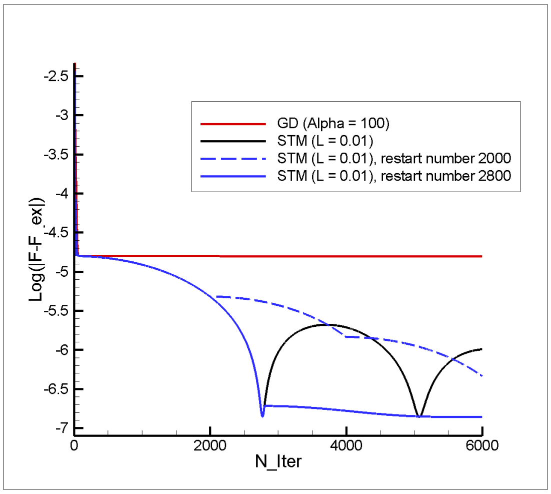


Figure 8: Test 2 - The residual functional (logarithmic scale)

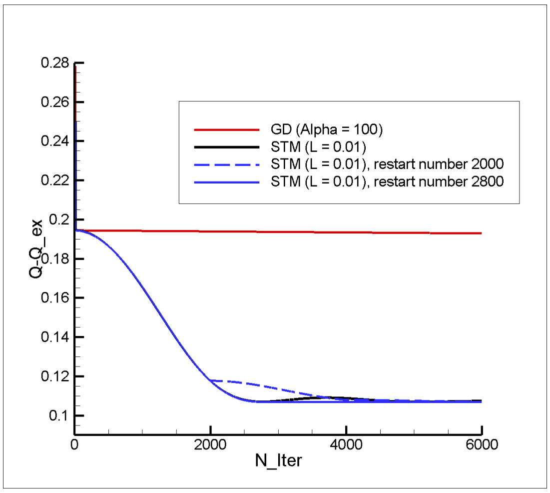


Figure 9: Test 2 (increased depth) - The errors of the methods

[5] Tyurin A., Mirror version of similar triangles method for constrained optimization problems, e-print (2017) arXiv:1705.09809

[6] Tseng P., On accelerated proximal gradient methods for convex-concave optimization // Submitted to SIAM Journal on Optimization, 2008.

[7] Nemirovski A.S., Yudin D.B., Problem complexity and method efficiency in optimization, Nauka publishers (1978) (in Russian)

[8] Fercoq O., Qu Z., Restarting accelerated gradient methods with a rough strong convexity estimate, e-print (2016). arXiv:1609.07358

[9] Devolder O., Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization: PhD thesis, CORE UCL, (2013).

[10] Dvurechensky P.E., Gasnikov A.V., Stochastic Intermediate Gradient Method for Convex Problems with Inexact Stochastic Oracle, J. Optim. Theory Appl., V. 171., no. 1., p. 121–145 (2016)

[11] Kabanikhin S.I., Bektemesov M.A., Nurseitova A.T., Iteration methods of solving inverse and ill-posed problems with data on the part of the boundary, Almaty-Novosibirsk (2006) (in Russian)

[12] Gasnikov A.V., Gasnikova E.V., Nesterov Yu.E., Chernov A.V., Efficient numerical methods for entropy-linear programming, Comp. math. & Math. Phys., V. 56., no. 4., p. 523–534 (2016)

[13] Gasnikov A., Zhukovski M., Kim S., Noskov F., Plaunov S., Smirnov D., Around power law for PageRank in Buckley–Osthus model of web graph, e-print, (2017) arXiv:1701.02595

[14] Dvurechensky P., Gradient method with inexact oracle for composite non-convex optimization, e-print (2017) arXiv:1703.09180

[15] Alekseev V.M., Tikhomirov V.M., Fomin S.V., Optimal control, Springer (1997)

[16] Ryaben'kii V.S., Introduction to computational mathematics, Nauka publishers (1994) (in Russian)

[17] Anikin A.S., Gasnikov A.V., Dvurechensky P.E., Turin A.I., Chernov A.V., Dual approaches to the strongly convex simple function minimization problem under affine restrictions, Comp. Math. & Math. Phys., V. 57., no. 8. (2017) arXiv:1602.01686

[18] Chernov A.V., Dvurechensky P.E., Gasnikov A.V., Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints, In: Kochetov, Yu. et all (eds.) DOOR-2016. LNCS, V. 9869. P. 391–403. Springer, Heidelberg, (2016)

[19] Gasnikov A.V., Dvurechensky P.E., Nesterov Yu.E., Stochastic gradient methods with inexact oracle, TRUDY MIPT, V. 8., no. 1., P. 41–91 (2016) [in Russian]

[20] Kamzolov D., Dvurechensky P., Gasnikov A., Universal intermediate gradient method for convex problems with inexact oracle, Optimization Methods and Software, 1-28, (2020)

[21] Nesterov Y., Gasnikov A., Guminov S. Dvurechensky, P., Primal–dual accelerated gradient methods with small-dimensional relaxation oracle, Optimization Methods and Software, 1-38, (2020)

[22] Dvurechensky P.E., Gasnikov A.V., Nurminski E.A., Stonyakin F.S., Advances in low-memory subgradient optimization. In Numerical Nonsmooth Optimization (pp. 19-59). Springer, Cham, (2020)

[23] Stonyakin F., Tyurin A., Gasnikov A., et al. Inexact Relative Smoothness and Strong Convexity for Optimization and Variational Inequalities by Inexact Model, e-print, (2020) arXiv:2001.09013

[24] Stonyakin F.S., Dvinskikh D., Dvurechensky P., et al. Gradient methods for problems with inexact model of the objective, International Conference on Mathematical Optimization Theory and Operations Research, Springer, Cham, 97-114, (2019)

[25] Gasnikov A., Universal gradient descent, e-print, (2017) arXiv:1711.00394

[26] Gasnikov A.V., Tyurin A.I., Fast gradient descent for convex minimization problems with an oracle producing a $(\delta,\ L)$-model of function at the requested point, Computational Mathematics and Mathematical Physics, 59(7), 1085-1097, (2019)

[27] Poljak B.T., Iterative algorithms for singular minimization problems, Nonlinear Programming 4 (O. L. Mangasarian, Ed.) – Academic Press, 147-166, (1981)

[28] Polyak B.T., Introduction to optimization. optimization software, Inc., Publications Division, New York, (1987)

[29] Kantorovich L.V., Functional analysis and applied mathematics, Uspekhi Mat. Nauk, 3:6(28), 89–185, (1948)

[30] Nocedal J., Wright S., Numerical optimization, Springer Science and Business Media, (2006)

[31] Nesterov Y., Universal gradient methods for convex optimization problems, Mathematical Programming, 152 (1-2), 381-404, (2015)

[32] Kabanikhin S.I., Definitions and Examples of Inverse and Ill-Posed Problems, Journal of Inverse and Ill-Posed Problems, V. 16, No. 4, 317–353 (2018)

[33] Belonosov, A., Shishlenin, M., Klyuchinskiy, D., A comparative analysis of numerical methods of solving the continuation problem for 1D parabolic equation with the data given on the part of the boundary, Advances in Computational Mathematics, V. 45, no. 2, 735–755 (2019)

[34] Shishlenin, M.A., Kasenov, S.E., Askerbekova, Z.A., Numerical algorithm for solving the inverse problem for the Helmholtz equation, Communications in Computer and Information Science, V. 998, 197–207 (2019)

[35] Hào, D.N., Thu Giang, L.T., Kabanikhin, S., Shishlenin, M., A finite difference method for the very weak solution to a Cauchy problem for an elliptic equation, Journal of Inverse and Ill-Posed Problems, V.26, no. 6, 835–857 (2018)

[36] Belonosov, A., Shishlenin, M., Regularization methods of the continuation problem for the parabolic equation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10187 LNCS, 220–226 (2017)

[37] Kabanikhin, S.I., Shishlenin, M.A., Regularization of the decision prolongation problem for parabolic and elliptic elliptic equations from border part, Eurasian Journal of Mathematical and Computer Applications, V. 2, no. 2, 81–91 (2014)

[38] Kabanikhin, S.I., Shishlenin, M.A., Nurseitov, D.B., Nurseitova, A.T., Kasenov, S.E. Comparative Analysis of Methods for Regularizing an Initial Boundary Value Problem for the Helmholtz Equation, Journal of Applied Mathematics, article id 786326 (2014)

[39] Kabanikhin, S.I., Nurseitov, D.B., Shishlenin, M.A., Sholpanbaev, B.B., Inverse problems for the ground penetrating radar, Journal of Inverse and Ill-Posed Problems, V. 21, no. 6, 885–892 (2013)

[40] Kabanikhin, S.I., Gasimov, Y.S., Nurseitov, D.B., Shishlenin, M.A., Sholpanbaev, B.B., Kasenov, S., Regularization of the continuation problem for elliptic equations, Journal of Inverse and Ill-Posed Problems, V. 21, no. 6, 871–884 (2013)

[41] Kabanikhin, S., Shishlenin, M., Quasi-solution in inverse coefficient problems, Journal of Inverse and Ill-Posed Problems, V.16, no. 7, 705–713 (2008)

[42] S. I. Kabanikhin. Inverse and Ill-posed Problems: Theory and Applications. De Gruyter, 2012.

[43] Nesterov Y. Lectures on convex optimization. – Berlin, Germany : Springer, 2018. – V. 137.

[44] 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).