Ministry of Science and Higher Education of the Russian Federation
Moscow Institute of Physics and Technology
(State University)
Department of Control and Applied Mathematics
Department of Intelligent Systems
at the A.A. Dorodnicyn RAS Computing Center.

Bachelor's thesis

"Error Accumulation in the conjugate gradient method for degenerate problems"

4th year student Ryabtsev Anton

Supervisor
Doctor of Physics and Mathematics Gasnikov A.V.

Moscow, 2020

Abstract

In this project, we consider the conjugate gradient method for solving the problem of minimizing a quadratic function with additive noise in the gradient. Three concepts of noise were considered: antagonistic noise in the linear term, stochastic noise in the linear term, and noise in the quadratic term, as well as combinations of the first and second with the last. It was experimentally obtained that error accumulation is absent for any of the considered concepts, which differs from the folklore opinion that, as in accelerated methods, error accumulation must take place. The paper gives motivation for why the error may not accumulate. The dependence of the solution error both on the magnitude (scale) of the noise and on the size of the solution using the conjugate gradient method was also experimentally investigated. Hypotheses about the dependence of the error in the solution on the noise scale and the size (2-norm) of the solution are proposed and tested for all the concepts considered. It turned out that the error in the solution (by function) linearly depends on the noise scale. The work contains graphs illustrating each individual study, as well as a detailed description of numerical experiments, which includes an account of the methods of the noise of both the vector and the matrix.

Contents

1 Introduction

In many applications it is necessary to solve a system of linear algebraic equations:

$$Ax = b.$$

If an exact solution is not required, and you just need to find some approximation, then this problem can be reduced to the problem of minimizing a quadratic function (we assume that the matrix A is symmetric and non-negative definite):

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \to \min_{x\in\mathbb{R}^n}$$

and solve it using well-known methods of numerical optimization, such as, for example, the accelerated Nesterov method [Nesterov, 2010], or the conjugate gradient method [Gasnikov, 2017].

1.1 Research motivation

Many problems that come from real applications turn out to be degenerate (the smallest eigenvalues of the matrix A are zero or close to zero), see, for example, [Kabanikhin S.I., 2012]. In this paper, we will also consider the degenerate case.

It is known[1], It is known that accelerated methods can turn out to be unstable to inaccuracies in gradients, which leads to accumulation of errors with an increase in the number of iterations. As far as I know, the conjugate gradient method has not yet been theoretically investigated in this vein, although, of course, it is also accelerated. Apparently, the reason is associated with the presence of negative results on the convergence of methods with one-dimensional search. So, for example, as follows from [Poljak, 1981] the steepest descent method with arbitrarily small additive imprecision in the gradient may eventually diverge. However, in this paper it is shown that typically such situations do not arise for the conjugate gradient method on quadratic problems and the noise does not accumulate.

It should be noted that the problem of minimizing a positive definite quadratic form is a classical convex optimization problem. The study of this family of problems can give an idea of the convergence (at least in the vicinity of the solution) of various methods in convex optimization problems. The conjugate gradient method is guaranteed to find an exact solution to this problem in $N = n$ iterations, where n — task size. This property is the distinguishing feature of conjugate gradient methods from all sorts of generalizations. However, difficulties appear when solving degenerate (incorrect) quadratic optimization problems. In this case, the number of required iterations can be close to n. Taking into

[1]More in the section Related works.

account that the cost of the iteration is $O(n^2)$, the total complexity is $O(n^3)$, which can be obtained by simpler algorithms, for example, the Gauss method.

1.2 Formulation of the problem

In this paper, the simplest problem is considered:

$$Ax = b \tag{1.1}$$

in the absence of exact values of the matrix A and/or vector b. Matrix $\tilde{A}$ is available instead of original A or vector $\tilde{b}$ instead of original b:

$$\|\tilde{A} - A\|_2 \leq \delta_A, \tag{1.2}$$

$$\|\tilde{b} - b\|_2 \leq \delta_b, \tag{1.3}$$

where $\left\|C\right\|_2 = \sqrt{\lambda_{max}(C^T C)}$.

Task 1.1 reduces to the problem of minimizing the quadratic form:

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle \tilde{b}, x\rangle \to \min_{x\in\mathbb{R}^n}, \tag{1.4}$$

given that 1.3;

$$f(x) = \frac{1}{2}\langle \tilde{A}x, x\rangle - \langle b, x\rangle \to \min_{x\in\mathbb{R}^n}, \tag{1.5}$$

given that 1.2;

$$f(x) = \frac{1}{2}\langle \tilde{A}x, x\rangle - \langle \tilde{b}, x\rangle \to \min_{x\in\mathbb{R}^n}, \tag{1.6}$$

given that 1.2 and 1.3.

2 Related works

It is known [Devolder O., 2013], [Dvinskikh D., Gasnikov A., 2019], that when using a noisy gradient $\tilde{\nabla} f(x)$,satisfactory for all x, y

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

fair is estimate

$$f(x^N) - f(x^*) = O\Big(\frac{LR^2}{N^p} + \delta_1 + N^{p-1}\delta_2\Big).$$

Here $p \in \{1, 2\}$ for non-accelerated and accelerated methods, respectively, and $R = \|x_0 - x^*\|_2$.

For the case of additive noise in a gradient

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

in a series of works by A.S. Nemirovsky [Nemirovskii A.S., Polyak B.T., 1984], [Nemirovskii, 1986], [Nemirovski A., 1992] interesting results were obtained on the regularizing properties of the conjugate gradient method for degenerate (ill-posed) problems of quadratic optimization. We will call degenerate a convex optimization problem for which the ratio of the maximum and minimum eigenvalues of the functional (conditionality of the problem) is much larger than the square of the dimension of the space in which the optimization occurs: $L/\mu \gg n^2$, and not less than the value inverse to the relative accuracy with which it is required to solve the problem. For example, this class of problems includes the problem of minimizing a quadratic form given by a matrix with a set of eigenvalues as in Fig. 1.

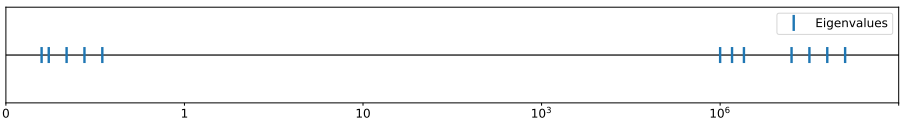


Figure 1: The spectrum of an ill-conditioned matrix.

Many problems that come from real applications turn out to be degenerate, see, for example, [Kabanikhin S.I., 2012]. It is generally impossible to construct algorithms converging in argument for such problems. The solution to the problem turns out to be unstable to inaccuracies in the data. To be able to correctly restore the solution, additional assumptions (source representability) are required. Here we restrict ourselves to the simplest task1.1 in conditions 1.2 and 1.3. According to the task, the following optimization tasks can be constructed

$$f_1(x) = \frac{1}{2}\left\|\tilde{A}x - \tilde{b}\right\|_2^2 \to \min_{x \in \mathbb{R}^n}, \tag{2.1}$$

$$f_2(x) = \frac{1}{2}\langle \tilde{A}x, x\rangle - \langle \tilde{b}, x\rangle \to \min_{x\in\mathbb{R}^n} \text{ (if } A^T = A,\ \tilde{A}^T = \tilde{A}). \tag{2.2}$$

Let's introduce the index $\tau \in \{1, 2\}$, which will correspond to the case under consideration. In work [Nemirovskii, 1986] it was shown that in the case when the source representability condition is satisfied

$$x_* = (A^T A)^{\sigma/2} y_*, \quad \|y_*\|_2 \le R_\sigma, \quad Ax_* = b,$$

conjugate gradient method with stopping criterion of the form

$$\left\|\tilde{A}x^N - \tilde{b}\right\|_2 \le 2\Big(\delta_A \left\|x^N\right\|_2 + \delta_b\Big),$$

starting from $x_0 = 0$, converges for the corresponding problem $\tau \in \{1, 2\}$ in the following way

$$\omega_N^2 = \left\|\tilde{A}x^N - \tilde{b}\right\|_2^2 = O\left(\frac{\tilde{L}^{2(1+\sigma)} R_\sigma^2}{N^{2\tau(1+\sigma)}} + \omega_*^2\right),\ \omega_* = \tilde{L}^\sigma R_\sigma \delta_A + \delta_b,$$

where $\tilde{L} = \max\Big\{\|A\|_2, \|\tilde{A}\|_2\Big\}$, and before the stop criterion is met

$$\left\|\tilde{A}x^N - \tilde{b}\right\|_2 \le 2\Big(\delta_A \left\|x^N\right\|_2 + \delta_b\Big)$$

at $\theta + 2\sigma > 0$, $\theta \in [0, 2]$ the following estimate is valid

$$\nu_{\theta,N}^2 = \left\|\left(A^T A\right)^{\theta/4}\left(x^N - x_*\right)\right\|_2^2 = O\Big(R_\sigma^{(2-\theta)/(1+\sigma)} w_N^{(\theta+2\sigma)/(1+\sigma)}\Big),$$

$$\left\|x^N\right\|_2 = O\big(\|x_*\|_2\big).$$

Note that in $\nu_{\theta,N}^2$ there is an original (not a noisy) matrix A. The above results are accurate and cannot be improved by using other methods. Moreover, they cannot be improved both in terms of the convergence rate and in terms of the achievable accuracy $O(\omega_*)$. It is surprising here, in particular, that the method of conjugate gradients can certainly be classified as a class of accelerated (optimal) methods, for which it is known that in the general case, the inaccuracy in calculating the gradient accumulates linearly with increasing iteration number [Devolder O., 2013]. However, the above result indicates that there is no accumulation of inaccuracies, which corresponds to non-accelerated methods.

Apparently, this is due to the specificity of noise – noise in the gradient is additive. In work [d'Aspremont A., 2008] considered the task $f(x) \to \min_{x\in Q}$ with compact Q. In this work, it is shown that for accelerated methods the additive noise in the gradient does not accumulate as the iterations grow. A more general result (not requiring compactness Q) could be found in work [Dvinskikh D., Gasnikov A., 2019].

Naturally, a hypothesis arises that for the conjugate gradient method, similar to what takes place for accelerated gradient methods, the accumulation of inaccuracies will not be observed as the

iterations grow, if the noise in the gradient is additive. This, in particular, is indicated by the results given above, in which the noise is additive and does not change with the iteration number. More precisely, in this work, we checked hypothesis, what for the case $\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \delta$ the conjugate gradient method satisfies the estimate

$$f(x^N) - f(x^*) = O\Big(\frac{LR^2}{N^2} + \delta R\Big).$$

Moreover, for the noise in the vector b, two concepts were separately considered: hostile noise, which depends on the direction of the gradient at a given moment, and random noise, which is added or subtracted from the vector b at each iteration with the probability $1/2$.

If we assume that the sequence generated by the method is limited, then the noise in the matrix can also be considered as additive noise in the gradient. The hypothesis is tested for this concept

$$\big|f(x^*_{noisy}) - f(x^*)\big| = O\big(\delta R^2\big).$$

Here $f(x^*_{noisy})$ — this is what functional of not noisy task strives for [2] on a sequence generated by conjugate gradients with an imprecise gradient.

I also find it important to note that the existing analysis of optimization algorithms focuses on worst-case analysis. [Nemirovski A., 1995], [Nesterov, 2013]. This type of analysis provides a bound

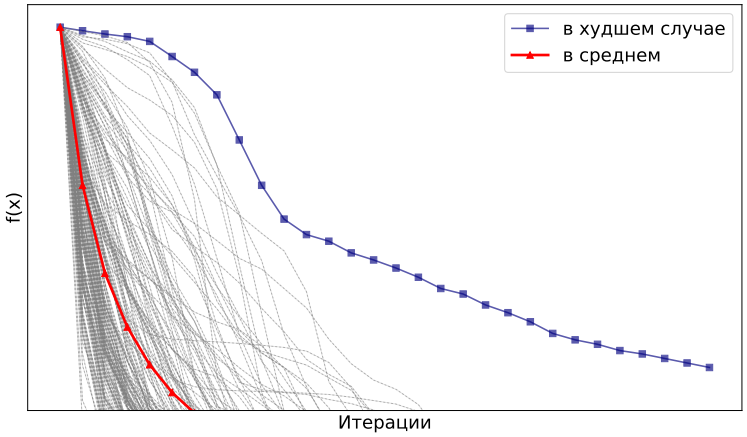


Figure 2: Worst-case analysis can lead to misleading results where the worst-case convergence rate is much worse than the observed average rate. On the graph: the convergence of the conjugate gradient method on individual quadratic forms in gray, as well as the average suboptimality (red), is significantly lower than the worst (blue).

on complexity given the possibility of any input from a function class, no matter how unlikely it is. This can lead to misleading results where worst-case runtimes are much worse than those observed in practice.

[2] It can be assumed that $f(x^k)$ strives for $f(x^*_{noisy})$ in the Cesaro sense (that is, in the sense of arithmetic means).

Average case analysis instead provides the expected complexity of the algorithm in the problem class and is more representative of the typical behavior of the algorithm. Unlike the more classical worst-case analysis, which requires knowledge of only the largest and smallest eigenvalue of the Hessian, average-case analysis requires more detailed knowledge of the spectrum.

Yu.E. Nesterov reported the following observation: methods like conjugate gradients converge just like accelerated methods with fixed steps only on special, as a rule, little interesting for practice examples, and on locally (in the vicinity of the minimum) quadratic functions with the spectrum concentrated around the largest and smallest eigenvalues. For example, for a quadratic form with a uniformly distributed spectrum and a randomly (equiprobably) chosen starting point, one can expect on average [Scieur D., Pedregosa F., 2020] convergence of special variants of conjugate gradient methods with speed (by function) $\sim k^{-6}$ instead of the expected speed $\sim k^{-2}$ (k – iteration number). In this regard, the Fletcher-Reeves algorithm used in this work was also investigated in this vein.

3 Noise concepts in vector b

Experiments are carried out for two concepts of noise in the vector b: antagonistic 3.1 and stochastic 3.2.

$$\delta^k = \delta sign(Ax^k - b) \tag{3.1}$$

$$\delta^k = \begin{cases} \delta, \text{ with probability } \frac{1}{2} \\ -\delta, \text{ with probability } \frac{1}{2} \end{cases} \tag{3.2}$$

The essence of the hostile method is to move the optimum point towards the antigradient at each iteration, which, as it were, creates "running away", which prevents the timely finding of a solution.

Note that when investigating the behavior of the method in the case of matrix noise, the noise at each iteration was introduced according to the idea of random noise for a vector, that is, the generated noise matrix was added to the original one with equal probability, either with a plus or with a minus sign.

More in the section Supplement materials.

4 Conjugate gradient method

For experiments, the method of conjugate gradients was used, namely — Fletcher-Reeves algorithm (algorithm 4.1).

4.1. Conjugate gradient method (Fletcher-Reeves algorithm).

Input: x_0 — start point, $d_0 = -\nabla f(x_0)$.

1: $i := 0$

2: **while** $\|d_i\|_2 \geq \varepsilon$

3: Calculate α_i, minimizing $f(x_i + \alpha_i d_i)$ according to the formula:

$$\alpha_i = -\frac{d_i^T(Ax_i + b)}{d_i^T A d_i} \tag{4.1}$$

Make a step:

$$x_{i+1} = x_i + \alpha_i d_i \tag{4.2}$$

Update direction:

$$d_{i+1} = -\nabla f(x_{i+1}) + \beta_i d_i, \tag{4.3}$$

where β_i is calculating according to the formula:

$$\beta_i = \frac{\nabla f(x_{i+1})^T A d_i}{d_i^T A d_i} \tag{4.4}$$

Output: x_N

5 Numerical experiments

This section presents the parameters of the tasks on which each of the studies was carried out.

5.1 Study of the exit of trajectories of a function to the asymptote

For the graphs illustrating the study of the exit of the trajectories of the function to the asymptote, the parameters from the table were used 1.

	n	δ_A	δ_b	R
Antagonistic noise	10^4	—	0.1	≈ 7000
Stochastic noise	10^4	—	0.1	≈ 7000
Noise in the matrix	10^3	$\{0.0025;\ 0.005\}$	—	≈ 2000
Noise in the matrix and in the vector	10^3	0.005	0.1	≈ 2000

Table 1: The parameters of the problem in the study of the exit of the trajectories of a function to the asymptote. Legend: dimension of the problem n, the size of the noise in the matrix and in the vector, respectively δ_A and δ_b, solution size $R = \|x^* - x_0\|$.

5.2 Study of the dependence of the error on δ

Graphs related to the study of the dependence of the error on δ when solving a noisy problem, correspond to experiments with parameters from the table 2.

	n	δ_A	δ_b	R
Antagonistic noise	10^4	—	$[0;\ 0.1]$	≈ 2000
Stochastic noise	10^4	—	$[0;\ 0.1]$	≈ 2000
Noise in the matrix	10^3	$[0;\ 0.01]$	—	$\{10;\ 50\}$

Table 2: Parameters in the study of dependence on δ.

5.3 Study of the dependence of the error on R.

In table 3 the parameters of the problems are indicated, for which the dependence of the error R.

	n	δ_A	δ_b	R
Noise in the matrix	10^4	$\{0.01;\ 0.1\}$	—	$[0;\ 20]$
Noise in the matrix and in the vector	10^3	0.001	0.01	$[0;\ 50]$

Table 3: Parameters in the study of dependence on R.

5.4 Investigation of the rate of convergence "on average" according to the choice of the starting point

The study of the rate of convergence "on average" according to the choice of the starting point was carried out with the parameters of the problem indicated in the table 4.

n	δ_A	δ_b	R_{ball}
10^3	0.0025	0.1	10^4

Table 4: Parameters in the study of the convergence rate "on average" at the choice of the starting point.

Here the starting point was chosen uniformly in a Euclidean ball of a given radius centered at the point x^*.

5.5 Investigation of the rate of convergence "on average" over the spectrum

In table 5 the parameters are indicated for the study of the convergence rate "on average" over the spectrum.

n	δ_A	δ_b	L/μ
10^3	0.0025	0.1	$[200;\ 220 \cdot 10^3]$

Table 5: Parameters in the study of the rate of convergence "on average" over the spectrum.

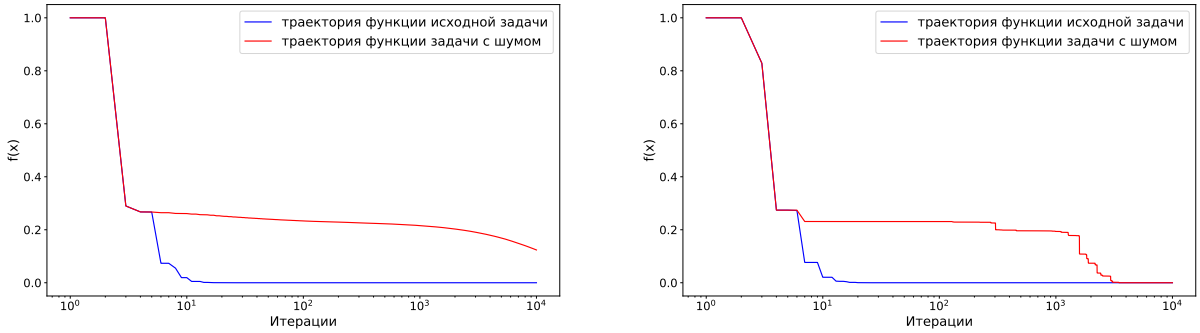
This study was carried out as follows: a diagonal matrix was generated with eigenvalues on the diagonal uniformly distributed on the segment [0; 1], further, it was framed by a randomly selected orthogonal matrix on the left and transposed to it on the right. The resulting matrix was the object with which the research was carried out. In this case, the condition number L/μ varied in the interval indicated in the column of the same name in the table.

The code is available for viewing in Google Colab (link).

6 Results

This section presents the results of each of the studies conducted. The results are accompanied by graphs illustrating them.

6.1 Noise in vector b

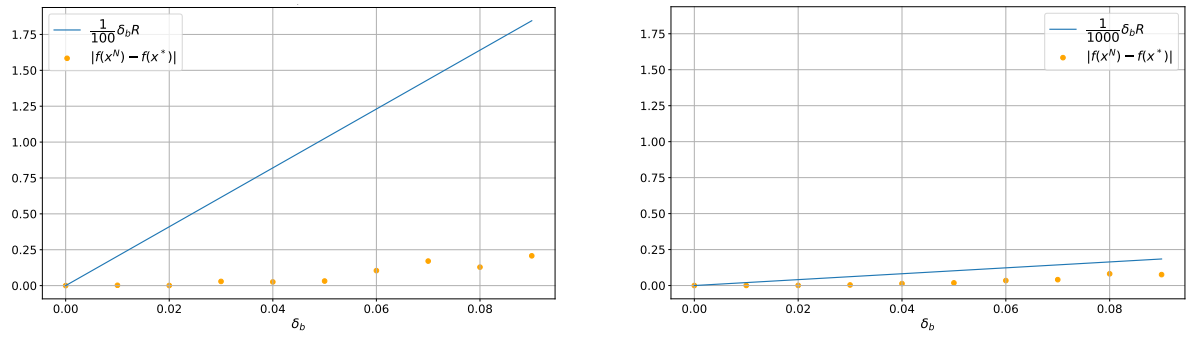


(a) Antagonistic noise in vector.

(b) Stochastic noise in vector.

Figure 3: Dependence of the magnitude of the function $f(x)$ (scaled by one) from the iteration number. The graphs show the approach to the asymptote, which indicates that there is no accumulation of errors with an increase in the number of iterations. Task parameters for this study: $n = 10^4,\ \delta_b = 0.1,\ R \approx 7000$.

The accumulation of error is absent both in the case of hostile noise and in the case of stochastic noise. The method converges, but it takes longer. This fact can be easily established from the graphs in fig. 3. They also show that hostile noise slows down the method.



(a) Antagonistic noise in vector.

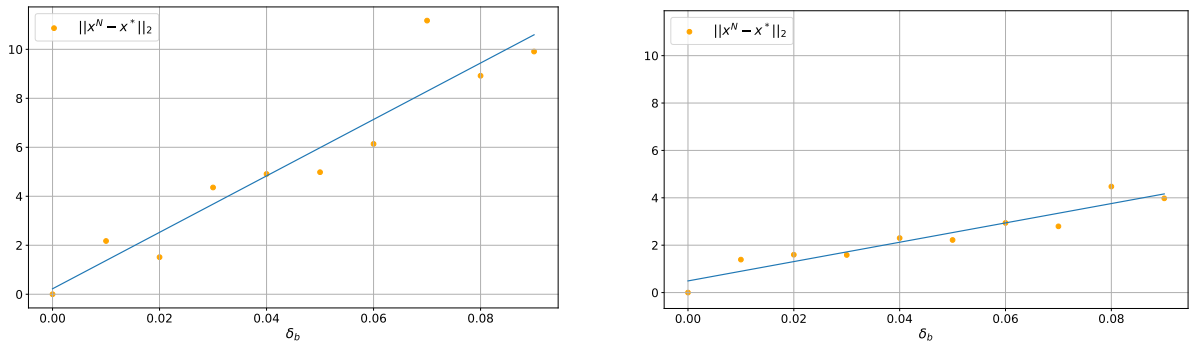
(b) Stochastic noise in vector.

Figure 4: Function residual dependence from δ_b. The graphs show that the error depends on δ_b linearly. Task parameters for this study: $n = 10^4,\ R \approx 2000$.

The graphs in fig. 4 confirm the hypothesis

$$|f(x^*_{noisy}) - f(x^*)| = O(\delta_b R)$$

at small[3] δ_b in the case of antagonistic and stochastic noise.



(a) Antagonistic noise in vector.

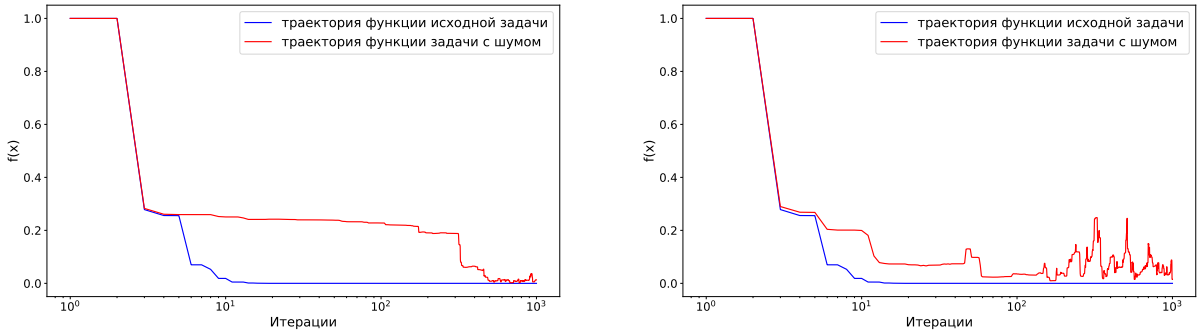
(b) Stochastic noise in vector.

Figure 5: Argument residual dependence from δ_b. Task parameters for this study: $n = 10^4$, $R \approx 2000$.

Since the problem is not strongly convex, convergence in argument is not expected. However, the argument error depends on the size of the noise linearly, as shown in the graphs from fig. 5. The blue line is plotted as an approximation of the experimental points by the least squares method.

All of the above pairs of graphs clearly show that hostile noise leads to worse results than stochastic noise, which explains its name.

6.2 Noise in matrix A



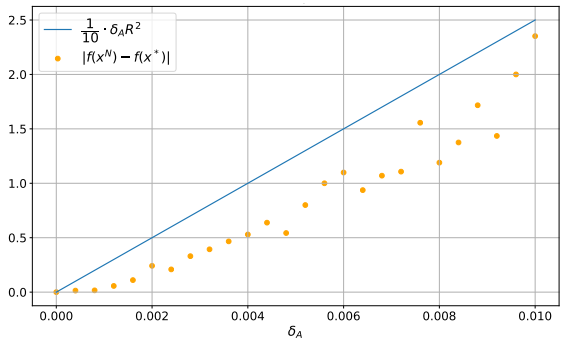
(a) Parameters of the task: $n = 10^3$, $\delta_A = 0.0025$, $R \approx 2000$.

(b) Parameters of the task: $n = 10^3$, $\delta_A = 0.005$, $R \approx 2000$.

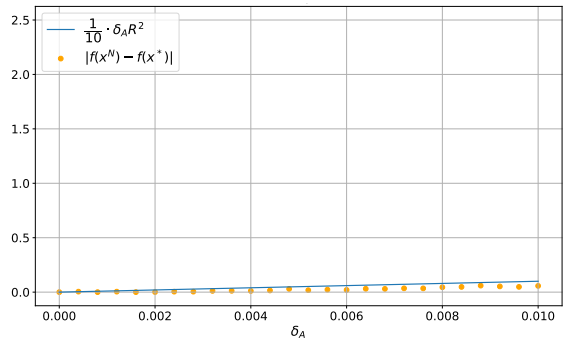
Figure 6: Dependence of the magnitude of the function $f(x)$ (scaled by one) from the iteration number with noise in the matrix. The graphs show the approach to the asymptote, which indicates that there is no accumulation of errors with an increase in the number of iterations. It is also seen that in the case of noise in the matrix, the method turns out to be more sensitive to the magnitude of the noise than in the case of noise in the vector.

[3]In this study, it makes sense to speak exclusively about small noises, since if the noise is comparable with the gradient norm, then it is impossible to find even an approximate solution to the original problem.

If the inaccuracy of the problem is due to noise in the matrix, then there will also be no accumulation of errors (graphs in fig.6).But in this case, the method is much more sensitive to size of δ. This is because the noise in the gradient will not be limited just by δ, as in the case of noise in the vector b, but $\delta\|x_k\|$, that in the vast majority of iterations (provided that $R \gg 1$) will be more than just δ. That is, having some specific δ, the gradient imprecision limit will actually be larger than δ.



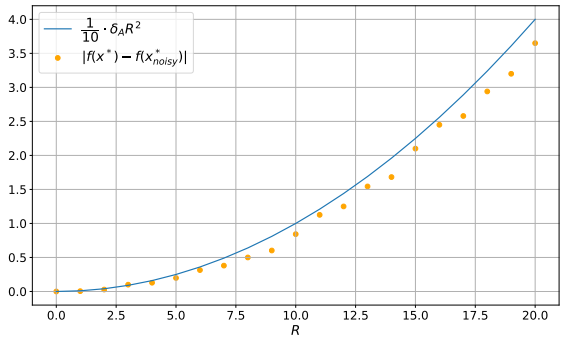
(a) Parameters of the task: $n = 10^4,\ R = 50.$



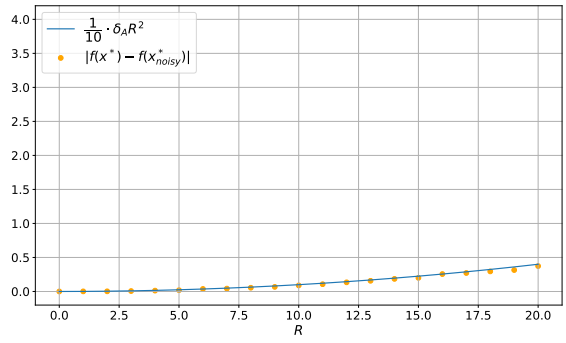
(b) Parameters of the task: $n = 10^4,\ R = 10.$

Figure 7: Function residual dependence from δ_A with noise in the matrix. The graphs show that the error depends on δ_A linearly. It is also seen that a slight increase in R leads to a significant increase in the inaccuracy of the solution for the function.

The graphs in fig. 7 illustrate the linearity of the function residual dependence on δ_A, and also a strong dependence on R — included in the estimate, obviously, not in the first degree.



(a) Parameters of the task: $n = 10^4,\ \delta = 0.1.$



(b) Parameters of the task: $n = 10^4,\ \delta = 0.01.$

Figure 8: Function residual dependence from R with noise in the matrix. The graphs show that the error depends on R quadratically. It is also seen that the increase of δ leads to an increase in the inaccuracy of the solution by the function.

Graphs in fig.8 illustrate the dependence on R more clearly. They show the results of studying the dependence of the residual by function on R at two different but fixed sizes δ_A. These graphs support the hypothesis of entry into the assessment

$$|f(x^*_{noisy}) - f(x^*)| = O(\delta_A R^2)$$

exactly a square of R contrary to the hypothesis

$$|f(x^*_{noisy}) - f(x^*)| = O(\delta_b R)$$

for the noise in vector b.

Together, these four graphs show that the experimental results are consistent with the hypothesis

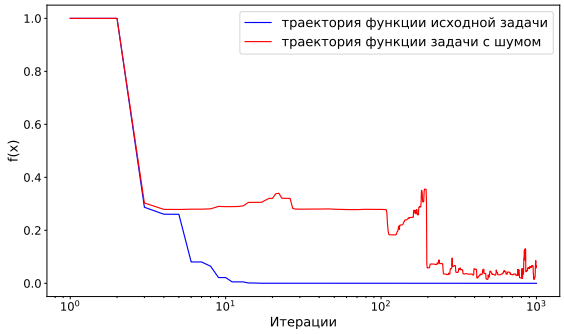
$$|f(x^*_{noisy}) - f(x^*)| = O(\delta_A R^2).$$

6.3 Noise in matrix A and in vector b

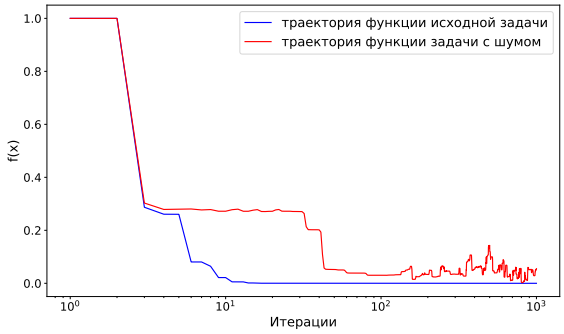
In this case, a hypothesis is a combination of hypotheses for the two previous options, namely

$$|f(x^*_{noisy}) - f(x^*)| = O(\delta_b R \ + \ \delta_A R^2). \tag{6.1}$$

It is quite expected that no accumulation of error is observed here too — the method reaches the asymptote and begins to oscillate around it, which is due to randomness[4] noise in the matrix for the graph in fig. 9(a) and the randomness of the noise in the matrix and the vector (to a greater extent, the randomness of the noise in the matrix) for the graph in fig. 9(b). It can also be seen here that the combination of matrix noise with hostile noise in the vector b slows down the method more than a combination of matrix noise with random noise in the vector b.



(a) Noise in matrix and antagonistic noise in vector.

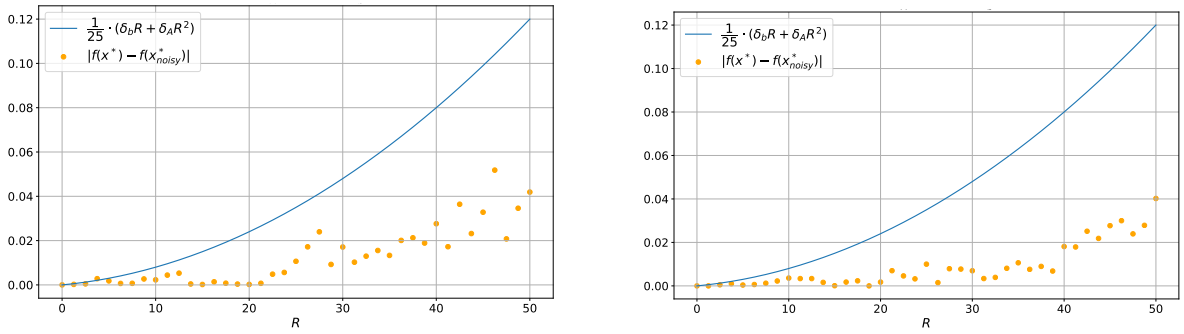


(b) Noise in matrix and stochastic noise in vector.

Figure 9: Dependence of the magnitude of the function $f(x)$ (scaled by one) from the iteration number with noise in the matrix and in the vector. The graphs show the approach to the asymptote, which indicates that there is no accumulation of errors with an increase in the number of iterations. It is also seen that in the case of hostile noise in the vector, the asymptote is reached later than in the case of random noise. Parameters of the task: $n = 10^3$, $\delta_A = 0.005$, $\delta_b = 0.1$, $R \approx 2000$.

[4]Here we are talking about the equiprobable addition or subtraction of the noise matrix or noise vector to the original ones at each iteration. More in the section Supplement materials.

Graphs in fig.10 are consistent with the hypothesis 6.1. Like the previous graphs, they illustrate that hostile noise interferes with the method more than random noise, and this leads to a larger error in the function of the found solution.



(a) Noise in matrix and antagonistic noise in vector.

(b) Noise in matrix and stochastic noise in vector.

Figure 10: Function residual dependence from R. The graphs show that the error depends on R is quadratic. It can also be seen that the noise in the matrix coupled with the hostile noise in the vector leads to worse results than the noise in the matrix together with random noise. Parameters of the task: $n = 10^3$, $\delta_A = 0.001$, $\delta_b = 0.01$.

6.3.1 Investigation of the rate of convergence in "average" according to the spectrum and the choice of the starting point

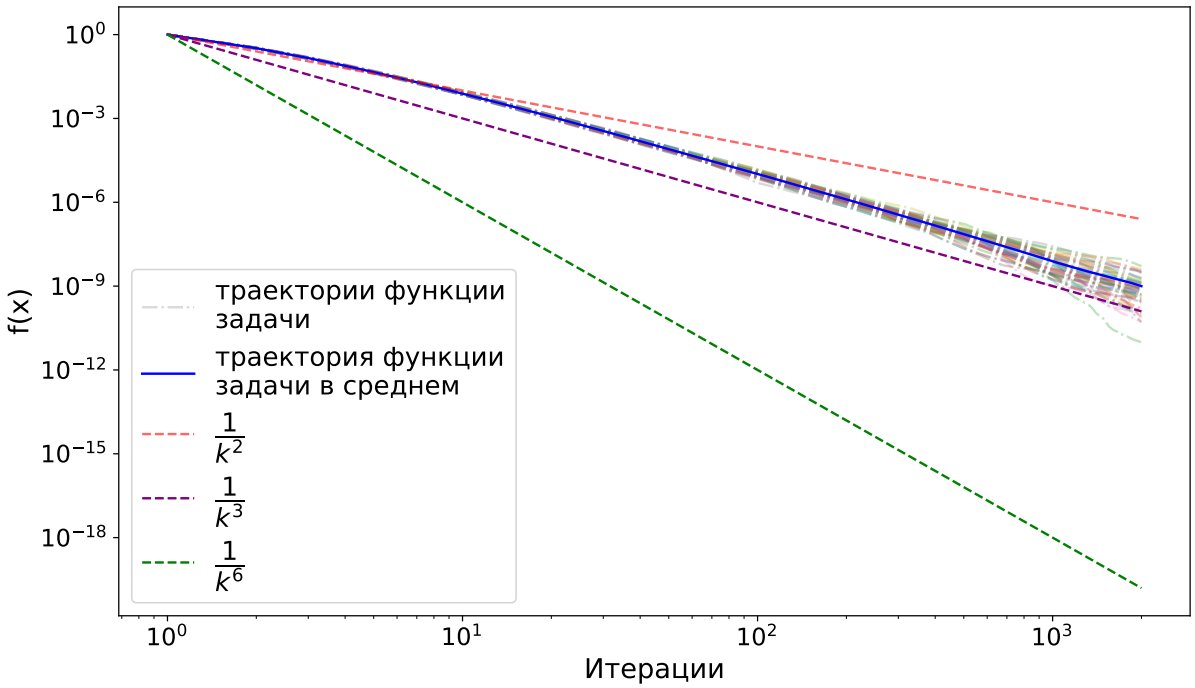
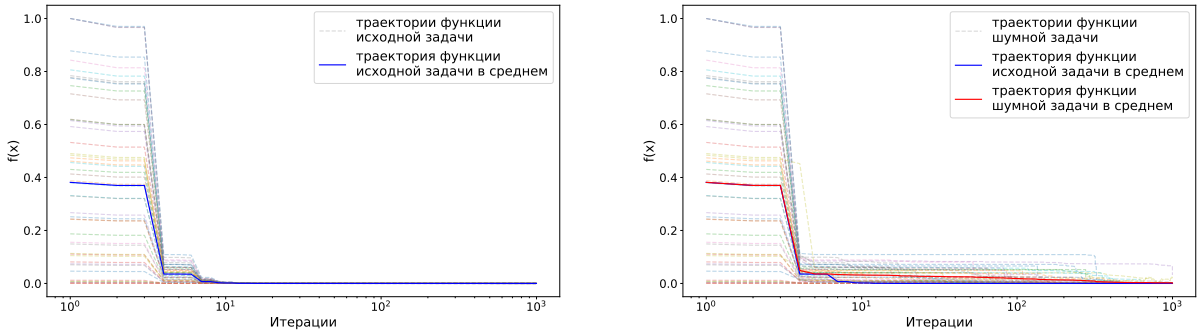


Figure 11: Convergence plot at log-log scale.

In graph 11 the results of the study are presented on average for the spectrum and for the choice

of the starting point. A quadratic form with a uniformly distributed spectrum and a randomly chosen starting point was investigated. The hypothesis was tested that the rate of convergence of the method will be $\sim\ k^{-6}$. The graph shows that the Fletcher-Reeves method does not belong to those "special" variants of conjugate gradient methods that Yu.E. Nesterov.

6.3.2 Investigation of the rate of convergence in "average" according to the choice of the starting point



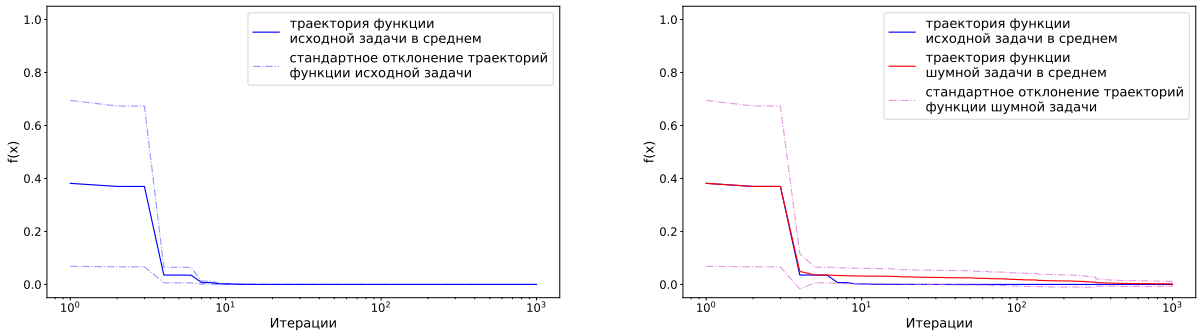
(a) Original task.

(b) Noisy task.

Figure 12: Dependence of the magnitude of the function $f(x)$ (scaled by one) from the iteration number with noise in the matrix and in the vector. The graphs show the outputs to the asymptote for both the exact oracle problem (a) and the noisy problem (b). Parameters of the task: $n = 10^3,\ \delta_A = 0.0025,\ \delta_b = 0.1,\ R = 10^4$.

In fig. 12 shows the graphs of the trajectories of the function for different start points. The graphs are normalized in such a way that the unit along the ordinate is the value of the function at the starting point farthest from the solution. It is easy to see that for the original problem (with an exact oracle), the method finds a solution practically at the same iteration. In this case, in the case of a problem with a noisy oracle, finding an exact solution, as a rule, is out of the question (the exception is cases when the starting point is very close to the solution, and it turns out that, δR may be negligible). As mentioned earlier, here we are more interested in the moment of reaching a certain asymptote. It, in turn, varies from case to case, which can be seen from the graphs in fig. 12(b).

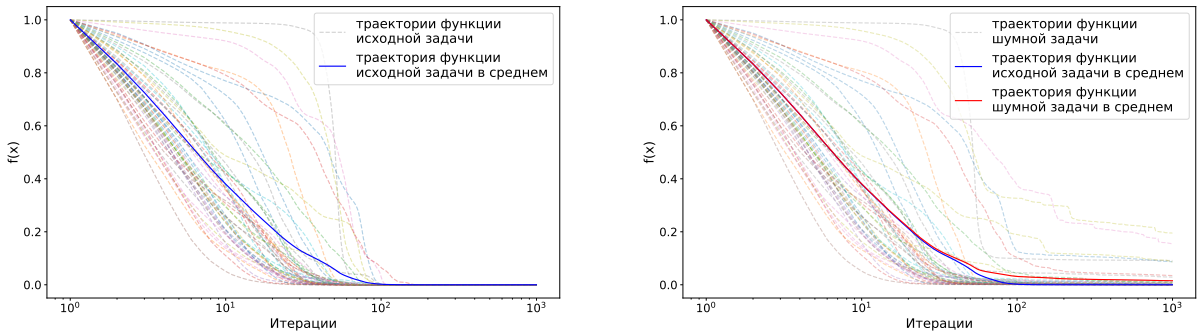
In fig. 13 instead of multiple trajectories, the standard deviation from the mean is shown.



(a) Original task. (b) Noisy task.

Figure 13: Dependence of the magnitude of the function $f(x)$ (scaled by one) from the iteration number with noise in the matrix and in the vector. The graphs show the outputs to the asymptote for both the exact oracle problem (a) and the noisy problem (b). Parameters of the task: $n = 10^3,\ \delta_A = 0.0025,\ \delta_b = 0.1,\ R = 10^4$.

6.3.3 Investigation of the rate of convergence in "average" over the spectrum

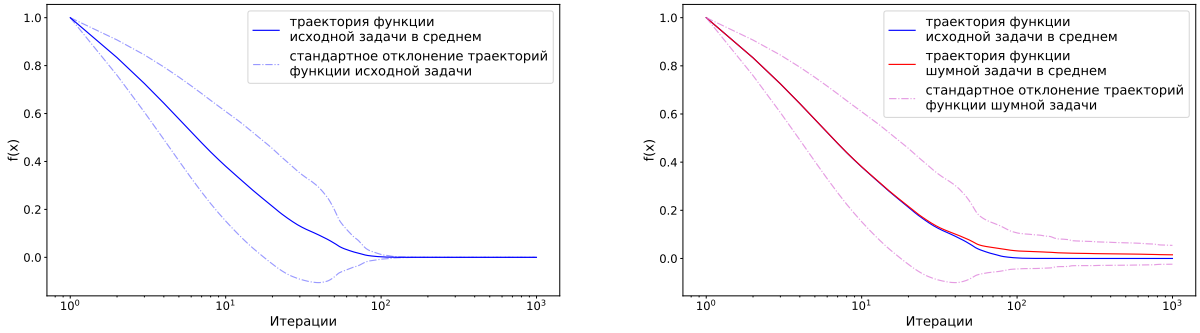


(a) Original task. (b) Noisy task.

Figure 14: Dependence of the magnitude of the function $f(x)$ (scaled by one) from the iteration number with noise in the matrix and in the vector. The graphs show the approach to the asymptote, which indicates that there is no accumulation of errors with an increase in the number of iterations. Parameters of the task: $n = 10^3,\ \delta_A = 0.001,\ \delta_b = 0.01$.

in fig. 14 the graphs of the trajectories of the function for problems with different spectra of the matrix A are shown. Here, since each trajectory corresponds to different tasks (different A and b), then each task has it's own $f(x_0)$ and $f(x^*)$. Therefore, the normalization was carried out for each trajectory separately. As a result, the graphs show the trajectory from the maximum of the function at (starting point) to the minimum (found solution) as a whole. It is easy to see that for the original problem (with an exact oracle), the method finds a solution before some iteration occurs, which is less than the dimension of the problem. In this case, in the case of a problem with a noisy oracle, finding an exact solution, as a rule, is out of the question (the exception is cases when the starting point is very close to the solution, and it turns out that, δRmay be negligible). As mentioned earlier, here we are more interested in the moment of reaching a certain asymptote. It, in turn, varies from

case to case, which can be seen from the graphs in Fig. 14(b).



(a) Original task.

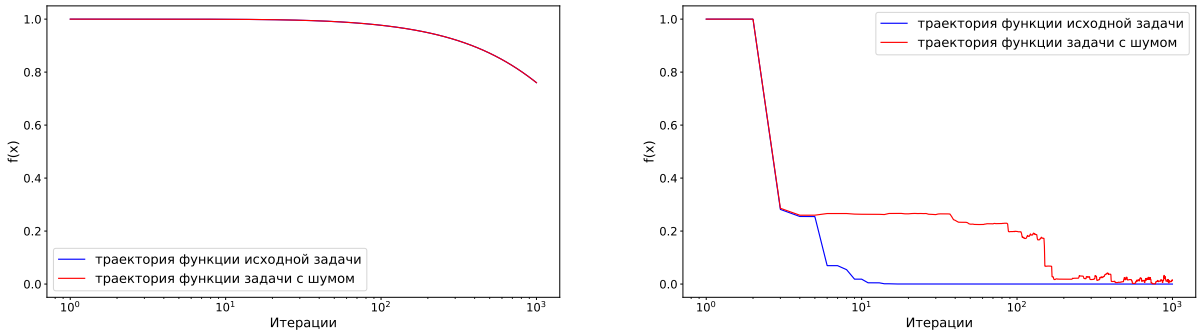
(b) Noisy task.

Figure 15: Dependence of the magnitude of the function $f(x)$ (scaled by one) from the iteration number with noise in the matrix and in the vector. The graphs show the approach to the asymptote, which indicates that there is no accumulation of errors with an increase in the number of iterations. Parameters of the task: $n = 10^3,\ \delta_A = 0.001,\ \delta_b = 0.01$.

In graph. 15 instead of multiple trajectories, the standard deviation from the mean is shown.

7 Conclusion

The result of the work is an experimental illustration of a nontrivial fact: when solving the problem of minimizing a positive definite quadratic form by the method of conjugate gradients with a noisy oracle, there is no accumulation of error. At the same time, it remains the most effective (fast) method for solving the problems of minimizing positive definite quadratic forms of large dimensions, which can be seen from fig. 12.



(a) Accelerated Nesterov method.

(b) Conjugate gradient method.

Figure 16: Dependence of the magnitude of the function $f(x)$ (scaled by one) from the iteration number with noise in the matrix and in the vector. The graphs show that the Nesterov method with noise in the oracle converges on ill-conditioned problems much more slowly than the conjugate gradient method with the same noise. Parameters of the task: $n = 10^3$, $\delta_A = 0.0025$, $\delta_b = 0.1$.

For specialists who are faced with the need to solve SLAEs or search for minima of positive definite quadratic forms with ill-conditioned matrices of large dimensions, these experimental results add confidence when using the conjugate gradient method in the absence of an exact gradient. Although the noise here leads to a greater inaccuracy of the obtained solution than when using other accelerated methods, the conjugate gradient method turns out to be tens and even hundreds of times faster. As a further development of this study, it would be interesting to generalize the result obtained to the case of a Hilbert space, where the inaccuracy in the solution arises naturally due to the impossibility of calculating the gradient in all directions.

8 Supplement materials

It is important to note that there are various approaches to generating noisy vectors and matrices. The methods we have chosen are due to the speed of computation and the relative ease of implementation.

8.1 Method of generating a noisy vector

- Basics
  - $\|\mathbf{b}\|_2 = \sqrt{b_1^2 + b_2^2 + ... + b_n^2}$
  - $\|\tilde{\mathbf{b}} - \mathbf{b}\|_2 = \sqrt{(\tilde{b}_1 - b_1)^2 + (\tilde{b}_2 - b_2)^2 + ... + (\tilde{b}_n - b_n)^2} \leq \delta_b$
  - $\|\tilde{\mathbf{b}} - \mathbf{b}\|_2 = \sqrt{\Delta_1^2 + \Delta_2^2 + ... + \Delta_n^2} \leq \delta_b$
  - $\Delta_1^2 + \Delta_2^2 + ... + \Delta_n^2 \leq \delta_b^2$
- Antagonistic noise
  - $\{\xi_i\}_{i=1}^n \in \mathcal{N}(0, 1)$
  - $\Delta_j^k = \sqrt{\frac{\xi_j^2 \cdot \delta_b^2}{\sum_{i=1}^n \xi_i^2}} \cdot \text{sign}\left(\left[\nabla f(x^k)\right]_j\right)$,
    in our case $\text{sign}\left(\left[\nabla f(x^k)\right]_j\right) = \text{sign}\left(\left[Ax^k - b\right]_j\right)$ – sign j-th gradient element at point x^k
  - $\mathbf{\Delta^k} = (\Delta_1^k, \Delta_2^k, \ldots, \Delta_n^k)^T$
  - $\tilde{\mathbf{b}}^{\mathbf{k}} = \mathbf{b} + \mathbf{\Delta^k}$
- Stochastic noise
  - $\{\xi_i\}_{i=1}^n \in \mathcal{N}(0, 1)$
  - $\Delta_j^k = \sqrt{\frac{\xi_j^2 \cdot \delta_b^2}{\sum_{i=1}^n \xi_i^2}}$ – normalization
  - $\mathbf{\Delta^k} = (\Delta_1^k, \Delta_2^k, \ldots, \Delta_n^k)^T$
  - $\tilde{\mathbf{b}}^{\mathbf{k}} = \mathbf{b} \pm \mathbf{\Delta^k}$ – with probability $\frac{1}{2}$

8.2 Method for generating a noisy matrix

- $\|\tilde{A} - A\|_2 \leq \delta_A$
- $\tilde{A} = A \pm M$ – with probability $\frac{1}{2}$, M – matrix of noise
- $\tilde{A} - A = \pm M$

– $\|M\|_2 \leq \delta_A$

– $\{\xi_i\}_{i=1}^{n\times n} \in \mathcal{N}(0,1)$

– $m_{1k} = \sqrt{\frac{\xi_k^2 \cdot \delta_A^2}{\sum_{i=1}^{n\times n} \xi_i^2}}$ – normalization of the first row of the noise matrix

– $m_{2k} = \sqrt{\frac{\xi_{(2-1)n+k}^2 \cdot \delta_A^2}{\sum_{i=1}^{n\times n} \xi_i^2}}$ – normalization of the second row of the noise matrix

– ...

– $m_{pk} = \sqrt{\frac{\xi_{(p-1)n+k}^2 \cdot \delta_A^2}{\sum_{i=1}^{n\times n} \xi_i^2}}$ – normalization of p-th row of the noise matrix

– ...

– $m_{nk} = \sqrt{\frac{\xi_{(n-1)n+k}^2 \cdot \delta_A^2}{\sum_{i=1}^{n\times n} \xi_i^2}}$ – normalization of the last row of the noise matrix

– $M = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1n} \\ m_{21} & m_{22} & \dots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n_2} & \cdots & m_{nn} \end{bmatrix}$

References

[Gasnikov, 2017] *Gasnikov A. V.* Universal gradient descent //arXiv preprint arXiv:1711.00394. – 2017.

[Poljak, 1981] *Poljak B. T.* Iterative algorithms for singular minimization problems //Nonlinear Programming 4. – Academic Press, 1981. – . 147-166.

[Nesterov, 2010] *Nesterov Y. E.* Vvedenie v vypukluyu optimizatsiyu. – 2010.

[Scieur D., Pedregosa F., 2020] *Scieur D., Pedregosa F.* Universal Average-Case Optimality of Polyak Momentum //arXiv preprint arXiv:2002.04664. – 2020.

[Gill, Myurrei, Rait, 1985] *Gill F., Myurrei U., Rait M.* Prakticheskaya optimizatsiya. – Mir, 1985. – T. 509.

[Nemirovskii, 1986] *Nemirovskii A. S.* O regulyarizuyushchikh svoistvakh metoda sopryazhennykh gradientov na nekorrektnykh zadachakh //Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki. – 1986. – T. 26. – . 3. – S. 332-347.

[Nemirovskii A.S., Polyak B.T., 1984] *Nemirovskii A. S., Polyak B. T.* Iteratsionnye metody resheniya lineinykh nekorrektnykh zadach pri tochnoi informatsii. II. // Izv. AN SSSR. Tekhnicheskaya kibernetika – 1984 – 3. – S. 18–25.

[Nemirovski A., 1992] *Nemirovski A.* Information-based complexity of linear operator equations // Journal of Complexity. – 1992. – V. 8. – P. 153–175.

[Nemirovski A., 1995] *Nemirovski A.* Information-based complexity of convex programming // Lecture Notes. – 1995.

[Nesterov, 2013] *Nesterov Yu. E.* Introductory lectures on convex optimization: A basic course. – Springer Science Business Media, 2013. – . 87.

[Kabanikhin S.I., 2012] *Kabanikhin S. I.* Inverse and ill-posed problems. – De Gruyter, 2012.

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

[Dvinskikh D., Gasnikov A., 2019] *Dvinskikh D., Gasnikov A.* Decentralized and parallelized primal and dual accelerated methods for stochastic convex programming problems //arXiv preprint arXiv:1904.09015. – 2019.

[d'Aspremont A., 2008] *d'Aspremont A.* Smooth optimization with approximate gradient //SIAM Journal on Optimization. – 2008. – . 19. – . 3. – . 1171-1183.

[Scieur D., Pedregosa F., 2020] *Scieur D., Pedregosa F.* Universal Average-Case Optimality of Polyak Momentum //arXiv preprint arXiv:2002.04664. – 2020.