National Research University Higher School of Economics

as a manuscript

Alexander Igorevich Tyurin

Development of a method for solving structural optimization problems

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow - 2020

The PhD dissertation was prepared at National Research University Higher School of Economics

Academic Supervisor:

Alexander Vladimirovich Gasnikov, Doctor of Sciences in Mathematical Modelling, Numerical Methods and Software Complexes, Senior Research Fellow at International Laboratory of Stochastic Algorithms and High-Dimensional Inference, National Research University Higher School of Economics

1 Introduction

Optimization methods have a significant impact on all spheres of human society. It is difficult to list all recent activities where optimization methods are used to solve practical problems. In many problems of economics, engineering, programming optimization methods are helpful. Optimization methods came up with computer engineering in the twentieth century. That is when the active development of the modern theory of optimization began. The pioneer is L. Kantorovich [1, 2], who considered linear programming problems in engineering and economics. In the 50s-60s, cutting edge works were done by G. Rubinstein, E. Ventsel, N. Vorobyov, D. Yudin, E. Golstein, N. Shor, B. Polyak. Yu. Ermoliev, L. Pontryagin, etc. In those years, researchers proposed the following methods: Pontryagin's maximum principle, projection gradient method, cutting plane method, penalty method, Newton's method for constrained optimization, subgradient method, the center of gravity method, etc. In 70th, A. Nemirovskii and D. Yudin have a significant impact on optimization development with work [3]. In this work, they used the oracle concept (black box), which for any input point returns, for example, function and gradient value. A. Nemirovski and D. Yudin obtained lower bounds for convergence rates for some general optimization problems classes (convex optimization problems, optimization problems with Lipchitz continuous functions, smooth strongly convex optimization problems, etc.). We should note that optimization methods that achieve corresponding lower bounds are proposed later for some classes of problems. In particular, Yu. Nesterov developed the fast gradient method [4], which has the convergence rate inversely proportional to the root of the accuracy of the solution by function in the class of functions with Lipchitz continuous gradient. This convergence rate is optimal in the sense of black-box oracle calls. The same result was obtained for smooth strongly convex functions.

It means that for many classes of optimization problems, optimal methods were developed; however, the progress did not stop. In practice, optimization tasks have some structure that allows developing new algorithms for every problem with faster convergence rates. Let us note some popular examples from structural optimization. Composite optimization solves optimization problems that can be represented as a sum of smooth and nonsmooth functions. Despite the fact that the sum is a nonsmooth function, with some additional assumptions about the nonsmooth part, we can develop methods that have convergence rates as in smooth optimization tasks [5]. Similarly, using the structure of optimization tasks, we can propose algorithms with more optimistic convergence rates for the following optimization problems: functions with Holder continuous gradients [6], superposition of functions (min-max problems) [7, 8], transportation problems [9, 10, 11], clustering by electorial model [12], etc.

We should note another development direction in structural optimiza-

tion theory that connected with different requirements about an oracle. In general, an oracle is a black-box framework that makes calculations. The complexity of optimization methods is estimated by the number of calls of an oracle. In classical optimization theory [3, 13] oracles for some query point return a function value (zero-order oracle), gradient / subgradient (first-order oracle), hessian (second-order oracle), etc. In particular, for a smooth convex function f with L–Lipchits gradient it is known that there is exist an oracle [13] such that for some query point y returns a pair $(f(y), \nabla f(y))$ and the following inequality holds:

$$0 \leq f(x) - (f(y) + \langle \nabla f(y), x - y \rangle) \leq \frac{L}{2} \|x - y\|^2 \quad \forall x \in Q, \tag{1}$$

where Q is a convex closed set on which a function f is defined. We have the left inequality from the convexity of a function f and the right inequality we have from L–Lipchits gradient assumption. Using this oracle, we can obtain the optimal convergence rate for this class of optimization problems. In detail, after N calls of the oracle, the guaranteed convergence rate is equal to $\mathcal{O}\left(\frac{LR^2}{N^2}\right)$ [4, 13], where a constant R is a distance between a method's starting point and the closest optimal point. In practice and theory, inequalities (1) do not always hold. Even for smooth optimization problems with Lipchits gradients, we can not get precise values of function and gradient due to calculation errors or the fact, that we obtain these values using some auxiliary problem. For these examples, inequalities (1) do not hold. Indeed, we can show that (1) holds only for some unique pair $(f(y), \nabla f(y))$. In [14] authors introduced (δ, L)–oracle that for some query point y returns a pair $(f_\delta(y), \nabla f_\delta(y))$ such that

$$0 \leq f(x) - (f_\delta(y) + \langle \nabla f_\delta(y), x - y \rangle) \leq \frac{L}{2} \|x - y\|^2 + \delta \quad \forall x \in Q. \tag{2}$$

Unlike (1), a pair $(f_\delta(y), \nabla f_\delta(y))$ is not unique and, in general, is not equal to $(f(y), \nabla f(y))$. The proposed oracle allows us to generalize the classical gradient and fast gradient type methods to a wider class of tasks. From [14], the guaranteed convergence rate for the fast gradient method is equal to $\mathcal{O}\left(\frac{LR^2}{N^2} + N\delta\right)$ and for the gradient method is equal to $\mathcal{O}\left(\frac{LR^2}{N} + \delta\right)$ using (δ, L)–oracle. We should note, that obtained convergence rates do not require smoothness of optimization problems. We can obtain (δ, L)–oracle for the following optimization problems [14]: functions with Holder continuous subgradient, functions obtained by smoothing techniques [15, 16, 17], Moreau-Yosida regularization [18], and composite optimization [5]. Note that there is another concept of inexact oracle [19] that is a particular case of (δ, L)–oracle [14].

A valuable property of optimization methods is to effectively resolve a dual solution from a primal solution (or vice versa) [20, 21, 22, 23], which

is called primal–duality. This property is advantageous in transportation problems [9, 10, 11], machine learning problems (ex. SVM), etc [24]. Another useful property of optimization methods is to be robust to cases when instead of a gradient oracle returns a stochastic gradient, random, unbiased vector w.r.t. a real gradient. Stochastic optimization methods are trendy because they allow reducing the calculation cost of a descent direction. It is impossible for some optimization problems [25, 26] to calculate a gradient in a reasonable amount of time, even at one point. Therefore, in this thesis, we focus on the extension of our results to primal–duality and stochasticity.

Object and goals of the dissertation: unification of previously proposed gradient-type methods into one method using a special concept of inexact model. Develop a series of methods that can solve generalized optimization problem statements and use its structure with the aid of the proposed concept of inexact model. Moreover, prove corresponding rates of convergence, when possible, in an optimal way for some classes of optimization problems.

The obtained results:

1. We propose concepts of inexact model for gradient-type methods.
2. We developed the adaptive gradient and fast gradient methods for optimizations tasks that support the concept of inexact model ($(\delta, L, \|\|\|)$–model).
3. We developed the gradient method for optimization tasks with relative smoothness that support the concept of inexact model ((δ, L, V)–model).
4. We developed the primal–dual adaptive gradient and fast gradient methods for optimizations tasks that support the concept of inexact model ($(\delta, L, \|\|\|)$–model).
5. We developed the stochastic gradient and fast gradient methods for stochastic optimizations tasks that support the concept of inexact model ($(\delta_1, \delta_2, L, \|\|\|)$–model).
6. We propose the heuristic (without theoretical guarantees) adaptive stochastic fast gradient method that is based on the adaptive fast gradient method and the stochastic fast gradient method.

Author's contribution includes the development of optimization methods for oracles that return an inexact model of a function, proving convergence rates of corresponding methods, and constructing of inexact models for problems from structural optimization. The heuristic adaptive algorithm is proposed for stochastic optimization problems.

Novelties: we developed the adaptive gradient and fast gradient methods for oracles that return an inexact model of a function. Further, we

constructed the gradient method for problems with relative smoothness, the primal–dual adaptive gradient and fast gradient methods, and the stochastic nonadaptive gradient methods that support an inexact model of a function. We attempted to enrich the stochastic nonadaptive fast gradient method with adaptivity. However, we were only able to develop the heuristic adaptive fast gradient method that showed high performance in practice.

As a result of the work of this thesis, 8 papers were published:

First-tier publications:

1. Gasnikov A., Tyurin A. (2019) Fast gradient descent for convex minimization problems with an oracle producing a (δ, L)-model of function at the requested point. Computational Mathematics and Mathematical Physics, 59, 7, 1085–1097, Scopus Q2 (main co-author; the author of this thesis formulated and proved convergence rate theorems for the gradient and fast gradient methods (Theorem 1 and 2), presented a description of examples (Section 4)).

2. Stonyakin F., Dvinskikh D., Dvurechensky P., Kroshnin A., Kuznetsova O., Agafonov A., Gasnikov A., Tyurin A., Uribe C., Pasechnyuk D., Artamonov S. (2019) Gradient methods for problems with inexact model of the objective. Lecture Notes in Computer Science, 11548, 97–114, Scopus Q2 (the author of this thesis prepared the text of section 2 and proved the convergence rate theorem for the gradient method for optimization problems with relative smoothness (Theorem 1)).

3. Ogaltsov A., Tyurin A. (2020) A heuristic adaptive fast gradient method in stochastic optimization problems. Computational Mathematics and Mathematical Physics, 60, 7, 1108–1115, Scopus Q2 (main co-author; the author of this thesis proposed the heuristic adaptive fast gradient method for stochastic optimization problems (Algorithm 2), did an analysis and justification).

4. Dvurechensky P., Gasnikov A., Omelchenko A., Tyurin A. (2020) A stable alternative to Sinkhorn's algorithm for regularized optimal transport. Lecture Notes in Computer Science, 12095, 406–423, Scopus Q2 (the author of this thesis helped with the development of Algorithm 1 and the proof of Theorem 1).

5. Dvinskikh D., Omelchenko A., Gasnikov A., Tyurin A. (2020) Accelerated gradient sliding for minimizing the sum of functions. Doklady Mathematics, 101, 3, Scopus Q2 (in press), (the author of this thesis helped with the text of this paper and the proofs of intermediate results).

Second-tier publications:

1. Tyurin A. (2020) Primal-dual fast gradient method with a model. Computer Research and Modeling, 12, 2, 263–274, Scopus Q3.

2. Dvinskikh D., Tyurin A., Gasnikov A., Omelchenko S. (2020) Accelerated and nonaccelerated stochastic gradient descent with model conception. Mathematical Notes, 108, 4, Scopus Q3 (in press), (main coauthor; the author of this thesis developed the fast gradient method for stochastic optimization tasks, provided description of examples).

3. Anikin A., Gasnikov A., Dvurechensky P., Tyurin A., Chernov A. (2017) Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints. Computational Mathematics and Mathematical Physics, 57, 8, 1262–1276, Scopus Q3. (the author of this thesis helped in writing of remarks).

Reports at conferences and seminars:

1. 8th Moscow International Conference on Operations Research, Russia, Moscow. (17.10.2016 - 22.10.2016). Dual fast gradient method for entropy–linear programming problems.

2. 59th MIPT Scientific Conference, Russia, Dolgoprudny. (21.11.2016 - 26.11.2016). Adaptive fast gradient method for convex min–max problems.

3. Workshop “Three oracles”, Russia, Skolkovo. (28.12.2016). On several extensions of similar triangles method.

4. Scientific conference "Modeling the Co-evolution of Nature and Society: problems and experience" devoted to the 100-th anniversary of N. N. Moiseev, Russia, Moscow. (7.11.2017 - 10.11.2017). Adaptive similar triangles method and its application in calculation of regularized optimal transport.

5. 60th MIPT Scientific Conference, Russia, Dolgoprudny. (20.11.2017 - 25.11.2017). The mirror triangle method with a generalized inexact oracle.

6. The 23rd International Symposium on Mathematical Programming, France, Bordeaux. (1.7.2018 - 6.7.2018). Universal Nesterov’s gradient method in general model conception.

7. 62th MIPT Scientific Conference, Russia, Dolgoprudny. (18.11.2019 - 23.11.2019). Primal–dual fast gradient method with a model.

The reported study was funded by RFBR, project number 19-31-90062 and project number 18-31-20005 mol-a-ved.

2 Convex optimization problem

Let us describe the mathematical formulation of a convex optimization problem [13]. Given an objective function $f(x) : Q \longrightarrow \mathbb{R}$, a set Q is a subset of finite–dimensional linear vector space $\mathbb{R}^n$, and a norm $\|\cdot\|$ in $\mathbb{R}^n$. Conjugate norm we define as

$$\|\lambda\|_* = \max_{\|\nu\| \leq 1, \nu \in \mathbb{R}^n} \langle \lambda, \nu \rangle \quad \forall \lambda \in \mathbb{R}^n.$$

Let us define prox–function and Bregman divergence [27], [28] (p. 327):

Definition 1. *$d(x) : Q \to \mathbb{R}$ is a prox–function if $d(x)$ is continuously differentiable on* int Q *and a function $d(x)$ is 1-strongly convex w.r.t. a norm $\|\|$ on* int Q.

Definition 2. *A function*

$$V[y](x) = d(x) - d(y) - \langle \nabla d(y), x - y \rangle$$

is called Bregman divergence, where $d(x)$ is a prox–function.[1]

The introduction of Bregman divergence allows us to obtain more general results for convergence rates. The classical example of Bregman divergence is the function $V[y](x) = \frac{1}{2}\|x - y\|_2^2$.

Further, we assume that

1. A set $Q \subseteq \mathbb{R}^n$ is a convex and closed set.
2. A function $f(x)$ is continuous and convex on Q.
3. A function $f(x)$ is lower bounded on Q and attains its minimum at some point from Q (not necessarily unique).

We consider the following optimization problem:

$$f(x) \to \min_{x \in Q}. \tag{3}$$

A point x_* is a solution of the optimization problem if inequality $f(x_*) \leq f(x)$ holds for all $x \in Q$. Also, we call a point x_ε as ε–solution if $f(x_\varepsilon) - f(x_*) \leq \varepsilon$ for all $x \in Q$. The main task of numerical convex optimization is to find ε–solution.

[1] In paper [29], $V[y](x)$ is denoted as $V(x, y)$.

2.1 The concept of inexact solution

Now we define the concept of inexact solution (see [28]) that we use in our methods.

Definition 3. *Given an optimization problem*

$$\psi(x) \to \min_{x \in Q},$$

where $\psi(x)$ is a convex function, then we denote $\operatorname{Arg\,min}^{\widetilde{\delta}}_{x \in Q} \psi(x)$ *by a set of $\widetilde{x}$ such that*

$$\exists h \in \partial\psi(\widetilde{x}), \;\; \langle h, x - \widetilde{x} \rangle \geq -\widetilde{\delta} \;\; \forall x \in Q,$$

where $\partial\psi(\widetilde{x})$ is a subderivative of a function ψ at a point $\widetilde{x}$. Any point from $\operatorname{Arg\,min}^{\widetilde{\delta}}_{x \in Q} \psi(x)$ *we denote as* $\arg\min^{\widetilde{\delta}}_{x \in Q} \psi(x)$.

This definition is stringent compared to the definition of δ–solution by function (see [29]). For instance, both definitions are equivalent when $\widetilde{\delta} = 0$. However, in some more general cases, we can derive solutions in terms of Definition 3 from δ–solution by function (see [30, 29]).

3 Contents

In this section, we describe results and conclusions in more detail.

3.1 Inexact model of a function

In [14], the authors proposed (δ, L)–oracle and corresponding methods that allow solving a vast number of optimization problems. Let us introduce concepts of inexact model ((δ, L)–model) of a function that generalizes (δ, L)–oracle.

Definition 4. *Given a convex function $\psi_\delta(x, y)$ w.r.t. x on a set Q such that $\psi_\delta(x, x) = 0$ for all $x \in Q$. The function $\psi_\delta(x, y)$ is $(\delta, L, \|\|)$–model of a function f at a point y w.r.t. $\|\cdot\|$ with value $f_\delta(y)$ if for all $x \in Q$ inequalities*

$$0 \leq f(x) - (f_\delta(y) + \psi_\delta(x, y)) \leq \frac{L}{2} \|x - y\|^2 + \delta \tag{4}$$

hold for some values $L, \delta \geq 0$.[2]

From the view of an oracle concept, we can assume that for a query point y, the oracle returns a pair $(f_\delta(y), \psi_\delta(x, y))$. Also, we can provide a more general definition by using so–called relative smoothness [32, 33, 34]:

[2] In papers [29, 31], $(\delta, L, \|\|)$–model is defined as (δ, L)–model.

Definition 5. *Given a convex function $\psi_\delta(x, y)$ w.r.t. x on a set Q such that $\psi_\delta(x, x) = 0$ for all $x \in Q$. The function $\psi_\delta(x, y)$ is $(\delta, L, \|\|)$–model of a function f at a point y w.r.t. Bregman divergence V with value $f_\delta(y)$ if for all $x \in Q$ inequalities*

$$0 \leq f(x) - (f_\delta(y) + \psi_\delta(x, y)) \leq LV[y](x) + \delta \tag{5}$$

hold for some values $L, \delta \geq 0$.[3]

We can obtain Definition 4 from Definition 5 if we take Bregman divergence $V[y](x) = \frac{1}{2}\|x - y\|^2$. The oracle that produces $(\delta, L, \|\|)$–model from Definition 4 is more universal than (δ, L)–oracle (see (2)). Indeed, it is enough to take $\psi_\delta(x, y) = \langle \nabla f_\delta(y), x - y\rangle$.

In [35], we propose a more general definition than Definition 4 by the introduction of additional noise, namely:

Definition 6. *Given a convex function $\psi_\delta(x, y)$ w.r.t. x on a set Q such that $\psi_\delta(x, x) = 0$ for all $x \in Q$. The function $\psi_\delta(x, y)$ is $(\delta_1, \delta_2, L, \|\|)$–model of a function f at a point y w.r.t. $\|\cdot\|$ if for all $x \in Q$ inequalities*

$$-\delta_1(x, y) \leq f(x) - (f(y) + \psi_\delta(x, y)) \leq \frac{L}{2}\|x - y\|^2 + \delta_2 \tag{6}$$

hold for some values $L \geq 0$, δ_2, and $\delta_1(x, y)$.

We can show, that $(\delta, L, \|\|)$–model is $(\delta, \delta, L, \|\|)$–model with $\delta_1(x, y) = \delta$ and $\delta_2 = \delta$. This concept is helpful in stochastic optimization problems. (see Section 3.8).

3.2 Examples of inexact models of a function

Let us provide some examples of inexact models for different optimization tasks.

1. *Smooth convex optimization with Lipchitz continuous gradient*

   Let us assume that a function $f(x)$ is a smooth convex function with L-Lipchitz gradient w.r.t. a norm $\|\cdot\|$, then

   $$0 \leq f(x) - f(y) - \langle \nabla f(y), x - y\rangle \leq \frac{L}{2}\|x - y\|^2 \quad \forall x, y \in Q. \tag{7}$$

   Thus, we have that $\psi_{\delta_k}(x, y) = \langle \nabla f(y), x - y\rangle$ is $(\delta, L, \|\|)$–model of the function f, $f_{\delta_k}(y) = f(y)$, and $\delta_k = 0$ for all $k \geq 0$.

[3] In paper [34] (δ, L, V)–model is defined as (δ, L)–model.

2. *Convex optimization with Holder continuous subgradients*

   Let us assume that a function f is a convex function with Holder continuous subgradients: exists $\nu \in [0,1]$ such that

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L_\nu \|x-y\|^\nu \quad \forall x, y \in Q.$$

   Then (see [6])

$$0 \leq f(x) - f(y) - \langle \nabla f(y), x-y \rangle \leq \frac{L(\delta)}{2}\|x-y\|^2 + \delta \quad \forall x, y \in Q,$$

   where

$$L(\delta) = L_\nu \left[\frac{L_\nu}{2\delta}\frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}$$

   and $\delta > 0$ is a controlling parameter. Hence, $\psi_{\delta_k}(x,y) = \langle \nabla f(y), x-y\rangle$ is $(\delta, L(\delta), \|\|)$–model of the function f.

3. *Composite optimization*

   Let us consider the composite optimization problem [36]:

$$f(x) := g(x) + h(x) \to \min_{x \in Q},$$

   where $g(x)$ is a smooth convex function with L–Lipchitz continuous gradients w.r.t. a $\|\|$ and $h(x)$ is a convex function (not necessarily smooth). For the optimization problem, we have that

$$0 \leq f(x) - f(y) - \langle \nabla g(y), x-y\rangle - h(x) + h(y) \leq \frac{L}{2}\|x-y\|^2 \quad \forall x, y \in Q.$$

   Hence, a function $\psi_{\delta_k}(x,y) = \langle \nabla g(y), x-y\rangle + h(x) - h(y)$ is $(\delta, L(\delta), \|\|)$–model of the function f, $f_{\delta_k}(y) = f(y)$, and $\delta_k = 0$ for all $k \geq 0$.

Note that in papers [29, 34], we give more examples, including the conditional gradient method (Frank–Wolfe) [28], superposition of functions [7, 8], min–min problem [11], saddle point problem [11]. There is an example of (δ, L, V)–model [34] for optimization problem which arises in an electoral model for clustering [12].

3.3 Gradient method

In [29], the following results were obtained using the concept of inexact model of a function. Consider the adaptive gradient method for the optimization problem (3). In Algorithm 1 we assume that we have a starting point x_0, a local approximation L_0 of Lipchitz parameter of a function gradient at a

point x_0. Also, as the input of the algorithm, we feed sequences $\{\delta_k\}_{k\geq 0}$ and $\{\widetilde{\delta}_k\}_{k\geq 0}$. We assume that on every step k, the method has access to $(\delta_k, \bar{L}_{k+1}, \|\|)$–model. In general, a constant $\bar{L}_{k+1}$ can vary from iteration to iteration; we only consider that $(\delta_k, \bar{L}_{k+1}, \|\|)$–model exists. We do not use the constant $\bar{L}_{k+1}$ in Algorithm 1 explicitly; furthermore , our method adapts to this constant. The sequence $\{\widetilde{\delta}_k\}_{k\geq 0}$ represents inexact solutions from Definition 3, which may be zero, constant, or vary from iteration to iteration in different problems.

Algorithm 1 Adaptive gradient method with $(\delta, L, \|\|)$–model

1: **Input:** Starting point x_0, sequences $\{\delta_k\}_{k\geq 0}$, $\{\widetilde{\delta}_k\}_{k\geq 0}$ and $L_0 > 0$.
2: $L_1 := \frac{L_0}{2}$.
3: **for** $k \geq 0$ **do**
4: Find a minimal integer $i_k \geq 0$ such that

$$f_{\delta_k}(x_{k+1}) \leq f_{\delta_k}(x_k) + \psi_{\delta_k}(x_{k+1}, x_k) + \frac{L_{k+1}}{2}\|x_k - x_{k+1}\|_2^2 + \delta_k, \quad (8)$$

where $L_{k+1} = 2^{i_k - 1}L_k$, $A_{k+1} := A_k + \frac{1}{L_{k+1}}$.

$$\phi_{k+1}(x) := \frac{1}{L_{k+1}}\psi_{\delta_k}(x, x_k) + V[x_k](x), \quad x_{k+1} := \arg\min_{x\in Q}^{\widetilde{\delta}_k} \phi_{k+1}(x).$$

5: **end for**

In [29], the following convergence rate is derived for Algorithm 1.

Theorem 1 ([29]). *Let $V[x_0](x_*) \leq R^2$, where x_0 is a starting point, x_* is the closest point to x_0 in terms of Bregman divergence, a function f is a convex function, for δ_k and x_k from Algorithm 1 we can always find a constant $\bar{L}_{k+1} > 0$ such that $(\delta_k, \bar{L}_{k+1}, \|\|)$–model $\psi_{\delta_k}(\cdot, x_k)$ exists at a point x_k, and $\bar{x}_N = \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}x_{k+1}$. For Algorithm 1, the following convergence rate holds:*

$$f(\bar{x}_N) - f(x_*) \leq \frac{R^2}{A_N} + \frac{1}{A_N}\sum_{k=0}^{N-1}\widetilde{\delta}_k + \frac{2}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k.$$

If we additionally assume that on every step k, inexact model $(\delta_k, L, \|\|)$–model exists with a fixed parameter L (in other words, $\bar{L}_k \leq L$ for all $k \geq 0$), then

$$f(\bar{x}_N) - f(x_*) \leq \frac{2LR^2}{N} + \frac{2L}{N}\sum_{k=0}^{N-1}\widetilde{\delta}_k + \frac{2}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k. \quad (9)$$

There are three terms in (9) from Theorem 1: convergence rate, accumulated error from auxiliary problems, and accumulated error from the inexact

model of a function. For the simplicity of the analysis, let us suppose that $\widetilde{\delta}_k = \widetilde{\delta}$ and $\delta_k = \delta$ for all $k \geq 0$, then from (9), we can get a more convenient convergence rate estimate:

$$f(\bar{x}_N) - f(x_*) \leq \frac{2LR^2}{N} + 2L\widetilde{\delta} + \delta. \tag{10}$$

From the last inequality, we can conclude that the derived convergence rate corresponds to the convergence rate of the nonaccelerated gradient method [13], while errors $\widetilde{\delta}$ and δ do not accumulate with the number of algorithm iterations. In Section 3.4, we consider the accelerated version of the proposed algorithm, which has a different nature of the convergence rate with respect to $\widetilde{\delta}$ and δ.

Note that we use brute–force search from 0 to infinity in order to find an integer i_k in Algorithm 1. However, the assumption about the existence of $(\delta_k, L_{k+1}, \|\|)$–model at a point x_k ensures that this process is finite. Moreover, we can show that in "average" an integer i_k for which (8) is satisfied is equal 1 (see [6], p. 7–8). Therefore, in "average", $\psi_{\delta_k}(\cdot, x_k)$ is requested 2 times in every iteration of Algorithm 1.

3.4 Fast gradient method

Let us consider the accelerated version of the algorithm from Section 3.3. In [29], we propose Algorithm 2 and prove the corresponding Theorem 2.

Theorem 2 ([29]). *Let $V[x_0](x_*) \leq R^2$, where x_0 is a starting point, x_* is the closest point to x_0 in terms of Bregman divergence, a function f is a convex function and for δ_k and y_{k+1} from Algorithm 2 we can always find a constant $\bar{L}_{k+1} > 0$ such that $(\delta_k, \bar{L}_{k+1}, \|\|)$–model $\psi_{\delta_k}(\cdot, y_{k+1})$ exists at a point y_{k+1}. For Algorithm 2, the following convergence rate holds:*

$$f(x_N) - f(x_*) \leq \frac{R^2}{A_N} + \frac{\sum_{k=0}^{N-1} \widetilde{\delta}_k}{A_N} + \frac{2\sum_{k=0}^{N-1} \delta_k A_{k+1}}{A_N}.$$

If we additionally assume that on every step k, inexact model $(\delta_k, L, \|\|)$–model exists with a fixed parameter L (in other words, $\bar{L}_k \leq L$ for all $k \geq 0$), then

$$f(x_N) - f(x_*) \leq \frac{8LR^2}{(N+1)^2} + \frac{8L\sum_{k=0}^{N-1} \widetilde{\delta}_k}{(N+1)^2} + \frac{2\sum_{k=0}^{N-1} \delta_k A_{k+1}}{A_N}. \tag{11}$$

As in Section 3.3, we assume that $\widetilde{\delta}_k = \widetilde{\delta}$ and $\delta_k = \delta$ for all $k \geq 0$, then from the inequality (11), we have:

$$f(x_N) - f(x_*) \leq \frac{8LR^2}{(N+1)^2} + \frac{8L\widetilde{\delta}}{N+1} + N\delta.$$

Algorithm 2 Adaptive fast gradient method with $(\delta, L, \|\|)$–model

Input: Starting point x_0, sequences $\{\delta_k\}_{k\geq 0}$, $\{\widetilde{\delta}_k\}_{k\geq 0}$ and $L_0 > 0$.

$y_0 := x_0$, $u_0 := x_0$, $L_1 := \frac{L_0}{2}$, $\alpha_0 := 0$, $A_0 := \alpha_0$.

for $k \geq 0$ **do**

Find a minimal integer $i_k \geq 0$ such that

$$f_{\delta_k}(x_{k+1}) \leq f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_{k+1}, y_{k+1}) + \frac{L_{k+1}}{2}\|x_{k+1} - y_{k+1}\|^2 + \delta_k,$$

where $L_{k+1} = 2^{i_k - 1}L_k$, α_{k+1} it the largest root of $A_k + \alpha_{k+1} = L_{k+1}\alpha_{k+1}^2$, $A_{k+1} := A_k + \alpha_{k+1}$.

$$y_{k+1} := \frac{\alpha_{k+1}u_k + A_k x_k}{A_{k+1}},$$

$$\phi_{k+1}(x) = V[u_k](x) + \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}),$$
$$u_{k+1} := \arg\min_{x\in Q}^{\widetilde{\delta}_k} \phi_{k+1}(x),$$

$$x_{k+1} := \frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}.$$

end for

Comparing the last inequality with (10) we can conclude that Algorithm 2 has the convergence rate of the fast gradient method, while an error δ linearly accumulates with the number of the algorithm iterations. In particular, the impact of an error $\widetilde{\delta}$ decreases. If we compare methods w.r.t. $\widetilde{\delta}$, then Algorithm 2 is more effective than Algorithm 1. While the conclusion w.r.t. an error δ is not so unequivocal and depends on an error δ. More details reader can find in the paper [14].

Similarly, as in Section 3.3, we can conclude, that in "average" an integer i_k is equal to 1 [6].

3.5 Gradient method with relative smoothness

Let us consider a simplified version of the algorithm from Section 3.3. The following method works with functions supported by an oracle from Definition 5 with relative smoothness. The optimization method from Section 3.3 (Algorithm 1) is not applicable to numerous optimization problems (see [33]). Further, we consider Algorithm 3 and corresponding Theorem 3.

In this section, we relax the assumption about a prox–function d and replace 1–strong convexity condition with the only convexity of a function d. This allows us to apply Theorem 3 in more general cases.

Algorithm 3 Gradient method with (δ, L, V)–model

1: **Input:** Starting point x_0, $L > 0$ and $\delta, \widetilde{\delta} > 0$.
2: **for** $k \geq 0$ **do**
3:

$$\phi_{k+1}(x) := \psi_\delta(x, x_k) + LV[x_k](x), \quad x_{k+1} := \arg\min_{x \in Q}^{\widetilde{\delta}} \phi_{k+1}(x). \quad (12)$$

4: **end for**

Theorem 3 ([34]). *Let $V[x_0](x_*) \leq R^2$, where x_0 is a starting point, is the closest point to x_0 in terms of Bregman divergence, a function f is a convex function, (δ, L, V)–model $\psi_\delta(\cdot, x_k)$ exists for a function f on a set Q, and $\bar{x}_N = \frac{1}{N}\sum_{k=0}^{N-1} x_{k+1}$. For Algorithm 3, the following convergence rate holds:*

$$f(\bar{x}_N) - f(x_*) \leq \frac{LR^2}{N} + \widetilde{\delta} + \delta.$$

It would be natural to develop the fast gradient descent with relative smoothness by analogy with Section 3.4. However, in general, for optimization problems supported by relative smoothness, convergences rate of nonaccelerated methods can not be improved up to a constant factor (see [37]).

3.6 Primal–dual adaptive gradient method

In this section, we consider the primal–dual gradient method. The main goal of primal–dual type methods is to find ε–solution of both the primal problem (3) and the corresponding dual problem. Let us introduce additional assumption on a set Q. Consider the following setup for a set Q:

$$Q = \{x \mid x \in \widetilde{Q},\ f_i(x) \leq 0\ \forall i \in [1, m]\}, \tag{13}$$

where for all i a function $f_i(x) : \widetilde{Q} \to \mathbb{R}$ is a convex function, and a set $\widetilde{Q}$ is a convex and closed set. Let us define a vector-valued function F:

$$F(x) = [f_1(x), \ldots, f_m(x)]^T.$$

Thus, we rewrite (3) as

$$f(x) \to \min_{x \in \widetilde{Q},\ F(x) \leq 0}. \tag{14}$$

Let us construct the Lagrange dual problem. Using a definition

$$g(z) = \max_{x \in \widetilde{Q}} [-f(x) - \langle z, F(x) \rangle]. \tag{15}$$

we obtain the dual problem for the primal problem (14):

$$g(z) \to \min_{z \in \mathbb{R}^m_+}. \tag{16}$$

From now on, consider the strong duality assumption [21] (p. 226).

The feature to restore a solution of a dual problem is proven to be very useful in various optimization tasks for which it is faster to find an optimal point in a primal problem than in a dual problem. For instance, this property is used in transportation tasks [9, 10, 11].

Definition 7. *Let a point x_* be a solution of a primal optimization problem*

$$p(x) \to \min_{x \in \widetilde{Q},\ G(x) \leq 0}. \tag{17}$$

A point z_* is a solution of a dual optimization problem

$$h(z) \to \min_{z \in \mathbb{R}^m_+},$$

for (17)*, where z are dual variable w.r.t. constraints $G(x) \leq 0$. We define operator* argdual *that depends on functions p and G and returns points x_* and z_*:*

$$(x_*, z_*) := \operatorname*{argdual}_{x \in \widetilde{Q}}(p(x), G(x)).$$

Algorithm 4 Primal–dual adaptive gradient method with $(\delta, L, \|\|)$–model

1: **Input:** Starting point x_0, $L_0 > 0$, and sequence $\{\delta_k\}_{k\geq 0}$.
2: $A_0 := 0$
3: **for** $k \geq 0$ **do**
4: Find a minimal integer $i_k \geq 0$ such that

$$f_{\delta_k}(x_{k+1}) \leq f_{\delta_k}(x_k) + \psi_{\delta_k}(x_{k+1}, x_k) + \frac{L_{k+1}}{2}\|x_{k+1} - x_k\|^2 + \delta_k,$$

where $L_{k+1} := 2^{i_k - 1}L_k$, $A_{k+1} := A_k + \frac{1}{L_{k+1}}$.

$$\begin{aligned} \phi_{k+1}(x) &:= \psi_{\delta_k}(x, x_k) + L_{k+1}V[x_k](x), \\ (x_{k+1}, z_{k+1}) &:= \underset{x\in\widetilde{Q}}{\text{argdual}}(\phi_{k+1}(x), F(x)). \end{aligned} \tag{18}$$

5: **end for**

In [31], we propose Algorithm 4 and corresponding Theorem 4.

Theorem 4 ([31]). *Let* $\bar{x}_N = \frac{1}{A_N}\sum_{k=0}^{N-1}\frac{x_{k+1}}{L_{k+1}}$, $\bar{z}_N = \frac{1}{A_N}\sum_{k=0}^{N-1}\frac{z_{k+1}}{L_{k+1}}$, $V[x_0](x(\bar{z}_N)) \leq R^2$, x_0 *is a starting point,* $x(\bar{z}_N)$ *is the maximum point in* (15) *with* $z = \bar{z}_N$, *a function* f *is a convex function, and for* δ_k *and* x_k *from Algorithm 4 we can always find a constant* $\bar{L}_{k+1} > 0$ *such that* $(\delta_k, \bar{L}_{k+1}, \|\|)$*–model* $\psi_{\delta_k}(\cdot, x_k)$ *exists at a point* x_k. *For Algorithm 4, the following convergence rate holds:*

$$f(\bar{x}_N) + g(\bar{z}_N) \leq \frac{R^2}{A_N} + \frac{1}{A_N}\sum_{k=0}^{N-1}\frac{2\delta_k}{L_{k+1}}.$$

The theorem fully agrees with results from Theorem 1, taking into account new assumptions about a set Q and condition, that $\widetilde{\delta}_k = 0$ for all $k \geq 0$. However, in Theorem 4, the convergence rate is proved for a duality gap $f(\bar{x}_N) + g(\bar{z}_N)$.

There are different approaches to restore a dual ε–solution while ε–solution of a primal task is calculating. In a series of our papers [20, 38], dual variables are recovering with the Lagrange function of optimization problem (16) as a method works. With this approach, conditions in the optimization problem are violated. In an alternative approach [23], dual variables are recovering using an auxiliary problem (see, for example, (18)); however, worser duality gap bounds can be obtained. Indeed, we have $V[x_0](x(\bar{z}_N))$ instead of $V[x_0](x_*)$. Methods from this and the next section inherit the idea from [23].

3.7 Primal–dual adaptive fast gradient method

In this section, we consider the accelerated version of the algorithm from Section 3.6. Let us study the same assumption on a set Q as in Section 3.6. In [31], we propose Algorithm 5 and prove the corresponding Theorem 5.

Algorithm 5 Primal–dual adaptive fast gradient method with $(\delta, L, \|\|)$–model

Input: Starting point x_0, sequence $\{\delta_k\}_{k\geq 0}$ and $L_0 > 0$.
$y_0 := x_0$, $u_0 := x_0$, $L_1 := \frac{L_0}{2}$, $\alpha_0 := 0$, $A_0 := \alpha_0$.
for $k \geq 0$ **do**
Find a minimal integer $i_k \geq 0$ such that

$$f_{\delta_k}(x_{k+1}) \leq f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_{k+1}, y_{k+1}) + \frac{L_{k+1}}{2}\|x_{k+1} - y_{k+1}\|^2 + \delta_k,$$

where $L_{k+1} = 2^{i_k - 1}L_k$, α_{k+1} is the largest root of $A_k + \alpha_{k+1} = L_{k+1}\alpha_{k+1}^2$, $A_{k+1} := A_k + \alpha_{k+1}$.

$$y_{k+1} := \frac{\alpha_{k+1}u_k + A_k x_k}{A_{k+1}},$$

$$\phi_{k+1}(x) := \psi_{\delta_k}(x, y_{k+1}) + L_{k+1}V[u_k](x),$$
$$(x_{k+1}, z_{k+1}) := \operatorname*{argdual}_{x\in\widetilde{Q}}(\phi_{k+1}(x), F(x)).$$

$$x_{k+1} := \frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}.$$

end for

Theorem 5 ([31]). *Let $\bar{z}_N = \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}z_{k+1}$, $V[x_0](x(\bar{z}_N)) \leq R^2$, x_0 is a starting point, $x(\bar{z}_N)$ is the maximum point in (15) with $z = \bar{z}_N$, a function f is a convex function, and for δ_k and y_{k+1} from Algorithm 5 we can always find a constant $\bar{L}_{k+1} > 0$ such that $(\delta_k, \bar{L}_{k+1}, \|\|)$–model $\psi_{\delta_k}(\cdot, y_{k+1})$ exists at a point y_{k+1}. For Algorithm 5, the following convergence rate holds:*

$$f(x_N) + g(\bar{z}_N) \leq \frac{R^2}{A_N} + \frac{2}{A_N}\sum_{k=0}^{N-1} A_{k+1}\delta_k.$$

With the assumption (13) about a set Q, the derived convergence rate fully agrees with the convergence rate from Theorem 2 if we consider, that $\widetilde{\delta}_k = 0$ for all $k \geq 0$.

3.8 Stochastic fast gradient method

Let us consider $(\delta_1, \delta_2, L, \|\|)$–model from Definition 6 that generalizes $(\delta, L, \|\|)$–model of a function. Similarly to Section 3.3 and 3.4, in paper [35], we provide convergence rates for methods that work with $(\delta_1, \delta_2, L, \|\|)$–model. One of the most important consequences is that this concept is surprisingly well–suited for stochastic optimization problems [39, 40]. In Algorithm 6, we provide the fast gradient method with $(\delta_1, \delta_2, L, \|\|_2)$–model.

Algorithm 6 Fast gradient method with $(\delta_1, \delta_2, L, \|\|_2)$–model

1: **Input:** Starting point x_0 and $L > 0$.
2: $y_0 := x_0$, $u_0 := x_0$, $\alpha_0 := 0$, $A_0 := \alpha_0$.
3: **for** $k \geq 0$ **do**
4: Constant α_{k+1} is the largest root of $A_k + \alpha_{k+1} = L\alpha_{k+1}^2$, $A_{k+1} := A_k + \alpha_{k+1}$.

$$y_{k+1} := \frac{\alpha_{k+1}u_k + A_k x_k}{A_{k+1}},$$

$$\phi_{k+1}(x) = \frac{1}{2}\|x - u_k\|_2^2 + \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}),$$
$$u_{k+1} := \arg\min_{x \in Q} \phi_{k+1}(x),$$

$$x_{k+1} := \frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}.$$

5: **end for**

Now, we formulate the convergence rate theorem for $(\delta_1, \delta_2, L, \|\|)$–model.

Theorem 6 ([35]). *Let $\frac{1}{2}\|x_* - x_0\|_2^2 \leq R^2$, where x_0 is a starting point, x_* is the closest point to x_0 in terms of euclidean distance, a function f is a convex function, $(\delta_1^k, \delta_2^k, L, \|\|_2)$–model $\psi_{\delta_k}(\cdot, y_{k+1})$ exists at a point y_{k+1} from Algorithm 6. For Algorithm 6, the following convergence rate holds:*

$$\begin{aligned} f(x_N) - f(x_*) \leq \frac{4LR^2}{N^2} &+ \frac{1}{A_N}\sum_{k=0}^{N-1} A_k \delta_1^k(x_k, y_{k+1}) \\ &+ \frac{1}{A_N}\sum_{k=0}^{N-1} \alpha_{k+1}\delta_1^k(x_*, y_{k+1}) + \frac{1}{A_N}\sum_{k=0}^{N-1} A_{k+1}\delta_2^k. \end{aligned} \tag{19}$$

If we additionally assume that $\{\delta_1^k\}_{k=0}^{N-1}$ and $\{\delta_2^k\}_{k=0}^{N-1}$ are random sequences with assumptions:

Assumption 1. *Given two sequences $\delta_1^k(y, x)$ and δ_2^k $(k \geq 0)$. Assume that*

$$\mathbb{E}\left[\delta_1^k(y,x)|\delta_{1,2}^{k-1},\delta_{1,2}^{k-2},...\right]=0,\ \textit{(conditionally unbiased)}$$

$\delta_1^k(y,x)$ *has* $\left(\hat{\delta}_1\right)^2$*–subgaussion conditional variance,* $\sqrt{\delta_2^k}$ *has* $\hat{\delta}_2$*–subgaussion conditional second moment.*

Assumption 2. *Given two sequences* $\delta_1^k(x,y)$ *and* δ_2^k *(*$k\geq 0$*). The random variable* $\delta_1^k(x,y)$ *has* $\left(\hat{\delta}_1^k(x-y)\right)^2$*–subgaussion conditional moment (*$\hat{\delta}_1^k(\cdot)$ *is a non–random function of one variable) such that*

1. $\hat{\delta}_1^k(\alpha z)\leq\alpha\hat{\delta}_1^k(z)$ *for all* $\alpha\geq 0$ *and* $z\in B(0,R)$.
2. $\hat{\delta}_1<+\infty$*, where* $\hat{\delta}_1\geq\sup_{z\in B(0,R)}\hat{\delta}_1^k(z)$.

With high probability [4]

$$f(x_N)-f(x_*)=\tilde{O}\left(\frac{LR^2}{N^2}+\frac{\hat{\delta}_1}{\sqrt{N}}+N\hat{\delta}_2\right).$$

Moreover,

$$\mathbb{E}[f(x_N)]-f(x_*)=O\left(\frac{LR^2}{N^2}+N\hat{\delta}_2\right).$$

The stochastic optimization problem is an important case that can be described by $(\delta_1,\delta_2,L,\|\|)$–model. Let us consider the following optimization task:

$$f(x)=\mathbb{E}[f(x,\xi)]\to\min_{x\in Q},\tag{20}$$

where a set Q is a convex and closed set, ξ is a random variable, the expected value $\mathbb{E}[f(x,\xi)]$ is well–defined and finite for all $x\in Q$, a function f is a convex function with L–Lipschitz continuous gradient, $\nabla f(y,\xi)$ has subgaussian distribution with subgaussian variance σ^2. For optimization tasks (20), we can take $\psi_\delta(x,y)=\langle\nabla f(y,\xi),x-y\rangle$, and we can show (see. [35]) that for sequences $\{\delta_k^1\}_{k=0}^{N-1}$ and $\{\delta_k^2\}_{k=0}^{N-1}$, the following bounds hold: $\hat{\delta}_1=O(\sigma R)$ and $\hat{\delta}_2=O(\sigma^2/L)$. The optimal convergence rate can be obtained for the task (20) with the help of a mini–batch technique (see [35]).

Note that the same reasoning can be applied to composite and min–max optimization tasks.

3.9 Heuristic adaptive stochastic fast gradient method

In [41], we propose the heuristic adaptive stochastic fast gradient method based on the adaptive fast gradient method (Algorithm 2) and the nonadaptive stochastic gradient method [35] (Algorithm 6). For now, it is an open

[4] it means that with probability $\geq 1-\gamma$, and $\tilde{O}(\cdot)$ means the same as $O(\cdot)$; however, a constant factor depends on $\ln(1/\gamma)$.

question, if it is possible to add adaptivity to the stochastic fast gradient method in order to preserve convergence rate estimates. Various attempts were made in [42, 43, 44, 45, 46, 47]. A more detailed analysis reader can find in [41]. Let us define a mini–batch of gradients as

$$\widetilde{\nabla}^{m_{k+1}} f(y) = \frac{1}{m_{k+1}} \sum_{j=1}^{m_{k+1}} \nabla f(y; \xi_j),$$

and mini–batch of functions values as

$$f^{m_{k+1}}(y) = \frac{1}{m_{k+1}} \sum_{j=1}^{m_{k+1}} f(y; \xi_j),$$

where ξ_j are random variables ($j = 1, ..., m_{k+1}$), $\nabla f(y; \xi_j)$ and $f(y; \xi_j)$ are unbiased estimates of $\nabla f(y)$ and $f(y)$, and m_{k+1} is the number of elements in the mini–batch. In Algorithm 7, we present the heuristic method (see [41]) that works with stochastic gradients and stochastic function values. In [35], using an inexact model from Definition 6, we have the theorem that shows the convergence rate estimate for the nonadaptive version of Algorithm 7.

Note that in paper [41], we have the approbation of our algorithm with the help of experiments. Using practical machine learning tasks MNIST [48] and CIFAR [49], we show that our algorithm convergence faster than popular optimization methods Adam [50] and AdaGrad [51] with logistic regression loss function and linear, neural network, and convolutional neural network backbones.

4 Conclusion

This thesis is based on published papers [29, 34, 31, 35, 41, 38, 20, 52].

In papers [29, 34, 31, 41], we developed optimization methods that exploit the concepts of inexact model. Also, we demonstrate various examples of optimization tasks supported by suitable inexact models. In addition to standard tasks from structural optimization like smooth optimization, composite optimization, optimization with Holder continuous subgradients, the proposed concepts of inexact model can describe transportation tasks [29, 11], optimization problem which arises in an electoral model for clustering [34, 12], etc.

Papers [41, 38, 20, 52] are milestones from the view of the development of the concept of inexact model. Moreover, they motivate further research.

Let us list the main results that are obtained in this thesis and submitted for defense.

1. Various concepts of inexact model are developed for gradient-type methods. As shown in the thesis, these concepts can represent a significant number of modern optimization problems.

Algorithm 7 Heuristic adaptive stochastic fast gradient method

1: **Input:** Starting point x_0, constants $\epsilon > 0$, $L_0 > 0$ and $\sigma_0^2 > 0$.
2: $y_0 := x_0$, $u_0 := x_0$, $L_1 := \frac{L_0}{2}$, $\alpha_0 := 0$, $A_0 := \alpha_0$.
3: **for** $k \geq 0$ **do**
4: Find a minimal integer $i_k \geq 0$ such that

$$f^{m_{k+1}}(x_{k+1}) \leq f^{m_{k+1}}(y_{k+1}) + \langle \widetilde{\nabla}^{m_{k+1}} f(y_{k+1}), x_{k+1} - y_{k+1} \rangle + \\ + \frac{L_{k+1}}{2} \|x_{k+1} - y_{k+1}\|^2 + \frac{\epsilon}{L_{k+1}\alpha_{k+1}},$$

where $L_{k+1} = 2^{i_k - 1} L_k$, $\widetilde{\alpha}_{k+1}$ it the largest root of $A_k + \alpha_{k+1} = L_k \alpha_{k+1}^2$, α_{k+1} it the largest root of $A_k + \alpha_{k+1} = L_{k+1}\alpha_{k+1}^2$, $A_{k+1} := A_k + \alpha_{k+1}$, $m_{k+1} := \left\lceil \frac{3\sigma_0^2 \tilde{\alpha}_{k+1}}{\epsilon} \right\rceil$. If $i_k = 0$, then generate i.i.d. ξ_j $(j = 1, ..., m_{k+1})$.

$$y_{k+1} := \frac{\alpha_{k+1} u_k + A_k x_k}{A_{k+1}},$$

$$\phi_{k+1}(x) := \frac{1}{2} \|x - u_k\|_2^2 + \alpha_{k+1} \left(f^{m_{k+1}}(y_{k+1}) + \langle \widetilde{\nabla}^{m_{k+1}} f(y_{k+1}), x - y_{k+1} \rangle \right),$$

$$u_{k+1} := \arg\min_{x \in Q} \phi_{k+1}(x),$$

$$x_{k+1} := \frac{\alpha_{k+1} u_{k+1} + A_k x_k}{A_{k+1}}.$$

5: **end for**

2. Different gradient methods are proposed which support concepts of inexact model. For methods from sections 3.3–3.8, we proved corresponding convergence rate theorems and performed analysis.

3. The heuristic adaptive stochastic fast gradient method is developed and justified.

4. Theoretical analysis of primal-dual methods for problems with strongly convex functionals with a simple structure under affine constraints, problems that calculate regularized optimal transport, and problems with the concept of inexact model is carried out.

It is worthwhile to say that some ideas and several examples of optimization problems that are well-described by the concept of inexact model are not listed:

1. In further research, we are planning to develop methods that work with strongly convex functions [30].

2. In [52], we consider the practical optimization task with an objective function that has the form of a sum of smooth strongly convex functions with a smooth regularizer. In this paper, we propose an approach that allows us to derive the optimal bound for the case when the composite part of an objective function is not proximal–friendly.[5]. In further research, we are planning to generalize this result with the concept of inexact model.

3. We are planning to combine the concept of inexact model with coordinate descent methods [53, 54]. As in general stochastic optimization, coordinate descent methods admit efficient optimization steps by calculating a descent direction via a subset of coordinates.

4. In the end of Section 3.6, we mention the described in [20, 38] approach of recovering of dual variables. It would be useful to generalize methods from [20, 38] using the concept of inexact model.

5 References

[1] *Kantorovich L.* Mathematical methods of organizing and planning production // Management Science. 1960. V. 6, №. 4. P. 366–422.

[2] *Polyak B.* History of mathematical programming in the USSR: analyzing the phenomenon // Math. Program. 2002. V. 91. № 3. P. 401–416.

[5] A proximal–friendly function is a function that, with a quadratic function, can be "simply" minimized.

[3] *Nemirovski A., Yudin D.* Problem complexity and method efficiency in optimization. Wiley–Interscience. 1983.

[4] *Nesterov Yu.* A method of solving a convex programming problem with convergence rate $O(1/k^2)$ // Dokl. Akad. Nauk SSSR. 1983. V. 269. № 3. P. 543–547.

[5] *Nesterov Yu.* Gradient methods for minimizing composite functions // Math. Program. 2013. V. 140, № 1. P. 125–161.

[6] *Nesterov Yu.* Universal gradient methods for convex optimization problems // Math. Program. 2015. V. 152. № 1–2. P. 381–404.

[7] *Nemirovski A.* Information-based complexity of convex programming. Technion. 1995.

[8] *Lan G.* Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization // Math. Program. 2015. V. 149. № 1–2. P. 1–45.

[9] *Baimurzina D., Gasnikov A., Gasnikova E., Kubentaeva M., Lagunovskaya A., Dvurechensky P., Ershov E.* Universal Method of Searching for Equilibria and Stochastic Equilibria in Transportation Networks // Computational Mathematics and Mathematical Physics. 2019. V. 59. № 1. P. 19–33.

[10] *Gasnikov A., Gasnikova E., Nesterov Yu.* Dual methods for finding equilibriums in mixed models of flow distribution in large transportation networks // Computational Mathematics and Mathematical Physics. 2017. V. 58. № 9. P. 1395–1403.

[11] *Gasnikov A.* Effective numerical methods for finding equilibrium in large transport networks. PhD thesis. MFTI, 2016.

[12] *Nesterov Yu.* Soft clustering by convex electoral model // CORE Discussion paper. 2018/01. 20p. URL: `https://alfresco.uclouvain.be/alfresco/service/guest/streamDownload/workspace/SpacesStore/ff42ec88-4339-4223-b05d-b768c71ef4e6/coredp2018_01web.pdf?guest=true`.

[13] *Nesterov Yu.* Lectures on convex optimization. Springer. 2018.

[14] *Devolder O., Glineur F., Nesterov Yu.* First-order methods of smooth convex optimization with inexact oracle // Math. Program. 2014. V. 146. № 1–2. P. 37–75.

[15] *Nesterov Yu.* Smooth minimization of non-smooth functions // Math. Program. 2005. V. 103. № 1. P. 127–152.

[16] *Nesterov Yu.* Excessive gap technique in nonsmooth convex minimization // SIAM J. Optimizat. 2005. V. 16, № 1. P. 235–249.

[17] *Nesterov Yu.* Smoothing technique and its applications in semidefinite optimization // Math. Program. 2007. V. 110. № 2. P. 245–259.

[18] *Lemarechal C., Sagaztizabal C.* Practice aspects of Moreau–Yosida regularization: Theoretical preliminaries // SIAM J. Optimizat. 1997. V. 7, № 2. P. 367–385.

[19] *D'Aspremont A,* Smooth optimization with approximate gradient // SIAM J. Optimizat. 2019. V. 19, № 3. P. 1171–1183.

[20] *Anikin A., Gasnikov A., Dvurechensky P., Tyurin A., Chernov A.* Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints // Computational Mathematics and Mathematical Physics. 2017. V. 57. № 8. P. 1262–1276.

[21] *Boyd S., Vandenberghe L.* Convex optimization. Cambridge University Press. 2004.

[22] *Nesterov Yu.* Complexity bounds for primal-dual methods minimizing the model of objective function // Mathematical Programming. 2018. V. 171, №. 1–2. P. 311–330.

[23] *Nesterov Yu.* Primal-dual subgradient methods for convex problems. // Mathematical Programming. 2009. V. 120, № 1. P. 221–259.

[24] *Gasnikov A.* Universal gradient descent // e-print. arXiv:1711.00394. 2020.

[25] *Goodfellow I., Bengio Y., Courville A.* Deep learning. MIT press. 2016.

[26] *Krizhevsky A., Sutskever I., Hinton G.* Imagenet classification with deep convolutional neural networks // In Advances in neural information processing systems. 2012. P. 1097–1105.

[27] *Bregman L.* The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming // USSR Computational Mathematics and Mathematical Physics. 1967. V. 7. № 3. P. 200–217.

[28] *Ben-Tal A., Nemirovski A.* Lectures on Modern Convex Optimization. Philadelphia: SIAM, 2015. URL: `http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf`.

[29] *Gasnikov A., Tyurin A.* Fast gradient descent for convex minimization problems with an oracle producing a (δ, L)-model of function at the requested point // Computational Mathematics and Mathematical Physics. 2019. V. 59. № 7. P. 1085–1097.

[30] *Stonyakin F., Tyurin A., Gasnikov A., Dvurechensky P., Agafonov A., Dvinskikh D., Pasechnyuk D., Artamonov S., Piskunova V.* Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model // e-print. arXiv:2001.09013. 2020.

[31] *Tyurin A.* Primal–dual fast gradient method with a model // Computer Research and Modeling. 2020. V. 12, № 2. P. 263–274.

[32] *Bauschke H., Bolte J., Teboulle M.* A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications // Mathematics of Operations Research. 2016. V. 42. № 2. P. 330–348.

[33] *Lu H., Freund R., Nesterov Yu.* Relatively smooth convex optimization by first-order methods, and applications // SIAM J. Optimizat. 2018. V. 28, № 1. P. 333–354.

[34] *Stonyakin F., Dvinskikh D., Dvurechensky P., Kroshnin A., Kuznetsova O., Agafonov A., Gasnikov A., Tyurin A., Uribe C., Pasechnyuk D., Artamonov S.* Gradient methods for problems with inexact model of the objective // Lecture Notes in Computer Science. 2019. V. 11548. P. 97–114.

[35] *Dvinskikh D., Tyurin A., Gasnikov A., Omelchenko S.* Accelerated and nonaccelerated stochastic gradient descent with model conception // Mathematical Notes. 2020. V. 108. № 4. In press.

[36] *Nesterov Yu.* Gradient methods for minimizing composite functions // Math. Program. 2013. V. 140, № 1. P. 125–161.

[37] *Dragomir R., Taylor A., D'Aspremont A., Bolte J.* Optimal complexity and certification of Bregman first-order methods // e-print. arXiv: 1911.08510. 2019.

[38] *Dvurechensky P., Gasnikov A., Omelchenko A., Tyurin A.* A stable alternative to Sinkhorn's algorithm for regularized optimal transport // Lecture Notes in Computer Science. 2020. V. 12095. P. 406–423.

[39] *Lan G.* Lectures on optimization. Methods for Machine Learning // e-print. 2019. URL: `http://pwp.gatech.edu/guanghui-lan/wp-content/uploads/sites/330/2019/08/LectureOPTML.pdf`.

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

[41] *Ogaltsov A., Tyurin A.* A heuristic adaptive fast gradient method in stochastic optimization problems // Computational Mathematics and Mathematical Physics. 2020. V. 60. № 7. P. 1108–1115.

[42] *Bach F., Levy K.Y.* A universal algorithm for variational inequalities adaptive to smoothness and noise // COLT, 2019.

[43] *Vaswani S., Mishkin A., Laradji I., Schmidt M., Gidel G., Lacoste-Julien S.* Painless Stochastic Gradient: interpolation, line-search, and convergence rates // NIPS, 2019.

[44] *Ward R., Wu X., Bottou L.* AdaGrad stepsizes: sharp convergence over nonconvex landscapes, from any initialization // ICML, 2019.

[45] *Deng Q., Cheng Y., Lan G.* Optimal adaptive and accelerated stochastic gradient descent // e-print. arXiv:1810.00553. 2018.

[46] *Levy K.Y., Yurtsever A., Cevher V.* Online adaptive methods, universality and acceleration // NIPS, 2018.

[47] *Iusem A.N., Jofre A., Oliveira R.I., Thompson P.* Variance-based extragradient methods with line search for stochastic variational inequalities // SIAM J. Optimizat. 2019. V. 29, № 1. P. 175–206.

[48] *LeCun Y., Bottou L., Bengio Y., Haffner P.* Gradient-based learning applied to document recognition // Proceedings of the IEEE. 1998. V. 86. № 11. P. 2278–2324.

[49] *Krizhevsky A.* Learning Multiple Layers of Features from Tiny Images. PhD thesis. University of Toronto, 2009.

[50] *Kingma D.P., Ba J.* Adam: a method for stochastic optimization // ICLR, 2015.

[51] *Duchi J., Hazan E., Singer Y.* Adaptive subgradient methods for online learning and stochastic optimization // Journal of Machine Learning Research. 2011. V. 12. № Jul. P. 2121–2159.

[52] *Dvinskikh D., Omelchenko A., Gasnikov A., Tyurin A.* Accelerated gradient sliding for minimizing the sum of functions // Doklady Mathematics. 2020. V. 101. № 3. In press.

[53] *Dvurechensky P., Gasnikov A., Tyurin A.* Randomized similar triangles method: a unifying framework for accelerated randomized optimization methods (coordinate descent, directional search, derivative–free method) // e-print. arXiv: 1707.08486. 2017.

[54] *Nesterov Yu.* Efficiency of coordinate descent methods on huge–scale optimization problems // SIAM J. Optimizat. 2012. V. 22, № 2. P. 341–362.