**Weierstraß-Institut
für Angewandte Analysis und Stochastik**

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model

Fedor Stonyakin[1,6], Alexander Gasnikov[1,2,3], Alexander Tyurin[3], Dmitry Pasechnyuk[4], Artem Agafonov[1], Pavel Dvurechensky[5], Darina Dvinskikh[5], Sergei Artamonov[3], Victorya Piskunova[6]

submitted: April 2, 2020

[1] Moscow Institute of Physics and Technology
Dolgoprudny, Russia
E-Mail: fedyor@mail.ru
gasnikov@yandex.ru
agafonov.ad@phystech.edu

[2] Institute for Information Transmission Problems
Moscow, Russia
E-Mail: gasnikov@yandex.ru

[3] Higher School of Economics
Moscow, Russia
E-Mail: alexandertiurin@gmail.com
gasnikov@yandex.ru
artamonov@yandex.ru

[4] Presidential Physics and Mathematics Lyceum No.239
St. Petersburg, Russia
E-Mail: pasechnyuk2004@gmail.com

[5] Weierstrass Institute
Berlin, Germany
E-Mail: darina.dvinskikh@wias-berlin.de
pavel.dvurechensky@wias-berlin.de

[6] V. Vernadsky Crimean Federal University
Simferopol, Crimea
E-Mail: piskunova@mmail.ru
fedyor@mail.ru

No. 2709
Berlin 2020

2010 *Mathematics Subject Classification.* 90C30, 90C25, 68Q25, 65K15.

Key words and phrases. Convex optimization, composite optimization, proximal method, level-set method, variational inequality, universal method, mirror prox, acceleration, relative smoothness, saddle-point problem.

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

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

Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model

Fedor Stonyakin, Alexander Gasnikov, Alexander Tyurin, Dmitry Pasechnyuk, Artem Agafonov, Pavel Dvurechensky, Darina Dvinskikh, Sergei Artamonov, Victorya Piskunova

Abstract

In this paper we propose a general algorithmic framework for first-order methods in optimization in a broad sense, including minimization problems, saddle-point problems and variational inequalities. This framework allows to obtain many known methods as a special case, the list including accelerated gradient method, composite optimization methods, level-set methods, Bregman proximal methods. The idea of the framework is based on constructing an inexact model of the main problem component, i.e. objective function in optimization or operator in variational inequalities. Besides reproducing known results, our framework allows to construct new methods, which we illustrate by constructing a universal conditional gradient method and universal method for variational inequalities with composite structure. These method works for smooth and non-smooth problems with optimal complexity without a priori knowledge of the problem smoothness. As a particular case of our general framework, we introduce relative smoothness for operators and propose an algorithm for VIs with such operator. We also generalize our framework for relatively strongly convex objectives and strongly monotone variational inequalities.

1 Introduction

In this paper we consider the following convex optimization problem

$$\min_{x\in Q} f(x), \tag{1}$$

where Q is a convex subset of finite-dimensional vector space E, f is generally a non-convex function.

Most of minimization methods for such problems are constructed using some model of the objective f at the current iterate x_k. This can be a quadratic model based on the L-smoothness of the objective

$$f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|_2^2. \tag{2}$$

The step of gradient method is obtained by the minimization of this model [56]. More general models are constructed based on regularized second-order Taylor expansion [59] or other Taylor-like models [16] as well as other objective surrogates [46]. Another example is the conditional gradient method [26], where a linear model of the objective is minimized on every iteration. Adaptive choice of the parameter of the model with provably small computational overhead was proposed in [59] and applied to first-order methods in [22, 52, 53]. Recently, first-order optimization methods were generalized to the so-called relative smoothness framework [6, 45, 60], where $\frac{1}{2}\|x-x_k\|_2^2$ in the quadratic model (2) for the objective is replaced with general Bregman divergence.

DOI 10.20347/WIAS.PREPRINT.2709 Berlin 2020

The literature on first-order methods [8, 15, 21] considers also gradient methods with inexact information, relaxing the model (2) to

$$f_\delta(x_k) + \langle \nabla f_\delta(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|_2^2 + \delta, \tag{3}$$

with $(f_\delta, \nabla f_\delta)$ called inexact oracle and this model being an upper bound for the objective. In particular, this relaxation allows to obtain universal gradient methods [53].

One of the goals of this paper is to describe and analyze first-order optimization methods which use a very general *inexact model* of the objective function, the idea being to replace the linear part in (3) by a general function $\psi_\delta(x, x_k)$ and the squared norm by general Bregman divergence. The resulting model includes as a particular case inexact oracle model and relative smoothness framework, and allows to obtain many optimization methods as a particular case, including conditional gradient method [26], Bregman proximal gradient method [11] and its application to optimal transport [69] and Wasserstein barycenter [65] problems, general Catalyst acceleration technique [44], (accelerated) composite gradient methods [7, 52], (accelerated) level methods [42, 50]. First attempts to propose this generalization were made in [29, 65] for non-accelerated methods and in [31] for accelerated methods, yet without relative smoothness paradigm. In this paper we propose the inexact model in a very general setting including adaptivity of the algorithms to the parameter L, possible relative strong convexity and relative smoothness. We also provide convergence rates for the gradient method and accelerated gradient method using inexact model of the objective. As an application of our general framework, we develop a universal conditional gradient method, providing a parameter-free generalization of the results in [54].

We believe that our model is flexible enough to be extended for problems with primal-dual structure[1] [49, 51, 54], e.g. for problems with linear constraints [2, 12, 34, 58]; for random block-coordinate descent [25]; for tensor methods [30, 55]; for distributed optimization setting [17, 18, 64, 68]; and adaptive stochastic optimization [36, 61].

Optimization problem (1) is tightly connected with variational inequality (VI)

$$\text{Find } x_* \in Q \text{ s.t. } \langle g(x_*), x_* - x \rangle \le 0, \ \forall x \in Q,$$

where $g(x) = \nabla f(x)$. A special VI is also equivalent to finding a saddle-point of a convex-concave function

$$\min_{u \in Q_1} \max_{v \in Q_2} f(u, v)$$

for $x = (u, v)$ and $g(x) = (\nabla_u f(u, v), -\nabla_v f(u, v))$. This motivates the second part of this paper, which consists in generalization of the inexact model of the objective function to an inexact model for an operator in variational inequality. In particular, we extend the relative smoothness paradigm to variational inequalities with monotone and strongly monotone operators and provide a generalization of Mirror-Prox method [48], its adaptive version [28] (see also [3]) and universal version [24] to variational inequalities with such general inexact model of the operator. As a partucular case, our approach allows to partially reproduce the results of [10]. We also apply the general framework for variational inequalities to saddle-point problems.

To sum up, we present a unified view on inexact models for convex optimization problems, variational inequalities, and saddle-point problems.

The structure of the paper is the following. In Section 2 we introduce inexact model of the objective in optimization and provide several examples to illustrate the flexibility and generality of the proposed

[1] see recent results on this generalization in [67].

model. In particular, we demonstrate that relative smoothness and strong convexity are particular cases of our general framework.

In Section 3 we consider adaptive gradient method (GM) and adaptive fast gradient method (FGM). FGM has better convergence rate, yet it is not adapted to the relative smoothness paradigm. In section 3.3, we construct universal conditional gradient (Frank–Wolfe) method using FGM with inexact projection. To the best of our knowledge, this is the first attempt to combine Frank–Wolfe method [35, 37] and universal method [53]. In Section 4 we generalize inexact model to variational inequalities and saddle-point problems for the case of monotone and strongly monotode operators. In the former case, we construct an adaptive generalization of the Mirror-Prox algorithm for variational inequalities and saddle-point problems with such inexact model. In the latter case the proposed algorithm is accelerated by the restart technique to have linear rate of convergence. We especially consider the case of m-strong convexity of the model. The natural motivation for such a formulation are composite saddle problems, and mixed variational inequalities with a m-strongly convex composite.

The contribution of this paper is follows:

1 We introduce inexact (δ, L, μ, m, V)-model for optimization problems and obtain convergence rates for adaptive GM for optimization problems with this model.

2 We introduce inexact $(\delta, L, \mu, m, V, \|\cdot\|)$-model for optimization problems and obtain convergence rates for adaptive FGM for optimization problems with such model. Using FGM with inexact model we construct a universal conditional gradient (Frank–Wolfe) method.

3 We propose generalizations of the above models, namely (δ, L, V)-model and (δ, L, μ, V)-model for variational inequalities and saddle-point problems. As a special case we introduce relative smoothness for operators in variational inequalities, thus, generalizing [45] from optimization to variational inequalities. We obtain convergence rates for adaptive versions of Mirror-Prox algorithm for problems with inexact model.

2 Inexact Model in Minimization Problems. Definitions and Examples

We start with the general notation. Let E be an n-dimensional real vector space and E^* be its dual. We denote the value of a linear function $g \in E^*$ at $x \in E$ by $\langle g, x\rangle$. Let $\|\cdot\|$ be some norm on E, $\|\cdot\|_*$ be its dual, defined by $\|g\|_* = \max\limits_{x}\big\{\langle g, x\rangle, \|x\| \leq 1\big\}$. We use $\nabla f(x)$ to denote any (sub)gradient of a function f at a point $x \in \mathrm{dom} f$. We define a continuous convex on Q function $d(x)$ to be distance generating function and $V[y](x) = d(x) - d(y) - \langle \nabla d(y), x - y\rangle$ to be the corresponding Bregman divergence. Most typically it is assumed that d is 1-strongly convex on Q w.r.t. $\|\cdot\|$-norm, which we refer to as (1-SC) assumption w.r.t. $\|\cdot\|$-norm. Namely, for all $x, y \in Q$, $d(x) - d(y) - \langle \nabla d(y), x - y\rangle \geqslant \frac{1}{2}\|x - y\|^2$. We underline that, in general, we do not make this assumption, and, in what follows, we explicitly write if this assumption is made.

Definition 1. Let $\delta, L, \mu, m \geq 0$. We say that $\psi_\delta(x, y)$ is a (δ, L, μ, m, V)-model of the function f at a given point y iff, for all $x \in Q$,

$$\mu V[y](x) \leqslant f(x) - (f_\delta(y) + \psi_\delta(x, y)) \leqslant LV[y](x) + \delta. \tag{4}$$

DOI 10.20347/WIAS.PREPRINT.2709

and $\psi_\delta(x,y)$ satisfies $\psi_\delta(x,x)=0$ for all $x\in Q$ and

$$\psi(x) \geqslant \psi(z) + \langle \nabla_z \psi(z), x-z\rangle + mV[z](x) \quad \forall x, z\in Q, \tag{5}$$

where for fixed $y\in Q$ and any $x\in Q$ we denote $\psi(x)=\psi_\delta(x,y)$.

Note that in Definition 1 we allow L to depend on δ. Definition 1 is a generalization of (δ,L)-model from [29, 31, 65], where $\mu=0$ and $m=0$. Further, we denote $(\delta,L,0,0,V)$-model as (δ,L)-model.

Let us illustrate the above definitions by several examples.

Example 2. Composite optimization, [7, 52]. Assume that in (1), $f(x)=g(x)+h(x)$ with L-smooth w.r.t. norm $\|\cdot\|$ part g and simple convex part h. In this case we assume that $V[y](x)$ satisfies (1-SC) condition w.r.t $\|\cdot\|$, and define $f_\delta(x)=f(x)+h(x)$ and $\psi_\delta(x,y)=\langle\nabla g(y),x-y\rangle+h(x)-h(y)$. It is clear that (4) holds with $\delta=0$ and $\mu=0$ and we are in the situation of Definition 1 with $m=0$. If h turns out to be relatively m-strongly convex [45] relatively to d, i.e. $h(x)-h(y)-\langle\nabla h(y),x-y\rangle\geq mV[y](x)$, then (5) holds, but in (4) $\mu=0$. On the other hand, if g turns out to be relatively μ-strongly convex [45] relatively to d, i.e. $g(x)-g(y)-\langle\nabla g(y),x-y\rangle\geq\mu V[y](x)$, then (4) holds with $\delta=0$, but in (5) $m=0$.

A particular example is the following minimization problem [1] motivated by traffic demands matrix estimation from link loads

$$f(x)=\frac{1}{2}\|Ax-b\|_2^2+m\sum_{k=1}^n x_k\ln x_k\to\min_{x\in S_n(1)}.$$

In this case $g(x)=\frac{1}{2}\|Ax-b\|_2^2$ and $h(x)=m\sum\limits_{k=1}^n x_k\ln x_k$. Choosing $\|\cdot\|=\|\cdot\|_1$ and $d(x)=\sum\limits_{k=1}^n x_k\ln x_k$, $V[y](x)=\sum\limits_{k=1}^n x_k\ln\left(x_k/y_k\right)$, we obtain that g is has Lipschitz gradient w.r.t. $\|\cdot\|_1$ with the constant $L=\max_{\|h\|_1\leq 1}\langle h,A^TAh\rangle=\max_{k=1,\dots,n}\|A_k\|_2^2$, where A_k is the k-th column of A. Finally, $\psi_\delta(x,y)=\langle\nabla g(y),x-y\rangle+h(x)-h(y)$ is a $(0,L,0,m,V)$-model. At the same time, the part g is not necessarily strongly convex. Thus, our framework allows to obtain (accelerated) gradient method for composite optimization and their counterparts for inexact oracle models.

Example 3. Relative smoothness and relative strong convexity, [6, 45]. Assume that in (1), the objective f is relatively smooth [6, 45] relative to d, i.e.

$$f(x)-f(y)-\langle\nabla f(y),x-y\rangle\leq(d(x)-d(y)-\langle\nabla d(y),x-y\rangle)=LV[y](x),\ \forall x,y\in Q$$

and relatively strongly convex [45] relative to d, i.e.

$$\mu V[y](x)=\mu\left(d(x)-d(y)-\langle\nabla d(y),x-y\rangle\right)\leq f(x)-f(y)-\langle\nabla f(y),x-y\rangle,\ \forall x,y\in Q.$$

Then, clearly, Definition 1 holds with $m=0$, $\delta=0$, $\psi_\delta(x,y)=\langle\nabla f(y),x-y\rangle$. Importantly, the function d is not necessarily strongly convex.

Note that if $V[y](x)\leq C_n\|x-y\|^2$ for some constant $C_n=O(\log n)$, the condition of (μC_n)-strong convexity w.r.t. norm $\|\cdot\|$, namely $\mu C_n\|x-y\|^2+f_\delta(y)+\psi_\delta(x,y)\leqslant f(x)$ implies the left inequality in (4).

One of the main applications of general relative smoothness and strong convexity is the step of tensor methods which use the derivatives of the objective of the order higher than 2 [30, 55]. Thus, our framework allows to obtain gradient method for optimization with relative smoothness and strong convexity and extend them to the case of inexact oracle setting.

Example 4. Superposition of functions, [50]. Assume that in (1) [42, 50] $f(x) := g(g_1(x), \ldots, g_m(x)) \to \min_{x\in Q}$, where each function $g_k(x)$ is a smooth convex function with L_k-Lipschitz gradient w.r.t. $\|\cdot\|$-norm for all k. Function $g(x)$ is a M-Lipschitz convex function w.r.t 1-norm, non-decreasing in each of its arguments. The chosen Bregman divergence $V[y](x)$ is assumed to satisfy (1-SC). From these assumptions we have [9, 42] that function $f(x)$ is also convex and

$$0 \leq f(x) - f(y) - g(g_1(y) + \langle \nabla g_1(y), x - y\rangle, \ldots, g_m(y) + \langle \nabla g_m(y), x - y\rangle) + f(y) \leq$$
$$\leq M\frac{\sum_{i=1}^m L_i}{2}\|x-y\|^2 \leq MV[y](x)\sum_{i=1}^m L_i, \quad \forall x, y \in Q.$$

Therefore,

$$\psi_\delta(x,y) = g(g_1(y) + \langle \nabla g_1(y), x - y\rangle, \ldots, g_m(y) + \langle \nabla g_m(y), x - y\rangle) - f(y),$$

is $(0, M \cdot (\sum_{i=1}^m L_i))$-model of f with $f_\delta(y) = f(y)$ at a given point y. Thus, our framework allows to obtain (accelerated) level gradient methods considered in [42, 50] as a special case. Moreover, we generalize these methods for the case of inexact oracle information.

Example 5. Proximal method, [11]. Let us consider optimization problem (1), where f is an arbitrary convex function (not necessarily smooth). Then, for arbitrary $L \geq 0$, $\psi_\delta(x,y) = f(x) - f(y)$ is $(0, L)$-model of f with $f_\delta(y) = f(y)$ at a given point y. Thus, our framework allows to obtain (Bregman) proximal gradient methods [11, 63] as a special case and extend them to the case of inexact oracle setting. In particular, based on this model (with Bregman divergence to be Kullback–Leibler divergence) we propose in [65] proximal Sinkhorn's algorithm for Wasserstein distance calculation problem and in [39] proximal IBP for Wasserstein barycenter problem.

Example 6. Min-min problem. Assume that in (1) $f(x) := \min_{z\in Q} F(z,x)$, the set Q is convex and bounded, function F is smooth and convex w.r.t. all variables. Moreover, assume that

$$\|\nabla F(z',x') - \nabla F(z,x)\|_2 \leq L\|(z',x') - (z,x)\|_2, \ \forall z, z' \in Q, \ x, x' \in \mathbb{R}^n.$$

Let $V[y](x) = \frac{1}{2}\|x-y\|_2^2$. If we can find a point $\widetilde{z}_\delta(y) \in Q$ such that

$$\langle \nabla_z F(\widetilde{z}_\delta(y), y), z - \widetilde{z}_\delta(y)\rangle \geq -\delta, \ \ \forall z \in Q,$$

then $F(\widetilde{z}_\delta(y), y) - f(y) \leq \delta$ and $\psi_\delta(x,y) = \langle \nabla_z F(\widetilde{z}_\delta(y), y), x - y\rangle$ is $(6\delta, 2L, 0, 0, V)$-model of f with $f_\delta(y) = F(\widetilde{z}_\delta(y), y) - 2\delta$ at a given point y.

Example 7. Saddle point problem, [15]

Assume that in (1) $f(x) = \max_{z\in Q}\left[\langle x, b - Az\rangle - \phi(z)\right] \to \min_{x\in\mathbb{R}^n}$, where $\phi(z)$ is a μ-strongly convex function w.r.t. p-norm ($1 \leq p \leq 2$). Then f is smooth and convex and its gradient is Lipschitz continuous with constant $L = \frac{1}{\mu}\max_{\|z\|_p\leq 1}\|Az\|_2^2$. If $z_\delta(y) \in Q$ is an approximate solution to auxiliary max-problem, i.e.

$$\max_{z\in Q}\left[\langle y, b - Az\rangle - \phi(z)\right] - \left[\langle y, b - Az_\delta(y)\rangle - \phi(z_\delta(y))\right] \leq \delta,$$

then $\psi_\delta(x,y) = \langle b - Az_\delta(y), x - y\rangle$ is $(\delta, 2L, 0, 0, V)$-model of f with $f_\delta(y) = \langle y, b - Az_\delta(y)\rangle - \phi(z_\delta(y))$ at the point y if we define $V[y](x) = \frac{1}{2}\|x-y\|_2^2$.

Example 8. Augmented Lagrangians, [15]. Let us consider

$$\min_{Az=b,\, z\in Q} \phi(z) + \frac{\mu}{2}\left\|Az-b\right\|_2^2$$

and the corresponding dual problem

$$\min_{x\in\mathbb{R}^n}\left\{ f(x) = \max_{z\in Q} \underbrace{\left(\langle x, b-Az\rangle - \phi(z) - \frac{\mu}{2}\left\|Az-b\right\|_2^2\right)}_{\Lambda(x,z)} \right\}.$$

If $z_\delta(y)$ is an approximate solution of auxiliary max-problem, i.e.

$$\max_{z\in Q}\langle \nabla_z \Lambda(y, z_\delta(y)), z - z_\delta(y)\rangle \leq \delta,$$

then $\psi_\delta(x,y) = \langle b - Az_\delta(y), x-y\rangle$ is $(\delta, \mu^{-1}, 0, 0, V)$-model of f with

$$f_\delta(y) = \langle y, b - Az_\delta(y)\rangle - \phi(z_\delta(y)) - \frac{\mu}{2}\left\|Az_\delta(y)-b\right\|_2^2$$

at the point y if we take $V[y](x) = \frac{1}{2}\|x-y\|_2^2$.

Example 9. Moreau envelope of the objective function, [15]. Let us consider optimization problem:

$$\min_{x\in\mathbb{R}^n}\left\{ f_L(x) := \min_{z\in Q} \underbrace{\left\{ f(z) + \frac{L}{2}\left\|z-x\right\|_2^2\right\}}_{\Lambda(x,z)} \right\}.$$

Assume that f is convex and, for some $z_L(y)$,

$$\max_{z\in Q}\left\{\Lambda(y, z_L(y)) - \Lambda(y,z) + \frac{L}{2}\left\|y - z_L(y)\right\|_2^2\right\} \leq \delta.$$

Then $\psi_\delta(x,y) = \langle L(y - z_L(y)), x-y\rangle$ is $(\delta, L, 0, 0, V)$-model of f with

$$f_\delta(y) = f(z_L(y)) + \frac{L}{2}\left\|z_L(y) - y\right\|_2^2 - \delta$$

at the point y if we take $V[y](x) = \frac{1}{2}\|x-y\|_2^2$.

Example 10. Clustering by Electorial Model, [65].

Another example of an optimization problem that allows for $(\delta, L, 0, m, V)$-model with strong convexity of the function $\psi_\delta(x,y)$ is proposed in [65] to address a *non-convex* optimization problem which arises in an electoral model for clustering introduced in [57]. In this model, voters (data points) select a party (cluster) iteratively by minimizing the following function

$$\min_{z\in S_n(1), p\in\mathbb{R}_+^m}\left\{ f_{\mu_1,\mu_2}(x = (z,p)) = g(x) + \mu_1\sum_{k=1}^n z_k \ln z_k + \frac{\mu_2}{2}\|p\|_2^2\right\}.$$

Let us choose $\|x\|^2 = \|z\|_1^2 + \|p\|_2^2$ and assume that, in general non-convex, $g(x)$ has L_g–Lipschitz continuous gradient

$$\|\nabla g(x) - \nabla g(y)\|_* \leq L_g \|x - y\| \quad \forall x, y \in S_n(1) \times \mathbb{R}_+^m$$

and $L_g \leq \mu_1$ and $L_g \leq \mu_2$. It can be shown (see [65]) that

$$\psi_\delta(x, y) = \langle \nabla g(y), x - y\rangle - L_g \cdot \mathrm{KL}(z_x|z_y) - \frac{L_g}{2}\|p_x - p_y\|_2^2$$
$$+\mu_1(\mathrm{KL}(z_x|\mathbf{1}) - \mathrm{KL}(z_y|\mathbf{1})) + \frac{\mu_2}{2}\left(\|p_x\|_2^2 - \|p_y\|_2^2\right)$$

is a $(0, 2L_g, 0, \min\{\mu_1, \mu_2\} - L_g, V)$-model of $f_{\mu_1,\mu_2}(x)$. Here $\mathrm{KL}(z_x|z_y) = \sum_{i=1}^m [z_x]_i \ln([z_x]_i/[z_y]_i)$ and

$$V[y](x) = \mathrm{KL}(z_x|z_y) + \frac{1}{2}\|p_x - p_y\|_2^2.$$

We finish this section by defining an approximate solution to an optimization problem. This definition will be used to allow inexact solutions of auxiliary minimization problems on each iteration of our algorithms.

Definition 11. For a convex optimization problem $\min_{x\in Q} \Psi(x)$, we denote by Arg $\min^{\widetilde{\delta}}_{x\in Q} \Psi(x)$ a set of such $\widetilde{x}$ that

$$\exists h \in \partial\Psi(\widetilde{x})\colon \forall x \in Q \to \langle h, x - \widetilde{x}\rangle \geq -\widetilde{\delta}.$$

We denote by $\mathrm{argmin}^{\widetilde{\delta}}_{x\in Q} \Psi(x)$ some element of Arg $\min^{\widetilde{\delta}}_{x\in Q} \Psi(x)$.

3 Gradient Method with Inexact Model.

In this section we consider adaptive gradient-type methods for problems with (δ, L, μ, m, V)-model of the objective. First, we consider non accelerated gradient method and then an accelerated version. We note that non-accelerated Algorithm 1 is suitable for the problems with relative smoothness and relative strong convexity, also there is no accumulation of errors. Accelerated Algorithm 2 gives a better estimate with errors close to zero, however, accumulation of errors is possible. We consider Algorithm 2 for the narrower class of problems with $(\delta, L, \mu, m, V, \|\cdot\|)$-models (see Definition 13) w.r.t norm $\|\cdot\|$. It means, that non-accelerated method (Algorithm 1) is suitable for a wider class of problems.

3.1 Adaptive Gradient Method with (δ, L, μ, m, V)-Model

In this section we consider adaptive gradient method for problem (1), which uses a (δ, L, μ, m, V)-model of the objective. For the case when $\mu + m > 0$ our method has linear convergence and for a more general case $\mu = 0$ and $m = 0$, we prove a sublinear convergence rate.

We assume that in each iteration k, the method has access to $(\delta, \bar{L}_{k+1}, \mu, m, V)$-model of f w.r.t $V[y](x)$ (see Definition 1). In general, constant $\bar{L}_{k+1}$ may vary from iteration to iteration and we only assume that the $(\delta, \bar{L}_{k+1}, \mu, m, V)$-model exists. We do not use $\bar{L}_{k+1}$ in Algorithm 1 explicitly and, moreover, our method is adaptive to this constant.

Algorithm 1 Adaptive gradient method with (δ, L, μ, m, V)-model

Input: x_0 is the starting point, $\mu \geq 0$ and δ.
Set $S_0 := 0$
for $k \geq 0$ **do**
    Find the smallest integer $i_k \geq 0$ such that

$$f_\delta(x_{k+1}) \leq f_\delta(x_k) + \psi_\delta(x_{k+1}, x_k) + L_{k+1} V[x_k](x_{k+1}) + \delta, \tag{6}$$

    where $L_{k+1} = 2^{i_k - 1} L_k$ for $L_k > 2\mu$ and $L_{k+1} = 2^{i_k} L_k$ for $L_k \leq 2\mu$,
    $\alpha_{k+1} := \frac{1}{L_{k+1}}$, $S_{k+1} := S_k + \alpha_{k+1}$.

$$\phi_{k+1}(x) := \psi_\delta(x, x_k) + L_{k+1} V[x_k](x), \quad x_{k+1} := \arg\min_{x \in Q}{}^{\widetilde{\delta}} \phi_{k+1}(x). \tag{7}$$

end for

We consider the case of m-strong convexity of the function $\psi_\delta(x,y)$ and prove convergence rate theorem for Algorithm 1, in particular, we prove a linear convergence for $\mu > 0$ or $m > 0$.

For $L_k > \mu$ and all $k \geq 0$, we denote

$$q_k \overset{\text{def}}{=} \frac{L_k - \mu}{L_k + m}, \quad Q_j^k \overset{\text{def}}{=} \prod_{i=j}^{k} q_i.$$

We assume that $Q_j^k = 1$ for $j > k$.

Theorem 12. *Assume that $\psi_\delta(x,y)$ is a (δ, L, μ, m, V)-model according to Definition 1. Denote by $y_N = \operatorname{argmin}_{k=1,\ldots,N} f(x_k)$. Then, after N iterations of Algorithm 1 we have*

$$f(y_N) - f(x_*) \leq \min\left\{(L_N + m) Q_1^N, \frac{1}{\sum_{i=1}^{N} \frac{1}{L_i + m}}\right\} V[x_0](x_*) + \widetilde{\delta} + 2\delta, \tag{8}$$

$$V[x_N](x_*) \leq Q_1^N V[x_0](x_*) + (\widetilde{\delta} + 2\delta) \sum_{i=1}^{N} \frac{Q_{i+1}^N}{L_i + m}. \tag{9}$$

To prove Theorem 12 we need the following lemma.

Lemma 1. *Let $\psi(x)$ be a m-strongly convex function, $m \geq 0$, and*

$$y = \operatorname*{argmin}_{x \in Q}{}^{\widetilde{\delta}} \{\psi(x) + \beta V[z](x)\},$$

where $\beta \geq 0$. Then

$$\psi(x) + \beta V[z](x) \geq \psi(y) + \beta V[z](y) + (\beta + m) V[y](x) - \widetilde{\delta}, \;\; \forall x \in Q.$$

Proof. By Definition 11:

$$\exists g \in \partial\psi(y), \;\; \langle g + \beta \nabla_y V[z](y), x - y \rangle \geq -\widetilde{\delta}, \;\; \forall x \in Q.$$

Then inequality

$$\psi(x) - \psi(y) \geq \langle g, x - y\rangle + mV[y](x) \geq \langle \beta \nabla_y V[z](y), y - x\rangle - \widetilde{\delta} + mV[y](x)$$

and equality

$$\begin{aligned}\langle \nabla_y V[z](y), y - x\rangle = \langle \nabla d(y) - \nabla d(z), y - x\rangle = d(y) - d(z) - \langle \nabla d(z), y - z\rangle + \\ + d(x) - d(y) - \langle \nabla d(y), x - y\rangle - d(x) + d(z) + \langle \nabla d(z), x - z\rangle = \\ = V[z](y) + V[y](x) - V[z](x)\end{aligned}$$

complete the proof. □

Proof of Theorem 12. Since by Definition 1 with $x = y$, $f(x) - \delta \leqslant f_\delta(x) \leqslant f(x)$, and (6), we have the following series of inequalities

$$f(x_N) \leq f_\delta(x_N) + \delta \leq f_\delta(x_{N-1}) + \psi_\delta(x_N, x_{N-1}) + L_N V[x_{N-1}](x_N) + 2\delta.$$

Using Lemma 1 for (7) we have

$$f(x_N) \leq f_\delta(x_{N-1}) + \psi_\delta(x, x_{N-1}) + L_N V[x_{N-1}](x) - (L_N + m)\, V[x_N](x) + \widetilde{\delta} + 2\delta.$$

In view of the left inequality (4), we have

$$f(x_N) \leq f(x) + (L_N - \mu)V[x_{N-1}](x) - (L_N + m)\, V[x_N](x) + \widetilde{\delta} + 2\delta. \tag{10}$$

Taking $x = x_*$ and using inequality $f(x_*) \leq f(x_N)$, we obtain

$$(L_N + m)\, V[x_N](x_*) \leq (L_N - \mu)V[x_{N-1}](x_*) + \widetilde{\delta} + 2\delta.$$

Thus, we have that

$$V[x_N](x_*) \leq q_N V[x_{N-1}](x_*) + \frac{\widetilde{\delta} + 2\delta}{L_N + m} \leq Q_1^N V[x_0](x_*) + (\widetilde{\delta} + 2\delta) \sum_{i=1}^{N} \frac{Q_{i+1}^N}{L_i + m}.$$

The last inequality proves (9). Now we rewrite (10) for $x = x_*$ as

$$V[x_N](x_*) \leq \frac{1}{L_N + m}(f(x_*) - f(x_N) + \widetilde{\delta} + 2\delta) + q_N V[x_{N-1}](x_*).$$

Recursively, we have

$$V[x_N](x_*) \leq \sum_{i=1}^{N} \left(\frac{Q_{i+1}^N}{L_i + m}(f(x_*) - f(x_i) + \widetilde{\delta} + 2\delta) \right) + Q_1^N V[x_0](x_*).$$

Using that $V[x_N](x_*) \geq 0$ and the definition of y_N, we get

$$\begin{aligned} Q_1^N V[x_0](x_*) &\geq \sum_{i=1}^{N} \left(\frac{Q_{i+1}^N}{L_i + m}(f(x_i) - f(x_*) - \widetilde{\delta} - 2\delta) \right) \\ &\geq (f(y_N) - f(x_*)) \sum_{i=1}^{N} \frac{Q_{i+1}^N}{L_i + m} - (\widetilde{\delta} + 2\delta) \sum_{i=1}^{N} \frac{Q_{i+1}^N}{L_i + m}. \end{aligned}$$

Dividing by $\sum_{i=1}^{N} \frac{Q_{i+1}^N}{L_i+m}$, we obtain

$$f(y_N) - f(x_*) \leq \frac{Q_1^N}{\sum_{i=1}^{N} \frac{Q_{i+1}^N}{L_i+m}} V[x_0](x_*) + \widetilde{\delta} + 2\delta.$$

Since $\sum_{i=1}^{N} \frac{Q_{i+1}^N}{L_i+m} \geq \frac{1}{L_N+m}$ and $Q_i^N \geq Q_1^N$ for all $i \geq 1$, we get

$$f(y_N) - f(x_*) \leq \min\left\{ (L_N+m)Q_1^N, \frac{1}{\sum_{i=1}^{N} \frac{1}{L_i+m}} \right\} V[x_0](x_*) + \widetilde{\delta} + 2\delta.$$

This proves (8). □

Remark 1. Let us assume that $L_0 \leq L$, and we know that $\bar{L}_{k+1} \leq L$ for all $k \geq 0$ (or in other words, (δ, L, μ, m, V)-model exists for all $k \geq 0$). This means that $L_k \leq 2L$ for all $k \geq 0$ due to (δ, L, μ, m, V)–model definition and L_k selection rule. From this fact we can obtain that $\sum_{i=1}^{N} \frac{1}{L_i+m} \geq \frac{N}{2L+m}$ and $q_k \leq q \overset{\text{def}}{=} \frac{2L-\mu}{2L+m}$. In view of the last two inequalities, we have

$$f(y_N) - f(x_*) \leq \min\left\{ \frac{2L+m}{N}, (2L+m)q^N \right\} V[x_0](x_*) + \widetilde{\delta} + 2\delta,$$

$$V[x_N](x_*) \leq q^N V[x_0](x_*) + (\widetilde{\delta} + 2\delta) \sum_{i=1}^{N} \frac{Q_{i+1}^N}{L_i+m}.$$

Remark 2. The advantage of Algorithm 1 is that there is no need to know the true values of the parameters L and m. Using the standard argument [20] one can show that the number of oracle calls is less than $2N + \log_2 \frac{2L}{L_0}$, where N is the number of iterations of Algorithm 1.

3.2 Adaptive Fast Gradient Method with $(\delta, L, \mu, m, V, \|\cdot\|)$-model

In this section we consider accelerated method for problems with $(\delta, L, \mu, m, V, \|\cdot\|)$-model of the objective (see Definition 13). The method is close to accelerated mirror-descent type of methods (see [22, 41, 66]). In contrast to the previous section, in this section we make a stronger assumption on the model, which is required to obtain acceleration of the gradient method. Namely, we use the square of the norm in the r.h.s. of (4) instead of the function V, which gives the following modification of Definition 1.

Definition 13. Let $\delta, L, \mu, m \geq 0$. We say that $\psi_\delta(x,y)$ is a $(\delta, L, \mu, m, V, \|\cdot\|)$-model of the function f at a given point y iff, for all $x \in Q$,

$$\mu V[y](x) \leqslant f(x) - (f_\delta(y) + \psi_\delta(x,y)) \leqslant \frac{L}{2}\|x-y\|^2 + \delta.$$

and $\psi_\delta(x,y)$ satisfies $\psi_\delta(x,x) = 0$ for all $x \in Q$ and

$$\psi(x) \geqslant \psi(z) + \langle \nabla_z \psi(z), x - z \rangle + mV[z](x) \quad \forall x, z \in Q,$$

where for fixed $y \in Q$ and any $x \in Q$ we denote $\psi(x) = \psi_\delta(x,y)$.

As in the previous subsection we assume that there exists some constant $\bar{L}_{k+1}$ such that $(\delta_k, \bar{L}_{k+1}, \mu, m, V, \|\cdot\|)$-model of f exists at k-th step ($k = 0, .., N-1$) of Algorithm 2. Unlike Algorithm 1, we assume that the errors $\widetilde{\delta}, \delta$ can depend on the iteration counter k, which is indicated by input sequences $\{\widetilde{\delta}_k\}_{k\geq 0}$ and $\{\delta_k\}_{k\geq 0}$. For instance, this allows to obtain Universal Fast Gradient Method in which different values of $\{\delta_k\}_{k\geq 0}$ are required (see [4, 53]).

Theorem 14. *Assume that $\psi_\delta(x, y)$ is a $(\delta, L, \mu, m, V, \|\cdot\|)$-model according to Definition 13. Also assume that $V[y](x)$ satisfies (1-SC) condition w.r.t. $\|\cdot\|$-norm. Then, after N iterations of Algorithm 2 we have*

$$f(x_N) - f(x_*) \leq \frac{V[u_0](x_*)}{A_N} + \frac{2\sum_{k=0}^{N-1} A_{k+1}\delta_k}{A_N} + \frac{\sum_{k=0}^{N-1}\widetilde{\delta}_k}{A_N}, \tag{11}$$

$$V[u_N](x_*) \leq \frac{V[u_0](x_*)}{(1 + A_N\mu + A_N m)} + \frac{2\sum_{k=0}^{N-1} A_{k+1}\delta_k}{(1 + A_N\mu + A_N m)} + \frac{\sum_{k=0}^{N-1}\widetilde{\delta}_k}{(1 + A_N\mu + A_N m)}. \tag{12}$$

Remark 3. Despite the adaptive structure of Algorithm 2 as in [53] it can be shown that in average the algorithm up to logarithmic terms requires four computations of function and two computations of $(\delta, L, \mu, m, V, \|\cdot\|)$-model per iteration.

Algorithm 2 Fast adaptive gradient method with $(\delta, L, \mu, m, V, \|\cdot\|)$-model

1: **Input:** x_0 is the starting point, $\mu \geq 0$, $m \geq 0$, $\{\widetilde{\delta}_k\}_{k\geq 0}$, $\{\delta_k\}_{k\geq 0}$ and $L_0 > 0$.
2: Set $y_0 := x_0$, $u_0 := x_0$, $\alpha_0 := 0$, $A_0 := \alpha_0$
3: **for** $k \geq 0$ **do**
4: Find the smallest integer $i_k \geq 0$ such that

$$f_{\delta_k}(x_{k+1}) \leq f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_{k+1}, y_{k+1}) + \frac{L_{k+1}}{2}\|x_{k+1} - y_{k+1}\|^2 + \delta_k, \tag{13}$$

where $L_{k+1} = 2^{i_k - 1}L_k$, α_{k+1} is the largest root of

$$A_{k+1}(1 + A_k\mu + A_k m) = L_{k+1}\alpha_{k+1}^2, \quad A_{k+1} := A_k + \alpha_{k+1}. \tag{14}$$

$$y_{k+1} := \frac{\alpha_{k+1}u_k + A_k x_k}{A_{k+1}}. \tag{15}$$

$$\phi_{k+1}(x) = \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}) + (1 + A_k\mu + A_k m)V[u_k](x) + \alpha_{k+1}\mu V[y_{k+1}](x).$$

$$u_{k+1} := \operatorname*{argmin}_{x\in Q}{}^{\widetilde{\delta}_k}\phi_{k+1}(x). \tag{16}$$

$$x_{k+1} := \frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}. \tag{17}$$

5: **end for**

In order to prove Theorem 14 we need the following lemma.

Lemma 2. *Let $\psi(x)$ be a m-strongly convex function, $m \geq 0$, and*

$$y = \operatorname*{argmin}_{x\in Q}{}^{\widetilde{\delta}}\{\psi(x) + \beta V[z](x) + \gamma V[u](x)\},$$

where $\beta \geq 0$ and $\gamma \geq 0$. Then

$$\psi(x) + \beta V[z](x) + \gamma V[u](x) \geq \psi(y) + \beta V[z](y) + \gamma V[u](y) + (\beta + \gamma + m)V[y](x) - \widetilde{\delta}, \quad \forall x \in Q.$$

We omit the proof of Lemma 2 since it is similar to the proof of Lemma 1.

Lemma 3. *For all $x \in Q$, we have*

$$A_{k+1}f(x_{k+1}) - A_k f(x_k) + (1 + A_{k+1}\mu + A_{k+1}m)V[u_{k+1}](x) - (1 + A_k\mu + A_k m)V[u_k](x) \\ \leq \alpha_{k+1}f(x) + 2\delta_k A_{k+1} + \widetilde{\delta}_k.$$

Proof. Since by Definition 13 with $x = y$, $f(x) - \delta \leqslant f_\delta(x) \leqslant f(x)$, and (13), we have

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

Using definitions (17) and (15) of sequences x_{k+1} and y_{k+1} we can show that

$$\begin{aligned} f(x_{k+1}) \leq\ & f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}\left(\frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}, y_{k+1}\right) \\ & + \frac{L_{k+1}}{2}\left\|\frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}} - y_{k+1}\right\|^2 + 2\delta_k \\ =\ & f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}\left(\frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}, y_{k+1}\right) + \frac{L_{k+1}\alpha_{k+1}^2}{2A_{k+1}^2}\|u_{k+1} - u_k\|^2 + 2\delta_k. \end{aligned}$$

Since $\psi_{\delta_k}(\cdot, y)$ is convex, we have

$$\begin{aligned} f(x_{k+1}) \leq\ & \frac{A_k}{A_{k+1}}\left(f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_k, y_{k+1})\right) + \frac{\alpha_{k+1}}{A_{k+1}}\left(f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(u_{k+1}, y_{k+1})\right) \\ & + \frac{L_{k+1}\alpha_{k+1}^2}{2A_{k+1}^2}\|u_{k+1} - u_k\|^2 + 2\delta_k. \end{aligned}$$

In view of definition (14) for the sequence α_{k+1}, we obtain

$$\begin{aligned} f(x_{k+1}) \leq\ & \frac{A_k}{A_{k+1}}\left(f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_k, y_{k+1})\right) + \frac{\alpha_{k+1}}{A_{k+1}}\Big(f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(u_{k+1}, y_{k+1}) \\ & + \frac{1 + A_k\mu + A_k m}{2\alpha_{k+1}}\|u_{k+1} - u_k\|^2\Big) + 2\delta_k. \end{aligned}$$

Using (1-SC) condition w.r.t. norm for V and the left inequality in (4), we get

$$\begin{aligned} f(x_{k+1}) \leq\ & \frac{A_k}{A_{k+1}}f_{\delta_k}(x_k) + \frac{\alpha_{k+1}}{A_{k+1}}\Big(f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(u_{k+1}, y_{k+1}) \\ & + \frac{1 + A_k\mu + A_k m}{\alpha_{k+1}}V[u_k](u_{k+1})\Big) + 2\delta_k. \end{aligned} \tag{18}$$

By Lemma 2 for the optimization problem in (16), it holds that

$$\begin{aligned} & \alpha_{k+1}\psi_{\delta_k}(u_{k+1}, y_{k+1}) + (1 + A_k\mu + A_k m)V[u_k](u_{k+1}) + \alpha_{k+1}\mu V[y_{k+1}](u_{k+1}) \\ & \quad + (1 + A_{k+1}\mu + A_{k+1}m)V[u_{k+1}](x) - \widetilde{\delta}_k \\ & \leq \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}) + (1 + A_k\mu + A_k m)V[u_k](x) + \alpha_{k+1}\mu V[y_{k+1}](x). \end{aligned}$$

From the fact that $V[y_{k+1}](u_{k+1}) \geq 0$, we have

$$\begin{aligned} & \alpha_{k+1}\psi_{\delta_k}(u_{k+1}, y_{k+1}) + (1 + A_k\mu + A_k m)V[u_k](u_{k+1}) \\ & \leq \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}) + (1 + A_k\mu + A_k m)V[u_k](x) \\ & \quad - (1 + A_{k+1}\mu + A_{k+1}m)V[u_{k+1}](x) + \alpha_{k+1}\mu V[y_{k+1}](x) + \widetilde{\delta}_k. \end{aligned} \tag{19}$$

Combining (18) and (19), we obtain

$$f(x_{k+1}) \le \frac{A_k}{A_{k+1}} f(x_k) + \frac{\alpha_{k+1}}{A_{k+1}} \Big(f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x, y_{k+1}) + \mu V[y_{k+1}](x) + \frac{1 + A_k\mu + A_k m}{\alpha_{k+1}} V[u_k](x) - \frac{1 + A_{k+1}\mu + A_{k+1} m}{\alpha_{k+1}} V[u_{k+1}](x) + \frac{\widetilde{\delta}_k}{\alpha_{k+1}} \Big) + 2\delta_k.$$

We finish the proof of Lemma 3 applying left inequality in (4)

$$f(x_{k+1}) \le \frac{A_k}{A_{k+1}} f(x_k) + \frac{\alpha_{k+1}}{A_{k+1}} f(x) + \frac{1 + A_k\mu + A_k m}{A_{k+1}} V[u_k](x) - \frac{1 + A_{k+1}\mu + A_{k+1} m}{A_{k+1}} V[u_{k+1}](x) + 2\delta_k + \frac{\widetilde{\delta}_k}{A_{k+1}}.$$

□

Proof of Theorem 14. We telescope the inequality in Lemma 3 for k from 0 to $N-1$ and take $x = x_*$:

$$A_N f(x_N) \le A_N f(x_*) + V[u_0](x_*) - (1 + A_N(\mu + m)) V[u_N](x_*) + 2\sum_{k=0}^{N-1} A_{k+1}\delta_k + \sum_{k=0}^{N-1} \widetilde{\delta}_k. \tag{20}$$

Since $V[u_{k+1}](x_*) \ge 0$ for all $k \ge 0$, we have

$$A_N f(x_N) - A_N f(x) \le V[u_0](x_*) + 2\sum_{k=0}^{N-1} A_{k+1}\delta_k + \sum_{k=0}^{N-1} \widetilde{\delta}_k.$$

The last inequality proves (11). Inequality (12) is straightforward from (20) since $f(x) \ge f(x_*)$ for all $x \in Q$. □

Next lemma is proved in Appendix B and gives the growth rate for A_N, see [14, 32, 52].

Lemma 4. *For all $N \ge 0$,*

$$A_N \ge \max\left\{ \frac{1}{2}\left(\sum_{k=0}^{N-1} \frac{1}{\sqrt{L_{k+1}}}\right)^2, \frac{1}{L_1} \prod_{k=1}^{N-1} \left(1 + \sqrt{\frac{\mu + m}{2L_{k+1}}}\right)^2 \right\}.$$

Remark 4. Let us assume that function f has L-Lipschitz continuous gradient. This means that for $L_k \ge L$ inequality (13) always holds, whence, $L_k \le 2L$, assuming that $L_0 \le L$. From Lemma 4 we have $A_N \ge \frac{N^2}{4L}$ and

$$A_N \ge \frac{1}{2L}\left(1 + \frac{1}{2}\sqrt{\frac{\mu + m}{L}}\right)^{2(N-1)} \ge \frac{1}{2L}\exp\left(\frac{N-1}{2}\sqrt{\frac{\mu + m}{L}}\right).$$

In the last inequality we used inequality $\log(1 + 2x) \ge x$ for all $x \in [0, \frac{1}{4}]$. Combining Theorem 14 and Lemma 4 we have

$$f(x_N) - f(x_*) \le \min\left\{ \frac{4L}{N^2}, 2L\exp\left(-\frac{N-1}{2}\sqrt{\frac{\mu + m}{L}}\right) \right\} V[u_0](x_*) + \frac{2\sum_{k=0}^{N-1} A_{k+1}\delta_k}{A_N} + \frac{\sum_{k=0}^{N-1}\widetilde{\delta}_k}{A_N}. \tag{21}$$

The first term in the right hand side of ineqaulity (21) up to a constant factor is optimal for μ–strongly convex functions with L-Lipschitz continuous gradient.

Note that in [14] for non-adaptive fast gradient method with (δ, L, μ)–oracle for the case when δ_k is a constant it is shown that

$$\frac{\sum_{k=0}^{N-1} A_{k+1}\delta}{A_N} \leq \min\left\{\left(\frac{1}{3}k + 2.4\right), \left(1 + \sqrt{\frac{L}{\mu}}\right)\right\}\delta.$$

This means that for $\mu > 0$ error δ does not accumulate.

Remark 5. In view of assumptions from Remark 4. For the case when $\mu = m = 0$ Algorithm 2 can guarantee the following convergence rate

$$f(x_N) - f_* \leq \frac{4LV[x_0](x_*)}{N^2} + 2N\delta + \frac{4L\widetilde{\delta}}{N}.$$

A similar result was shown in [31].

Remark 6. Let us analyze the convergence rate of the argument (12) from Theorem 14. There are two different scenarios:

1 $\mu = m = 0$. In this case we have:

$$V[u_N](x_*) \leq V[u_0](x_*) + 2\sum_{k=0}^{N-1} A_{k+1}\delta_k + \sum_{k=0}^{N-1} \widetilde{\delta}_k.$$

For non-strongly convex case we can only bound $V[u_N](x_*)$ by $V[u_0](x_*)$ up to additive noise.

2 $\mu + m > 0$. Using Lemma 4 we can see that Theorem 14 guarantees linear convergence in argument up to additive noise.

Note that convergence rates for the objective and for the argument are obtained for different sequences x_N and u_N, respectively.

3.3 Universal conditional gradient (Frank–Wolfe) method

Let us show an example of $(\delta, L, \mu, m, V, \|\cdot\|)$–model application. We use Algorithm 2 as a proxy method for universal Frank–Wolfe method with $\mu = 0$ and $m = 0$. In order to construct universal Frank–Wolfe method let us introduce the following constraints to the optimization problem (1):

1 The set Q is bounded w.r.t $V[y](x)$: $\exists R_Q \in \mathbb{R}: \ V[y](x) \leq R_Q^2 \quad \forall x, y \in Q.$

2 The function $f(x)$ has Holder-continuous subgradients:

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

From this we can get an inequality (see [53]):

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

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 free parameter.

First, let us take $\delta_k = \epsilon\frac{\alpha_{k+1}}{4A_{k+1}}$. With this choice of δ_k and the fact that the objective function has Holder continues subgradient as in Theorem 3 from [53] we can get the following inequality for A_N:

$$A_N \geq \frac{N^{\frac{1+3\nu}{1+\nu}}\epsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{3+5\nu}{1+\nu}}L_\nu^{\frac{2}{1+\nu}}}. \tag{22}$$

It it shown in [31] that in order to construct the classical Frank–Wolfe method instead of an auxiliary problem $\phi_{k+1}(x) = \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}) + V[u_k](x)$ in Algorithm 2 for $m = 0$ and $\mu = 0$ (see also section 3, [31]) we can take an auxiliary problem $\widetilde{\phi}_{k+1}(x) = \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1})$. Let us look at this substitution from the view of $\widetilde{\delta}_k$–precision from Definition 11. As in [31] we can show that an error in sense of Definition 11 would not be greater than $2R_Q^2$. Therefore, we can take $\widetilde{\delta}_k = 2R_Q^2$. From Theorem 14 we can get the following inequality:

$$f(x_N) - f(x_*) \leq \frac{R_Q^2}{A_N} + \frac{\varepsilon}{2} + \frac{2R_Q^2 N}{A_N} \leq \frac{3R_Q^2 N}{A_N} + \frac{\varepsilon}{2}.$$

Using inequality (22), we can finally get the following upper bound for the number of steps in order to get ε-solution:

$$N \leq \inf_{\nu\in(0,1]}\left[2^{\frac{3+4\nu}{\nu}}\left(\frac{L_\nu R_Q^{1+\nu}}{\varepsilon}\right)^{\frac{1}{\nu}}\right].$$

This inequality for $\nu = 1$ has the same convergence rate as in the classical Frank–Wolfe method, however, universal Frank–Wolfe method can work with any function that has Holder continuous subgradients with constant $\nu > 0$. Note that in the classical Frank–Wolfe method $\psi_{\delta_k}(x, y_{k+1}) = \langle\nabla f(y_{k+1}), x - y_{k+1}\rangle$. However, here we assume that $\psi_{\delta_k}(x, y_{k+1})$ can have a more general representation (see Definition 13).

4 Inexact Model for Variational Inequalities

In this section, we go beyond minimization problems and propose an abstract inexact model counterpart for variational inequalities. As a special case in Example 17 we introduce relative smoothness for operators in the spirit of [45], where it was introduced for optimization problems. Further, we propose a generalization of the Mirror-Prox algorithm for this general case of inexact model of the operator and abstract variational inequalities.One of the main features of our algorithm is its adaptation to generalized inexact parameter of smoothness. As a special case we propose a universal method for variational inequalities with complexity $O\left(\inf_{\nu\in[0,1]}\left(\frac{1}{\varepsilon}\right)^{\frac{2}{1+\nu}}\right)$, where ε is the desired accuracy of the solution and ν is the Hölder exponent of the operator. According to the lower bounds in [62], this algorithm is optimal for $\nu = 0$ (bounded variation of the operator) and $\nu = 1$ (Lipschitz continuity of the operator). Based on the model for VI and functions, we introduce inexact model for saddle-point problems (see Definition 20). We are also motivated by mixed variational inequalities [5, 38] and composite saddle-point problems [10].

Formally speaking, we consider the problem of finding the solution $x_* \in Q$ for VI in the following abstract form

$$\psi(x, x_*) \geqslant 0 \quad \forall x \in Q \tag{23}$$

for some convex compact set $Q \subset \mathbb{R}^n$ and some function $\psi : Q \times Q \to \mathbb{R}$. Assuming the abstract monotonicity of the function ψ

$$\psi(x, y) + \psi(y, x) \leqslant 0 \quad \forall x, y \in Q, \tag{24}$$

any solution to (23) is a solution of the following inequality

$$\max_{x \in Q} \psi(x_*, x) \leqslant 0 \tag{25}$$

In the general case, we make an assumption about the existence of a solution x_* of the problem (23). As a particular case, if for some operator $g : Q \to \mathbb{R}^n$ we set $\psi(x, y) = \langle g(y), x - y\rangle \;\; \forall x, y \in Q$, then (23) and (25) are equivalent, respectively, to a standard strong and weak variational inequality with the operator g.

We propose an adaptive proximal method for the problems (23) and (25). We start with a concept of (δ, L, V)-model for such problems.

Definition 15. We say that function ψ has (δ, L, V)-model $\psi_\delta(x, y)$ for some fixed values $\delta > 0$ and $L = L(\delta) > 0$ if the following properties hold for each $x, y, z \in Q$:

(i) $\psi(x, y) \leq \psi_\delta(x, y) + \delta$;

(ii) $\psi_\delta(x, y)$ convex in the first variable;

(iii) $\psi_\delta(x, x) = 0$;

(iv) (*abstract δ-monotonicity*)

$$\psi_\delta(x, y) + \psi_\delta(y, x) \leq \delta; \tag{26}$$

(v) (*generalized relative smoothness*)

$$\psi_\delta(x, y) \leqslant \psi_\delta(x, z) + \psi_\delta(z, y) + LV[z](x) + LV[y](z) + \delta. \tag{27}$$

Example 16. For some operator $g : Q \to \mathbb{R}^n$ and a convex function $h : Q \to \mathbb{R}^n$ choice

$$\psi(x, y) = \langle g(y), x - y\rangle + h(x) - h(y)$$

leads to a *mixed variational inequality from* [5, 38]

$$\langle g(y), y - x\rangle + h(y) - h(x) \leqslant 0,$$

which in the case of the monotonicity of the operator g implies

$$\langle g(x), y - x\rangle + h(y) - h(x) \leqslant 0.$$

Remark 7. Similarly to Definition 1 above, in general case, we do not need the (1-SC) assumption for $V[y](x)$ in Definition 15. In some situations we make (1-SC) assumption for $V[y](x)$ (see Example 18 and Section 5).

Note that for $\delta = 0$ the following analogue of (27) for some fixed $a, b > 0$

$$\psi(x,y) \leqslant \psi(x,z) + \psi(z,y) + a\|z-y\|^2 + b\|x-z\|^2 \quad \forall x,y,z \in Q \tag{28}$$

was introduced in [47]. Condition (28) is used in many works on equilibrium programming. Our approach allows us to work with non-Euclidean set-up without (1-SC) assumption and inexactness δ, that is important for the ideology of universal methods [53] (see Example 18 below).

One can directly verify that if $\psi_\delta(x,y)$ is $(\delta/3, L, 0, 0, V)$-model of the function f at a given point y then $\psi_\delta(x,y)$ is (δ, L, V)-model in the sense of Definition 15.

Let us consider some examples.

Example 17. Relative smoothness for optimization and VI. Let us consider a minimization problem (1) with the function f being convex and relatively L-smooth w.r.t. to d [45], i.e., for all $x, y \in Q$,

$$f(x) - f(y) - \langle \nabla f(y), x-y\rangle \leq LV[y](x).$$

In this case, (1) is equivalent to abstract VI (25) with $\psi_\delta(x,y) := \langle \nabla f(y), x-y\rangle$. Properties (i)-(iv) in Definition 15 obviously hold with $\delta = 0$. Let us check that (v) also holds. Indeed,

$$\begin{aligned}
&\psi_\delta(x,y) - \psi_\delta(x,z) - \psi_\delta(z,y) = \langle \nabla f(y), x-y\rangle - \langle \nabla f(z), x-z\rangle - \langle \nabla f(y), z-y\rangle \\
&\quad = (f(x) - f(z) - \langle \nabla f(z), x-z\rangle) + (f(z) - f(y) - \langle \nabla f(y), z-y\rangle) \\
&\quad - (f(x) - f(y) - \langle \nabla f(y), x-y\rangle) \leq LV[z](x) + LV[y](z),
\end{aligned}$$

where we used relative L-smoothness and convexity of f. This example shows that our inexact model for VI as a particular case contains the concept of relative smoothness introduced in optimization. In this particular case we say that an operator g is relatively L-smooth if

$$\langle g(y) - g(z), x-z\rangle \leq LV[z](x) + LV[y](z), \ \ \forall x,y,z \in Q.$$

Example 18. Variational Inequalities with monotone Hölder continuous operator. Assume that V satisfies (1-SC) condition w.r.t. some norm $\|\cdot\|$ and for a monotone operator g there exists $\nu \in [0,1]$ such that

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

Then we have: $\langle g(z) - g(y), z-x\rangle \leq \|g(z) - g(y)\|_* \|z-x\| \leq L_\nu \|z-y\|^\nu \|z-x\|$

$$\leq \frac{L(\delta)}{2}||z-x||^2 + \frac{L(\delta)}{2}||z-y||^2 + \delta \leq LV[z](x) + LV[y](z) + \delta \tag{29}$$

for

$$L(\delta) = \left(\frac{1}{2\delta}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}} \tag{30}$$

with arbitrary $\delta > 0$. In this case $\psi_\delta(x,y) := \langle g(y), x-y\rangle$ is (δ, L, V)-model.

Note that for the previous two examples in Algorithm 3 and Theorem 19 we need $V[z](x)$ to satisfy (1-SC) condition.

Next, we introduce our novel adaptive method (Algorithm 3) for abstract variational inequalities with inexact (δ, L, V)-model. This method adapts to the local values of L and allows us to construct universal method for variational inequalities by applying it to VI with Hölder interpolation (29) for $\delta = \frac{\varepsilon}{2}$ and $L = L\left(\frac{\varepsilon}{2}\right)$.

Algorithm 3 Generalized Mirror Prox for VI

Require: accuracy $\varepsilon > 0$, oracle error $\delta > 0$, initial guess $L_0 > 0$, prox set-up: $d(x)$, $V[z](x)$.

1: Set $k = 0$, $z_0 = \arg\min_{u \in Q} d(u)$.

2: **repeat**

3: Find the smallest integer $i_k \geq 0$ such that

$$\psi_\delta(z_{k+1}, z_k) \leq \psi_\delta(z_{k+1}, w_k) + \psi_\delta(w_k, z_k) + L_{k+1}(V[z_k](w_k) + V[w_k](z_{k+1})) + \delta, \tag{31}$$

where $L_{k+1} = 2^{i_k - 1} L_k$ and

$$w_k = \underset{x \in Q}{\operatorname{argmin}} \left\{ \psi_\delta(x, z_k) + L_{k+1} V[z_k](x) \right\}. \tag{32}$$

$$z_{k+1} = \underset{x \in Q}{\operatorname{argmin}} \left\{ \psi_\delta(x, w_k) + L_{k+1} V[z_k](x) \right\}. \tag{33}$$

4: **until** $S_N := \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \geqslant \frac{\max\limits_{x \in Q} V[x^0](x)}{\varepsilon}$.

Ensure: $\widehat{w}_N = \frac{1}{\sum_{k=0}^{N-1} \frac{1}{L_{k+1}}} \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} w_k$.

Next we state convergence rate result for the proposed method.

Theorem 19. *For Algorithm 3 the following inequality holds*

$$-\frac{1}{S_N} \sum_{k=0}^{N-1} \frac{\psi_\delta(x, w_k)}{L_{k+1}} \leq \frac{V[z_0](x)}{S_N} + \delta + 2\widetilde{\delta} \quad \forall x \in Q.$$

It means that:

$$\max_{u \in Q} \psi(\widehat{w}_N, u) \leq \frac{2L \max_{u \in Q} V[z_0](u)}{N} + 3\delta + 2\widetilde{\delta}$$

Note that the Algorithm 3 works no more than

$$\left\lceil \frac{2L \max_{u \in Q} V[z_0](u)}{\varepsilon} \right\rceil \tag{34}$$

iterations.

Proof. After $(k+1)$-th iteration ($k = 0, 1, 2 \ldots$) from (32) and (33) we have for each $u \in Q$:

$$\psi_\delta(w_k, z_k) \leqslant \psi_\delta(u, z_k) + L_{k+1} V[z_k](u) - L_{k+1} V[w_k](u) - L_{k+1} V[z_k](w_k) + \widetilde{\delta}$$

and

$$\psi_\delta(z_{k+1}, w_k) \leq \psi_\delta(u, w_k) + L_{k+1} V[z_k](u) - L_{k+1} V[z_{k+1}](u) - L_{k+1} V[z_k](z_{k+1}) + \widetilde{\delta}.$$

The first inequality means that

$$\psi_\delta(w_k, z_k) \leqslant \psi_\delta(z_{k+1}, z_k) + L_{k+1} V[z_k](z_{k+1}) - L_{k+1} V[w_k](z_{k+1}) - L_{k+1} V[z_k](w_k) + \widetilde{\delta}.$$

Taking into account (31), we obtain for all $u \in Q$

$$-\psi_\delta(u, w_k) \leq L_{k+1} V[z_k](u) - L_{k+1} V[z_{k+1}](u) + \delta + 2\widetilde{\delta}.$$

So, the following inequality holds:

$$-\sum_{k=0}^{N-1}\frac{\psi_\delta(u,w_k)}{L_{k+1}} \leq V[z_0](u) - V[z_N](u) + S_N(\delta + 2\widetilde{\delta}).$$

By virtue of (27) and the choice of $L_0 \leqslant 2L$, it is guaranteed that $L_{k+1} \leqslant 2L \quad \forall k = \overline{0, N-1}$ and we have from Definition 15

$$\max_{u\in Q}\psi(\widehat{w}_N, u) \leqslant \max_{u\in Q}\psi_\delta(\widehat{w}_N, u) + \delta$$
$$\leqslant -\frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\psi_\delta(u,w_k)}{L_{k+1}} + 2\delta \leqslant \frac{2L\max_{u\in Q}V[z_0](u)}{N} + 3\delta + 2\widetilde{\delta}.$$

□

Remark 8. For universal method to obtain precision ε we can choose $\delta = \frac{\varepsilon}{2}$ and $L = L\left(\frac{\varepsilon}{2}\right)$ according to (29) and (30) and the estimate (34) reduces to

$$\left\lceil 2\inf_{\nu\in[0,1]}\left(\frac{2L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}} \cdot \max_{u\in Q} V[z_0](u)\right\rceil. \tag{35}$$

Note that estimate (35) is optimal for variational inequalities and saddle-point problems in the cases $\nu = 0$ and $\nu = 1$.

Thus, the introduced concept of the (δ, L, V)-model for variational inequalities allows us to extend the previously proposed universal method for VI to a wider class of problems, including *mixed variational inequalities* [5, 38] and *composite saddle-point problems* [10].

Now we introduce inexact model for saddle-point problems. The solution of variational inequalities reduces the so-called saddle points problems, in which for a convex in u and concave in v functional $f(u,v) : \mathbb{R}^{n_1+n_2} \to \mathbb{R}$ ($u \in Q_1 \subset \mathbb{R}^{n_1}$ and $v \in Q_2 \subset \mathbb{R}^{n_2}$) needs to be found the point (u_*, v_*) such that:

$$f(u_*, v) \leqslant f(u_*, v_*) \leqslant f(u, v_*) \tag{36}$$

for arbitrary $u \in Q_1$ and $v \in Q_2$. Let $Q = Q_1 \times Q_2 \subset \mathbb{R}^{n_1+n_2}$. For $x = (u,v) \in Q$, we assume that $||x|| = \sqrt{||u||_1^2 + ||v||_2^2}$ ($||\cdot||_1$ and $||\cdot||_2$ are the norms in the spaces $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$). We agree to denote $x = (u_x, v_x),\ y = (u_y, v_y) \in Q$.

It is well known that for a sufficiently smooth function f with respect to u and v the problem (36) reduces to VI with an operator $g(x) = (f'_u(u_x, v_x),\ -f'_v(u_x, v_x))$.

For saddle-point problems we propose some adaptation of the concept of the (δ, L, V)-model for abstract variational inequality.

Definition 20. We say that the function $\psi_\delta(x,y)$ ($\psi_\delta : \mathbb{R}^{n_1+n_2} \times \mathbb{R}^{n_1\times n_2} \to \mathbb{R}$) is a (δ, L, V)-model for the saddle-point problem (36) if the conditions (ii) – (v) of Definition 15 hold and in addition

$$f(u_y, v_x) - f(u_x, v_y) \leqslant -\psi_\delta(x,y) + \delta \quad \forall x, y \in Q.$$

Example 21. The proposed concept of the (δ, L, V)-model for saddle-point problems is quite applicable, for example, for composite saddle point problems of the form considered in the popular article [10]:

$$f(u,v) = \tilde{f}(u,v) + h(u) - \varphi(v)$$

for some convex in u and concave in v subdifferentiable functions $\tilde{f}$, as well as convex functions h and φ. In this case, we can put

$$\psi_\delta(x,y) = \langle \tilde{g}(y), x-y\rangle + h(u_x) + \varphi(v_x) - h(u_y) - \varphi(v_y),$$

where

$$\tilde{g}(y) = \begin{pmatrix} \tilde{f}'_u(u_y,v_y) \\ -\tilde{f}'_v(u_y,v_y) \end{pmatrix}.$$

Theorem 19 implies

Theorem 22. *If for the saddle problem* (36) *there is a* (δ, L, V)*-model* $\psi_\delta(x,y)$*, then after stopping the algorithm we get a point*

$$\widehat{y}_N = (u_{\widehat{y}_N}, v_{\widehat{y}_N}) := (\widehat{u}_N, \widehat{v}_N) := \frac{1}{S_N}\sum_{k=0}^{N-1}\frac{y_k}{L_{k+1}},$$

for which the following inequality is true:

$$\max_{v\in Q_2} f(\widehat{u}_N, v) - \min_{u\in Q_1} f(u, \widehat{v}_N) \leqslant \frac{2L\max_{(u,v)\in Q} V[u_0,v_0](u,v)}{N} + 2\tilde{\delta} + 2\delta.$$

5 Inexact Model for Strongly Monotone VI

In this section similarly with the concept of (δ, L, μ, m, V)-model in optimization we consider inexact model for VI with a stronger version of monotonicity condition (26).

Definition 23. We say that functional ψ has (δ, L, μ, V)-model $\psi_\delta(x,y)$ at a given point y if the following properties hold for each $x, y, z \in Q$:

(i) $\psi(x,y) \leq \psi_\delta(x,y) + \delta$;

(ii) $\psi_\delta(x,y)$ convex in the first variable;

(iii) $\psi_\delta(x,y)$ continuous in x and y;

(iii) $\psi_\delta(x,x) = 0$;

(iv) (μ*-strong* δ*-monotonicity*)

$$\psi_\delta(x,y) + \psi_\delta(y,x) + \mu\|x-y\|^2 \leq \delta; \quad (37)$$

(v) (*generalized relative smoothness*)

$$\psi_\delta(x,y) \leqslant \psi_\delta(x,z) + \psi_\delta(z,y) + LV[z](x) + LV[y](z) + \delta$$

for some fixed values $L > 0$, $\delta > 0$.

Remark 9. We note that we cannot replace $\|x-y\|^2$ on $V[y](x)$ in (37) since it is essentially used in the proof of Theorem 24.

Now we propose method with linear rate of convergence for VI with (δ, L, μ, V)-model. We slightly modify the assumptions on prox-function $d(x)$. Namely, we assume that $\operatorname{argmin}_{x\in Q} d(x) = 0$ and that d is bounded on the unit ball in the chosen norm $\|\cdot\|$, that is

$$d(x) \leq \frac{\Omega}{2}, \quad \forall x \in Q : \|x\| \leq 1, \tag{38}$$

where Ω is some known constant. Note that for standard proximal setups $\Omega = O(\ln \dim E)$. Finally, we assume that we are given a starting point $x_0 \in Q$ and a number $R_0 > 0$ such that $\|x_0 - x_*\|^2 \leq R_0^2$, where x^* is the solution to abstract VI. The procedure of restating of Algorithm 3 is applicable for abstract strongly monotone variational inequalities.

Algorithm 4 Restarted Generalized Mirror Prox

Require: accuracy $\varepsilon > 0, \mu > 0, \Omega$ s.t. $d(x) \leq \frac{\Omega}{2}\, \forall x \in Q : \|x\| \leq 1; x_0, R_0\ s.t. \|x_0 - x_*\|^2 \leq R_0^2$.

1: Set $p = 0, d_0(x) = d\left(\frac{x-x_0}{R_0}\right)$.
2: **repeat**
3: Set x_{p+1} as the output of Algorithm 3 after N_p iterations of Algorithm 3 with prox-function $d_p(\cdot)$ and stopping criterion $\sum_{k=0}^{N_p-1} \frac{1}{L_{k+1}} \geq \frac{\Omega}{\mu}$.
4: Set $R_{p+1}^2 = R_0^2 \cdot 2^{-(p+1)}$.
5: Set $d_{p+1}(x) \leftarrow d\left(\frac{x-x_{p+1}}{R_{p+1}}\right)$.
6: Set $p = p + 1$.
7: **until** $p > \log_2 \frac{R_0^2}{\varepsilon}$

Ensure: x_{p+1}.

Theorem 24. *Assume that ψ_δ is a (δ, L, μ, V)-model for ψ. Also assume that the prox function $d(x)$ satisfies* (38) *and the starting point $x_0 \in Q$ and a number $R_0 > 0$ are such that $\|x_0 - x_*\|^2 \leq R_0^2$, where x_* is the solution to* (25). *Then, for each $p \geq 0$*

$$\|x_p - x_*\|^2 \leq R_0^2 \cdot 2^{-p} + \frac{\delta}{\mu} + \frac{2\widetilde{\delta}}{\mu} \leq \varepsilon + \frac{\delta}{\mu} + \frac{2\widetilde{\delta}}{\mu}.$$

The total number of iterations of the inner Algorithm 3 does not exceed

$$\left\lceil \frac{2L\Omega}{\mu} \cdot \log_2 \frac{R_0^2}{\varepsilon} \right\rceil, \tag{39}$$

where Ω satisfies (38).

Proof. We show by induction that for $p \geq 0$

$$\|x_p - x_*\|^2 \leq R_0^2 \cdot 2^{-p} + \frac{\delta}{\mu} + \frac{2\widetilde{\delta}}{\mu},$$

which leads to the statement of the Theorem. For $p = 0$ this inequality holds by the Theorem assumption. Assuming that it holds for some $p \geq 0$, our goal is to prove it for $p + 1$ considering the outer iteration $p + 1$. Observe that the function $d_p(x)$ defined in Algorithm 4 is 1-strongly convex w.r.t. the norm $\|\cdot\|/R_p$.

This means that, at each step k of inner Algorithm 3, L_{N_p} changes to $L_{N_p} \cdot R_p^2$. Using the definition of $d_p(\cdot)$ and (38), we have, since $x_p = \operatorname{argmin}_{x \in Q} d_p(x)$

$$V_p[x_p](x_*) = d_p(x_*) - d_p(x_p) - \langle \nabla d_p(x_p), x_* - x_p \rangle \leq d_p(x_*) \leq \frac{\Omega}{2}.$$

Denote by

$$S_{N_p} := \sum_{k=0}^{N_p - 1} \frac{1}{L_{k+1}}.$$

Thus, by Theorem 19, taking $u = x_*$, we obtain

$$-\frac{1}{S_{N_p}} \sum_{k=0}^{N_p-1} \frac{\psi_\delta(x_*, w_k)}{L_{k+1}} \leq \frac{R_p^2 V_p[x_p](x_*)}{S_{N_p}} + \delta + 2\widetilde{\delta} \leq \frac{\Omega R_p^2}{2 S_{N_p}} + \delta + 2\widetilde{\delta}.$$

Since the operator ψ is continuous and abstract monotone, we can assume that the solution to weak VI (23) is also a strong solution and $-\psi(w_k, x_*) \leq 0$, $k = 0, ..., N_p - 1$ and, by Definition 23 (i), $-\psi_\delta(\omega_k, x_*) \leq \delta$ ($k = 0, \ldots, N_p - 1$). This and (37) gives, that for each $k = 0, ..., N_p - 1$,

$$\begin{aligned} -\psi_\delta(x_*, w_k) &\geq -\delta - \psi_\delta(x_*, w_k) - \psi_\delta(w_k, x_*) \geq -\delta + \mu \|w_k - x_*\|^2, \\ -\psi_\delta(x_*, \omega_k) &\geq -\delta - \psi_\delta(x_*, \omega_k) - \psi_\delta(\omega_k, x_*) \geq -\delta + \mu \|\omega_k - x_*\|^2. \end{aligned}$$

Thus, by convexity of the squared norm, we obtain

$$\begin{aligned} -2\delta + \mu \|x_{p+1} - x_*\|^2 &= -2\delta + \mu \left\| \frac{1}{S_{N_p}} \sum_{k=0}^{N_p-1} \frac{w_k}{L_{k+1}} - x_* \right\|^2 \leq -2\delta + \frac{\mu}{S_{N_p}} \sum_{k=0}^{N_p-1} \frac{\|w_k - x_*\|^2}{L_{k+1}} \\ &\leq -2\delta - \frac{1}{S_{N_p}} \sum_{k=0}^{N_p-1} \frac{\psi_\delta(x_*, w_k)}{L_{k+1}} \leq \frac{\Omega R_p^2}{2 S_{N_p}} - \delta + 2\widetilde{\delta}. \end{aligned}$$

Using the stopping criterion $S_{N_p} \geq \frac{\Omega}{\mu}$ we have

$$\begin{aligned} \|x_{p+1} - x_*\|^2 &\leq \frac{R_p^2}{2} + \frac{\delta + 2\widetilde{\delta}}{\mu} = \frac{1}{2} R_0^2 \cdot 2^{-p} + \frac{\delta + 2\widetilde{\delta}}{\mu} \\ &= R_0^2 \cdot 2^{-(p+1)} + \frac{\delta + 2\widetilde{\delta}}{\mu}, \end{aligned}$$

which finishes proof by induction. □

Remark 10. If for some $m > 0$ $\psi_\delta(x, y)$ is a m-strongly convex function in x then for Algorithm 4 we can prove estimate

$$\|x_p - x_*\|^2 \leq R_0^2 \cdot 2^{-p} + \frac{\delta}{m + \mu} + \frac{2\widetilde{\delta}}{m + \mu} \leq \varepsilon + \frac{\delta}{m + \mu} + \frac{2\widetilde{\delta}}{m + \mu}$$

for each $p \geq 0$ and instead of (39) we obtain

$$\left\lceil \frac{2L\Omega}{m + \mu} \cdot \log_2 \frac{R_0^2}{\varepsilon} \right\rceil.$$

DOI 10.20347/WIAS.PREPRINT.2709

6 Conclusion

In this paper, we consider convex optimization problem (1). It is well known (see [15, 21, 33]) that if there is an inexact gradient $\nabla_\delta f(y)$ of f, s.t., for all $x, y \in Q$,

$$f(y) + \langle \nabla_\delta f(y), x - y\rangle - \delta_1 \le f(x) \le f(y) + \langle \nabla_\delta f(y), x - y\rangle + \frac{L}{2}\|x-y\|_2^2 + \delta_2, \quad (40)$$

then the corresponding versions of Gradient Method (GM) and Fast Gradient Method (FGM) have convergence rate

$$f(x_N) - f(x_*) = O\left(\frac{LR^2}{N^p} + \delta_1 + N^{p-1}\delta_2\right), \quad (41)$$

where $p = 1$ corresponds to GM and $p = 2$ corresponds to FGM, x_* – is a solution of (1), R is an upper bound for $\|x_0 - x_*\|_2$. We show[2] that under an appropriate generalization of (40) to

$$f(y) + \psi_\delta(x, y) - \delta_1 \le f(x) \le f(y) + \psi_\delta(x, y) + \frac{L}{2}\|x-y\|_2^2 + \delta_2$$

as well as appropriate generalizations of GM and FGM, the sequence generated by these methods satisfy (41). It should be noted that, despite there are many variants of FGM, we are aware of only one which can be generalized for problems with inexact model, namely accelerated mirror descent type of FGM [23, 41, 66]. An important feature of this method is that it requires only one projection step on each iteration. A primal-dual extension of the proposed framework is made in [67].

We also show that in the case of μ-strongly convex objective (model) the estimate (41) can be improved to

$$f(x_N) - f(x_*) = O\left(\Delta f \exp\left(-O(1)\left(\frac{\mu}{L}\right)^{\frac{1}{p}} N\right) + \delta_1 + \left(\frac{L}{\mu}\right)^{\frac{p-1}{2}} \delta_2\right),$$

where $\Delta f = f(x^0) - f(x_*)$, $p = 1$ for GM and $p = 2$ for restarted FGM.

In this paper we also propose a generalization of this inexact model framework for saddle-point problems and variational inequalities. We consider universal (adaptive) generalizations in the spirit of [53] and relative smoothness generalizations, generalizing the framework [6, 45] from optimization problems to saddle-point problems and VI. We also investigate the sensitivity of the convergence results to the accuracy of auxiliary minimization on each iteration.

Due to the lack of the space we only briefly mention here an extension of our framework for block-coordinate descent using the randomized version of FGM in [25] and stochastic optimization problems using the ideas from [29]. For the latter case we indicate that if we additionally assume that δ_1, δ_2 are independently chosen at each iteration random variables such that $\mathbb{E}\delta_1 = 0$ and $\delta_1, \sqrt{\delta_2}$ have correspondingly $(\delta_1')^2$-subgaussian variance and δ_2'-subgaussian second moment [33] then with high probability (41) changes to

$$f(x_N) - f(x_*) = \tilde{O}\left(\frac{LR^2}{N^p} + \frac{\delta_1'}{\sqrt{N}} + N^{p-1}\delta_2'\right).$$

From this result and mini-batch trick [29] one can obtain the main estimates for convex and strongly convex stochastic optimization problems [19, 27, 33, 40].

As further generalizations we point a generalization for tensor methods [30, 55] and for incremental and variance reduction methods for finite-sum minimization [13, 43].

[2]For simplicity in the paper we consider the case $\delta_1 = \delta_2 = \delta$, but one can easily rewrite all the results of this paper to obtain (41). See [33] for details.

References

[1] A. Anikin, P. Dvurechensky, A. Gasnikov, A. Golov, A. Gornov, Y. Maximov, M. Mendel, and V. Spokoiny, *Modern efficient numerical approaches to regularized regression problems in application to traffic demands matrix calculation from link loads*, in *Proceedings of International conference ITAS-2015. Russia, Sochi*. 2015. arXiv:1508.00858.

[2] A.S. Anikin, A.V. Gasnikov, P.E. Dvurechensky, A.I. Tyurin, and A.V. Chernov, *Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints*, Computational Mathematics and Mathematical Physics 57 (2017), pp. 1262–1276.

[3] K. Antonakopoulos, V. Belmega, and P. Mertikopoulos, *An adaptive mirror-prox method for variational inequalities with singular operators*, in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché Buc, E. Fox, and R. Garnett, eds., Curran Associates, Inc., 2019, pp. 8455–8465.

[4] D. Baimurzina, A. Gasnikov, E. Gasnikova, P. Dvurechensky, E. Ershov, M. Kubentaeva, and A. Lagunovskaya, *Universal similar triangulars method for searching equilibriums in traffic flow distribution models*, Journal of Computational Mathematics and Mathematical Physics 59 (2019), pp. 21–36.

[5] T.Q. Bao and P.Q. Khanh, *Some algorithms for solving mixed variational inequalities*, Acta Mathematica Vietnamica 31 (2006), pp. 77–98.

[6] H.H. Bauschke, J. Bolte, and M. Teboulle, *A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications*, Mathematics of Operations Research 42 (2016), pp. 330–348.

[7] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences 2 (2009), pp. 183–202. Available at `https://doi.org/10.1137/080716542`.

[8] L. Bogolubsky, P. Dvurechensky, A. Gasnikov, G. Gusev, Y. Nesterov, A.M. Raigorodskii, A. Tikhonov, and M. Zhukovskii, *Learning supervised pagerank with gradient-based and gradient-free optimization methods*, in *Advances in Neural Information Processing Systems 29*, D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett, eds., Curran Associates, Inc., 2016, pp. 4914–4922. arXiv:1603.00717.

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*, NY Cambridge University Press, 2004.

[10] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Mathematical Imaging and Vision 40 (2011), pp. 120–145.

[11] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using bregman functions*, SIAM Journal on Optimization 3 (1993), pp. 538–543.

[12] A. Chernov, P. Dvurechensky, and A. Gasnikov, *Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints*, in *Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings*, Y. Kochetov, M. Khachay, V. Beresnev, E. Nurminski, and P. Pardalos, eds. Springer International Publishing, 2016, pp. 391–403.

[13] A. Defazio, *A simple practical accelerated method for finite sums*, in *Advances in neural information processing systems*. 2016, pp. 676–684.

[14] O. Devolder, F. Glineur, and Y. Nesterov, *First-order methods with inexact oracle: the strongly convex case*, CORE Discussion Papers 2013/16 (2013).

[15] O. Devolder, F. Glineur, and Y. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Mathematical Programming 146 (2014), pp. 37–75. Available at `http://dx.doi.org/10.1007/s10107-013-0677-5`.

[16] D. Drusvyatskiy, A.D. Ioffe, and A.S. Lewis, *Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria*, Mathematical Programming (2019). arXiv:1610.03446.

[17] D. Dvinskikh, E. Gorbunov, A. Gasnikov, P. Dvurechensky, and C.A. Uribe, *On Primal-Dual Approach for Distributed Stochastic Convex Optimization over Networks*, in *2019 IEEE Conference on Decision and Control (CDC)*. 2019. (accepted), arXiv:1903.09844.

[18] P. Dvurechensky, D. Dvinskikh, A. Gasnikov, C.A. Uribe, and A. Nedić, *Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters*, in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds. Curran Associates, Inc., NIPS'18, 2018, pp. 10783–10793. Available at `http://papers.nips.cc/paper/8274-decentralize-and-randomize-faster-algorithm-for-wasserstein-baryc` `pdf`, arXiv:1802.04367.

[19] P. Dvurechensky and A. Gasnikov, *Stochastic intermediate gradient method for convex problems with stochastic inexact oracle*, Journal of Optimization Theory and Applications 171 (2016), pp. 121–145. Available at `http://dx.doi.org/10.1007/s10957-016-0999-6`.

[20] P. Dvurechensky, A. Gasnikov, E. Gasnikova, S. Matsievsky, A. Rodomanov, and I. Usik, *Primal-Dual Method for Searching Equilibrium in Hierarchical Congestion Population Games*, in *Supplementary Proceedings of the 9th International Conference on Discrete Optimization and Operations Research and Scientific School (DOOR 2016) Vladivostok, Russia, September 19 - 23, 2016*. 2016, pp. 584–595. arXiv:1606.08988.

[21] P. Dvurechensky, A. Gasnikov, and D. Kamzolov, *Universal intermediate gradient method for convex problems with inexact oracle*, arXiv:1712.06036, Opt. Meth. & Software (accepted) (2019). Available at `http://dx.doi.org/10.1080/10556788.2019.1711079`.

[22] P. Dvurechensky, A. Gasnikov, and A. Kroshnin, *Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm*, in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, eds., Proceedings of Machine Learning Research Vol. 80. 2018, pp. 1367–1376. arXiv:1802.04367.

[23] P. Dvurechensky, A. Gasnikov, S. Omelchenko, and A. Tiurin, *Adaptive similar triangles method: a stable alternative to sinkhorn's algorithm for regularized optimal transport*, arXiv:1706.07622 (2017).

[24] P. Dvurechensky, A. Gasnikov, F. Stonyakin, and A. Titov, *Generalized Mirror Prox: Solving variational inequalities with monotone operator, inexact oracle, and unknown Hölder parameters*, arXiv:1806.05140 (2018).

[25] P. Dvurechensky, A. Gasnikov, and A. Tiurin, *Randomized similar triangles method: A unifying framework for accelerated randomized optimization methods (coordinate descent, directional search, derivative-free method)*, arXiv:1707.08486 (2017).

[26] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly 3 (1956), pp. 95–110.

[27] A.V. Gasnikov and P.E. Dvurechensky, *Stochastic intermediate gradient method for convex optimization problems*, Doklady Mathematics 93 (2016), pp. 148–151.

[28] A.V. Gasnikov, P.E. Dvurechensky, F.S. Stonyakin, and A.A. Titov, *An adaptive proximal method for variational inequalities*, Computational Mathematics and Mathematical Physics 59 (2019), pp. 836–841. Available at `https://doi.org/10.1134/S0965542519050075`.

[29] A. Gasnikov, *Universal gradient descent*, arXiv preprint arXiv:1711.00394 (2017).

[30] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, C.A. Uribe, B. Jiang, H. Wang, S. Zhang, S. Bubeck, Q. Jiang, Y.T. Lee, Y. Li, and A. Sidford, *Near Optimal Methods for Minimizing Convex Functions with Lipschitz p-th Derivatives*, in *Proceedings of the Thirty-Second Conference on Learning Theory*, A. Beygelzimer and D. Hsu, eds., Proceedings of Machine Learning Research Vol. 99, 25–28 Jun, Phoenix, USA. PMLR, 2019, pp. 1392–1393. Available at `http://proceedings.mlr.press/v99/gasnikov19b.html`, arXiv:1809.00382.

[31] A. Gasnikov and A. Tyurin, *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 (2019), pp. 1085–1097.

[32] A.V. Gasnikov and Y.E. Nesterov, *Universal method for stochastic composite optimization problems*, Computational Mathematics and Mathematical Physics 58 (2018), pp. 48–64. First appeared in arXiv:1604.05275.

[33] E. Gorbunov, D. Dvinskikh, and A. Gasnikov, *Optimal decentralized distributed algorithms for stochastic convex optimization*, arXiv preprint arXiv:1911.07363 (2019).

[34] S.V. Guminov, Y.E. Nesterov, P.E. Dvurechensky, and A.V. Gasnikov, *Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems*, Doklady Mathematics 99 (2019), pp. 125–128.

[35] Z. Harchaoui, A. Juditsky, and A. Nemirovski, *Conditional gradient algorithms for norm-regularized smooth convex optimization*, // Mathematical Programming 152 (2015), pp. 75–112.

[36] A.N. Iusem, A. Jofré, R.I. Oliveira, and P. Thompson, *Variance-based extragradient methods with line search for stochastic variational inequalities*, SIAM Journal on Optimization 29 (2019), pp. 175–206. arXiv:1703.00262.

[37] M. Jaggi, *Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization.*, in *ICML (1)*. 2013, pp. 427–435.

[38] I. Konnov and R. Salahutdin, *Two-level iterative method for non-stationary mixed variational inequalities*, Izvestija vysshih uchebnyh zavedenij. Matematika 61 (2017), pp. 50–61.

[39] A. Kroshnin, N. Tupitsa, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and C. Uribe, *On the Complexity of Approximating Wasserstein Barycenters*, in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, eds., Proceedings of Machine Learning Research Vol. 97, 09–15 Jun, Long Beach, California, USA. PMLR, 2019, pp. 3530–3540. arXiv:1901.08686.

[40] A. Kulunchakov and J. Mairal, *Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise*, arXiv preprint arXiv:1901.08788 (2019).

[41] G. Lan, *An optimal method for stochastic composite optimization*, Mathematical Programming 133 (2012), pp. 365–397. Available at `https://doi.org/10.1007/s10107-010-0434-y`, Firs appeared in June 2008.

[42] G. Lan, *Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization*, Mathematical Programming 149 (2015), pp. 1–45.

[43] G. Lan and Y. Zhou, *Random gradient extrapolation for distributed and stochastic optimization*, SIAM Journal on Optimization 28 (2018), pp. 2753–2782.

[44] H. Lin, J. Mairal, and Z. Harchaoui, *A Universal Catalyst for First-order Optimization*, in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA. MIT Press, NIPS'15, 2015, pp. 3384–3392. Available at `http://dl.acm.org/citation.cfm?id=2969442.2969617`.

[45] H. Lu, R.M. Freund, and Y. Nesterov, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM Journal on Optimization 28 (2018), pp. 333–354.

[46] J. Mairal, *Optimization with first-order surrogate functions*, in *International Conference on Machine Learning*. 2013, pp. 783–791.

[47] G. Mastroeni, *On auxiliary principle for equilibrium problems*, Publicatione del Departimento di Mathematica Dell'Universita di Pisa 3 (2000), pp. 1244–1258.

[48] A. Nemirovski, *Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM Journal on Optimization 15 (2004), pp. 229–251.

[49] A. Nemirovski, S. Onn, and U.G. Rothblum, *Accuracy certificates for computational problems with convex structure*, Mathematics of Operations Research 35 (2010), pp. 52–78.

[50] A. Nemirovskii and Y. Nesterov, *Optimal methods of smooth convex minimization*, USSR Computational Mathematics and Mathematical Physics 25 (1985), pp. 21 – 30. Available at `http://www.sciencedirect.com/science/article/pii/0041555385901004`.

[51] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Mathematical Programming 120 (2009), pp. 221–259. Available at `https://doi.org/10.1007/s10107-007-0149-x`, First appeared in 2005 as CORE discussion paper 2005/67.

[52] Y. Nesterov, *Gradient methods for minimizing composite functions*, Mathematical Programming 140 (2013), pp. 125–161. First appeared in 2007 as CORE discussion paper 2007/76.

[53] Y. Nesterov, *Universal gradient methods for convex optimization problems*, Mathematical Programming 152 (2015), pp. 381–404. Available at http://dx.doi.org/10.1007/s10107-014-0790-0.

[54] Y. Nesterov, *Complexity bounds for primal-dual methods minimizing the model of objective function*, Math. Program. 171 (2018), pp. 311–330. Available at https://doi.org/10.1007/s10107-017-1188-6.

[55] Y. Nesterov, *Implementable tensor methods in unconstrained convex optimization*, Tech. Rep., CORE UCL, 2018. Available at https://alfresco.uclouvain.be/alfresco/service/guest/streamDownload/workspace/SpacesStore/aabc2323-0bc1-40d4-9653-1c29971e7bd8/coredp2018_05web.pdf, CORE Discussion Paper 2018/05.

[56] Y. Nesterov, *Lectures on convex optimization*, Vol. 137, Springer International Publishing, 2018.

[57] Y. Nesterov, *Soft clustering by convex electoral model*, CORE Discussion Papers 2018001, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2018. Available at https://ideas.repec.org/p/cor/louvco/2018001.html.

[58] Y. Nesterov, A. Gasnikov, S. Guminov, and P. Dvurechensky, *Primal-dual accelerated gradient methods with small-dimensional relaxation oracle*, arXiv:1809.05895 (2018). Submitted to Optimization Methods & Software.

[59] Y. Nesterov and B. Polyak, *Cubic regularization of newton method and its global performance*, Mathematical Programming 108 (2006), pp. 177–205. Available at http://dx.doi.org/10.1007/s10107-006-0706-8.

[60] P. Ochs, J. Fadili, and T. Brox, *Non-smooth non-convex bregman minimization: Unification and new algorithms*, arXiv preprint arXiv:1707.02278 (2017).

[61] A. Ogaltsov, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and V. Spokoiny, *Adaptive gradient descent for convex and non-convex stochastic optimization*, arXiv:1911.08380 (2019). Submitted to IFAC 2020 Congress.

[62] Y. Ouyang and Y. Xu, *Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems*, arXiv preprint arXiv: 1808.02901 (2018).

[63] N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends® in Optimization 1 (2014), pp. 127–239. Available at http://dx.doi.org/10.1561/2400000003.

[64] K. Scaman, F. Bach, S. Bubeck, Y.T. Lee, and L. Massoulié, *Optimal Algorithms for Smooth and Strongly Convex Distributed Optimization in Networks*, in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y.W. Teh, eds., Proceedings of Machine Learning Research Vol. 70, 06–11 Aug, International Convention Centre, Sydney, Australia. PMLR, 2017, pp. 3027–3036. Available at http://proceedings.mlr.press/v70/scaman17a.html.

[65] F.S. Stonyakin, D. Dvinskikh, P. Dvurechensky, A. Kroshnin, O. Kuznetsova, A. Agafonov, A. Gasnikov, A. Tyurin, C.A. Uribe, D. Pasechnyuk, and S. Artamonov, *Gradient Methods for Problems with Inexact Model of the Objective*, in *Mathematical Optimization Theory and Operations Research*, M. Khachay, Y. Kochetov, and P. Pardalos, eds., Cham. Springer International Publishing, 2019, pp. 97–114. arXiv:1902.09001.

[66] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, Tech. Rep., MIT, 2008. Available at http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf.

[67] A. Tyurin, *Primal-dual fast gradient method with a model*, arXiv preprint arXiv:1906.10107 (2019).

[68] C.A. Uribe, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and A. Nedić, *Distributed Computation of Wasserstein Barycenters over Networks*, in *2018 IEEE 57th Annual Conference on Decision and Control (CDC)*. 2018. Accepted, arXiv:1803.02933.

[69] Y. Xie, X. Wang, R. Wang, and H. Zha, *A fast proximal point method for wasserstein distance*, arXiv preprint arXiv:1802.04307 (2018).

A Auxiliary facts

Let us commenton the inexact solution of the auxiliary problem.

Remark 11. We can show that if $\widetilde{x} \in \mathrm{Arg}\min^{\widetilde{\delta}}_{x\in Q} \Psi(x)$, then $\Psi(\widetilde{x}) - \Psi(x_*) \leq \delta$. Indeed, we have $\Psi(x_*) \geq \Psi(\widetilde{x}) + \langle h, x_* - \widetilde{x}\rangle \geq \Psi(\widetilde{x}) - \widetilde{\delta}$. The converse statement is not always true. However, for some general cases we can resolve the problem (see [31] and Example 25).

Example 25. Let us show an example, how we can resolve the problem in Remark 11. Note, that if $\Psi(x)$ is μ-strongly convex; has L-Lipschitz continuous gradient in $\|\cdot\|$ norm (To say more precisely

$$L = \max_{\|h\|\leq 1, x\in[\widetilde{x},x_*]} \langle h, \nabla^2\Psi(x)h\rangle.$$

and $R = \max_{x,y\in Q}\|x-y\|$, then $\Psi(\widetilde{x}) - \Psi(x_*) \leq \widetilde{\epsilon}$ entails that [65]

$$\widetilde{\delta} \leq (LR + \|\nabla\Psi(x_*)\|_*)\sqrt{2\tilde{\varepsilon}/\mu}, \tag{42}$$

where $x_* = \mathrm{argmin}_{x\in Q}\Psi(x)$. If one can guarantee that $\nabla\Psi(x_*) = 0$, then (42) can be improved $\widetilde{\delta} \leq R\sqrt{2L\widetilde{\varepsilon}}$.

B Proof for Lemma 4

Proof. In view of definition (14) of sequence α_{k+1}, we have:

$$\begin{aligned} A_N &\leq A_N(1+\mu A_{N-1} + mA_{N-1}) = L_N(A_N - A_{N-1})^2 \\ &\leq L_N(A_N^{1/2} - A_{N-1}^{1/2})^2(A_N^{1/2} + A_{N-1}^{1/2})^2 \leq 2L_N A_N (A_N^{1/2} - A_{N-1}^{1/2})^2. \end{aligned}$$

We can see that

$$A_N^{1/2} \geq A_{N-1}^{1/2} + \frac{1}{2L_N}$$

and

$$A_N \geq \frac{1}{2}\left(\sum_{k=0}^{N-1}\frac{1}{\sqrt{L_{k+1}}}\right)^2.$$

For the case when $\mu + m > 0$ we obtain:

$$(\mu + m)A_{N-1}A_N \leq A_N(1 + \mu A_{N-1} + mA_{N-1}) \leq 2L_N A_N(A_N^{1/2} - A_{N-1}^{1/2})^2.$$

From the fact that $A_1 = 1/L_1$ and the last inequality we can show that

$$A_N^{1/2} \geq \left(1 + \sqrt{\frac{\mu + m}{2L_N}}\right) A_{N-1}^{1/2} \geq \frac{1}{\sqrt{L_1}} \prod_{k=1}^{N-1} \left(1 + \sqrt{\frac{\mu + m}{2L_{k+1}}}\right).$$

□

C Fast gradient method with $(\delta, L, \mu, m, V, \|\cdot\|)$-model. Restart technique.

Let us consider the case of a strongly convex functional f and show how to accelerate the work of the Algorithm 1 using the restart technique. Let us assume that

$$\psi_\delta(x, x_*) \geq 0 \ \ \forall x \in Q.$$

Note that this assumption is natural, e.g. $\psi_\delta(x, y) := \langle \nabla f(y), x - y\rangle \ \ \forall x, y \in Q$. We also modify the concept of μ-strong convexity in the following way

Definition 26. Say that the function f is a left relative μ-strongly convex if the following inequality

$$\mu V[x](y) \leq f(x) - f(y) - \psi_\delta(x, y) \quad \forall x, y \in Q$$

holds.

Remark 12. Let us remind that if $d(x-y) \leq C_n \|x - y\|^2$ for $C_n = O(\log n)$, (where n is dimension of vectors from Q) then $V[y](x) \leq C_n \|x - y\|^2$. This assumption is true for many standard proximal setups. In this case the condition of (μC_n)-strong convexity

$$\mu C_n \|x - y\|^2 + f_\delta(y) + \psi_\delta(x, y) \leqslant f(x)$$

entails right relative strong convexity:

$$\mu V[y](x) + f_\delta(y) + \psi_\delta(x, y) \leqslant f(x).$$

Note that concepts of right and left relative strongly convexity from Definitions 1 and 26 are equivalent in the case of an assumption from Remark 12 ($V[x](y) \leq C_n\|x - y\|^2$ for each $x, y \in Q$).

We show that using the restart technique can also accelerate the work of non-adaptive version of Algorithm 1 ($L_{k+1} = L$) for $(\delta, L, 0, 0, V, \|\cdot\|)$-model and relative μ-strogly convex function f in sense Definition 26:

$$\mu V[x](y) + f(y) + \psi_\delta(x, y) - \delta \leq f(x) \leq f(y) + \psi_\delta(x, y) + \frac{L}{2}\|x - y\|^2 + \delta.$$

for each $x, y \in Q$. By Theorem 14 and Remark 5:

$$f(x_N) - f(x_*) \leqslant \frac{4LV[x_0](x_*)}{N^2} + \frac{4L\widetilde{\delta}}{N} + 2N\delta. \tag{43}$$

Consider the case of relatively μ-strongly convex function f. We will use the restart technique to obtain the method for strongly convex functions.

Theorem 27. *Let f be a left relative μ-strongly convex function and $\psi_\delta(x,y)$ is a $(\delta, L, 0, 0, V, \|\cdot\|)$-model. Let δ and $\widetilde{\delta}$ satisfy $\frac{4\mu\sqrt{10}}{L}\left(5\delta\left\lceil\sqrt{\frac{L}{\mu}}\right\rceil^3 + \widetilde{\delta}L\left\lceil\sqrt{\frac{L}{\mu}}\right\rceil\right) \leqslant \varepsilon$. Then, using the restarts of Algorithm 1, we need*

$$N = \left\lceil \log_2 \frac{\mu R^2}{\varepsilon}\right\rceil \cdot \left\lceil\sqrt{\frac{10L}{\mu}}\right\rceil.$$

iterations to achieve ε accuracy by function: $f(x_N) - f(x_*) \leqslant \varepsilon$.

Proof. By (43) and Definition 26:

$$\mu V[x_{N_1}](x_*) \le f(x_{N_1}) - f(x_*) \le \frac{4LV[x_0](x_*)}{N^2} + \frac{4L\widetilde{\delta}}{N} + 2N\delta. \tag{44}$$

Let's choose N_1 so that the following inequality holds:

$$\frac{4L\widetilde{\delta}}{N_1} + 2N_1\delta \le \frac{LV[x_0](x_*)}{N_1^2}. \tag{45}$$

We restart method as $V[x_{N_1}](x_*) \le \frac{V[x_0](x_*)}{2}$. Using (44), we obtain an estimation for the number of iterations on the first restart: $\frac{5L}{\mu N_1^2} \le \frac{1}{2}$. Therefore, let's choose

$$N_1 = \left\lceil\sqrt{\frac{10L}{\mu}}\right\rceil. \tag{46}$$

Then after N_1 iterations we restart method. Similarly, we restart after N_2 iterations, such that $V[x_{N_2}](x_*) \le \frac{V[x_{N_1}](x_*)}{2}$. We obtain $N_2 = \left\lceil\sqrt{\frac{10L}{\mu}}\right\rceil$. So, after p-th restart the total number of iterations is $M = p \cdot \left\lceil\sqrt{\frac{10L}{\mu}}\right\rceil$.

Now let's consider how many iterations is needed to achieve accuracy $\varepsilon = f(x_{N_p}) - f(x_*)$. From (43) and (46) we take $p = \left\lceil \log_2 \frac{\mu R^2}{\varepsilon}\right\rceil$ and the total number of iterations is $M = \left\lceil \log_2 \frac{\mu R^2}{\varepsilon}\right\rceil \cdot \left\lceil\sqrt{\frac{10L}{\mu}}\right\rceil$.

We have chosen our errors as δ and $\widetilde{\delta}$ to satisfy (45). Indeed, from (45) using $N_k = \left\lceil\sqrt{\frac{10L}{\mu}}\right\rceil$ we can deduce the following inequality:

$$\varepsilon \ge \frac{4\sqrt{10}\mu}{L}\left(5\delta\left\lceil\sqrt{\frac{L}{\mu}}\right\rceil^3 + \widetilde{\delta}L\left\lceil\sqrt{\frac{L}{\mu}}\right\rceil\right).$$

One can see that such a choice of δ and $\widetilde{\delta}$ as above satisfies that inequality. □

Remark 13. Partially, we can choose $\delta = O\left(\frac{\varepsilon L}{\mu\left\lceil\sqrt{\frac{10L}{\mu}}\right\rceil^3}\right)$ and $\widetilde{\delta} = O\left(\frac{\varepsilon}{\mu\left\lceil\sqrt{\frac{10L}{\mu}}\right\rceil}\right)$.