Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Inexact model: A framework for optimization and variational inequalities

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

submitted: January 27, 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

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

[5] Weierstrass Institute
Mohrenstr. 39
10117 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

No. 2679
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.

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 model: A framework for optimization and variational inequalities

Fedor Stonyakin, Alexander Gasnikov, Alexander Tyurin, Dmitry Pasechnyuk, Artem Agafonov, Pavel Dvurechensky, Darina Dvinskikh, 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, 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 method for variational inequalities with composite structure. This method works for smooth and non-smooth problems with optimal complexity without a priori knowledge of the problem smoothness. We also generalize our framework for strongly convex objectives and strongly monotone variational inequalities.

1 Introduction

Let us consider the following convex optimization problem

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

where f is a convex function and Q is a convex subset of finite-dimensional vector space E. 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 gradient

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

The step of gradient method is obtained by the minimization of this model [35]. More general models are constructed based on regularized second-order Taylor expansion [37] or other Taylor-like models [9] or objective surrogates [25]. Another example is the conditional gradient method [15], 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 first proposed in [37] and applied to first-order methods in [32, 33, 11]. Recently, first-order optimization methods were generalized to the so-called relative smoothness framework [2, 24, 38], where $\frac{1}{2}\|x_k - y\|_2^2$ in the quadratic model (2) for the objective is replaced with general Bregman divergence.

The literature on first-order methods [8, 10] considers also gradient methods with inexact information, relaxing the model (2) to

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

DOI 10.20347/WIAS.PREPRINT.2679 Berlin 2020

with $(f_\delta, \nabla f_\delta)$ called inexact oracle and this model being an upper bound for the objective.

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. This 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 [15] and proximal gradient method [7]. First attempts to propose this generalization were made in [43, 16] for non-accelerated methods, yet without proofs, and in [17] for accelerated methods, yet without relative smoothness paradigm. In this paper we make a review of these results for completeness and extend them to more general cases, including proofs, adaptivity to the parameter L, extension for strongly convex case and relative smoothness. As an application of our general framework, we also develop a universal conditional gradient method, providing a parameter-free generalization of the results of [34]. We believe that our model is flexible enough to be extended for problems with primal-dual structure [31, 28], e.g. for problems with linear constraints [12], and also for random block-coordinate descent [14].

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

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

where $g(x) = \nabla f(x)$. This problem is also equivalent to finding saddle-point of a 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, more novel 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 [27], its adaptive version [18] and universal version [13] to variational inequalities with such general inexact model of the operator. As a partucular case, our approach allows to partially reproduce the results of [6]. We also apply the general framework for variational inequalities to saddle-point problems.

In general, 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 consider minimization problems and inexact model of the objective function. 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 2.4, we construct universal conditional gradient (Frank–Wolfe) method using FGM with inexact projection.

In Section 3, we present inexact (δ, L, μ)-model which is compatible with relative smoothness paradigm. We obtain convergence rates of adaptive and non-adaptive GM for this case. We especially consider the case of m-strong convexity of the model, which is motivated by composite optimization problems [32] with strongly convex composite term. We also illustrate the definition of inexact (δ, L, μ)-model by several examples.

In Section 4 we generalize inexact (δ, L)-model and (δ, L, μ)-model to variational inequalities and saddle-point problems. In the former case, we construct an adaptive generalization (Algorithm 5) of the Mirror-Prox algorithm for variational inequalities and saddle-point problems with such inexact models. In the case of (δ, L, μ)-model 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.

In the Appendix we give some numerical experiments to compare adaptive and non-adaptive versions of gradient method with (δ, L, μ)-model and to compare Algorithms 5 and 6.

The contribution of this paper is follows:

1 Using FGM with inexact model we construct a universal conditional gradient (Frank–Wolfe) method.

2 We present (δ, L, μ)-model for optimization problems. Also, we derive convergence rates for non-adaptive and adaptive GM for optimization problems with (δ, L, μ)-model. Specially we consider the case of m-strong convexity of such model.

3 We propose generalizations of (δ, L)-model and (δ, L, μ)-model for variational inequalities and saddle-point problems. Specially we consider the case of m-strong convexity of such model. We obtain convergence rates for adaptive versions of Mirror-Prox algorithm for problems with inexact model.

2 Inexact Model for Minimization

2.1 Definitions and Examples

We start with the general notation. Let E be a finite-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 $\|g\|_* = \max\limits_x \{\langle g, x\rangle, \|x\| \leq 1\}$. We use $\nabla f(x)$ to denote any (sub)gradient of a function f at a point $x \in \mathrm{dom} f$.

Definition 1. Suppose that for a given point $y \in Q$ and for all $x \in Q$ the inequality

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

holds for some $\psi_\delta(x, y)$, $f_\delta(y) \in [f(y) - \delta; f(y)]$, L, $\delta > 0$ and $V[y](x) = d(x) - d(y) - \langle \nabla d(y), x - y\rangle$, where $d(x)$ is convex function on Q . Let $\psi_\delta(x, y)$ be convex for $x \in Q$ and satisfy $\psi_\delta(x, x) = 0$ for all $x \in Q$. Then we say that $\psi_\delta(x, y)$ is (δ, L)-model of the function f at a given point y with respect to (w.r.t.) $V[y](x)$.

Remark 1. Function $V[y](x)$, defined above as $V[y](x) = d(x) - d(y) - \langle \nabla d(y), x - y\rangle$ is often called Bregman divergence [4]. But typically it should be added the (1-SC) assumption in definition: $d(x)$ is 1-strongly convex on Q w.r.t. $\|\cdot\|$-norm. Note that in Definition 1 we do not need such assumption. But sometimes we also use the definition of $V[y](x)$ in the description of algorithms below and corresponding theorems of convergences rates separately. If additionally the condition (1-SC) is required we write it explicitly.

Remark 2. We change 'w.r.t $V[y](x)$' to 'w.r.t. $\|\cdot\|$-norm' in Definition 1 if we use $\frac{1}{2}\|x - y\|^2$ instead of $V[y](x)$.

Definition 2. For a convex optimization problem

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

we denote by Arg $\min^{\widetilde{\delta}}_{x\in Q} \Psi(x)$ a set of $\widetilde{x}$:

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

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

Remark 3. We can show that if $\widetilde{x} \in$ 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 [17] and Example 3).

Example 3. Let us show an example, how we can resolve the problem in Remark 3. 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 [43]

$$\widetilde{\delta} \leq (LR + \|\nabla\Psi(x_*)\|_*)\sqrt{2\widetilde{\varepsilon}/\mu}, \tag{3}$$

where $x_* = \operatorname{argmin}_{x\in Q} \Psi(x)$. If one can guarantee that $\nabla\Psi(x_*) = 0$, then (3) can be improved $\widetilde{\delta} \leq R\sqrt{2L\widetilde{\varepsilon}}$.

Let us recall some examples in which the concept of (δ, L)-model of objective function is useful. For the following optimization problems and methods: smooth optimization problem, composite (accelerated) gradient methods [3, 32], level (accelerated) gradient method [23], min-min problem, Proximal method [7], universal method [33] refer to [17]. In section 2.4 we consider (δ, L)–model for universal Frank–Wolfe method. As far as we know, this is the first attempt to combine Frank–Wolfe method [19, 20] and universal method [33].

2.2 Gradient Method with Inexact Model

In this section we consider a simple non-accelerated method for optimization problems with (δ, L)-model. This method is a variant of the standard gradient method [41] with adaptive Lipschitz constant tuning of the gradient of the objective function.

We assume that at each iteration k, the method has access to $(\delta, \bar{L}_{k+1})$-model of f w.r.t $V[y](x)$ (see Definition 1 and Remark 2). In general, we consider that constant $\bar{L}_{k+1}$ may vary from iteration to iteration, we only assume that we can find some constant $\bar{L}_{k+1}$ such that $(\delta, \bar{L}_{k+1})$-model exists at k–step of Algorithm 1 and we do not use $\bar{L}_{k+1}$ in Algorithm 1 explicitly.

Theorem 4. *Let $V[x_0](x_*) \leq R^2$, where x_0 is the starting point, and x_* is the nearest minimum point to the point x_0 in the sense of Bregman divergence (see Remark 1). We assume that $\bar{L}_{k+1} \leq L$ for all $k \geq 0$. Then, for the sequence, generated by Algorithm 1 the following holds*

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

The theorem is proved in [43].

Remark 4. Despite the adaptive structure of Algorithm 1 as in [33] it can be shown that in average the algorithm up to logarithmic terms requires two computations of function and one computation of (δ, L)-model per iteration.

Algorithm 1 Gradient method with an oracle using the (δ, L)-model

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

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

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

$$\phi_{k+1}(x) = \psi_\delta(x, x_k) + L_{k+1}V[x_k](x), \quad x_{k+1} := \operatorname*{argmin}_{x \in Q}{}^{\widetilde{\delta}}\phi_{k+1}(x). \tag{4}$$

5: **end for**
Ensure: $\bar{x}_N = \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}x_{k+1}$

2.3 Fast Gradient Method with Inexact Model

In this section we consider accelerated method for problems with (δ, L)-model. The method is close to accelerated mirror-descent type of methods by [44, 22, 11]. On each iteration, the inexact model is used to make a mirror-descent-type of step. In this section, we assume that the (δ, L)-model of f is given w.r.t. $\|\cdot\|$-norm and $V[u](x)$ satisfies (1-SC) condition w.r.t. this norm (see Remarks 1, 2). As in section 2.2 we assume that we can find some constant $\bar{L}_{k+1}$ such that $(\delta_k, \bar{L}_{k+1})$-model of f w.r.t. $\|\cdot\|$-norm exists at k–step ($k = 0, .., N-1$) of Algorithm 2. Unlike Algorithm 1, we have sequences $\{\widetilde{\delta}_k\}_{k\geq 0}$ and $\{\delta_k\}_{k\geq 0}$ for input instead of constants in Algorithm 2.

Theorem 5. *Let $V[x_0](x_*) \leq R^2$, where x_0 is the starting point and x_* is the nearest minimum point to x_0 in the sense of Bregman divergence. Then, for the sequence, generated by Algorithm 2,*

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

The theorem is proved in [17].

Remark 5. Despite the adaptive structure of Algorithm 2 as in [33] it can be shown that in average the algorithm up to logarithmic terms requires four computations of function and two computations of (δ, L)-model per iteration.

Remark 6. For the case when we know that $\bar{L}_{k+1} \leq L$ for all $k \geq 0$ (or in other words, (δ_k, L)-model exists for all $k \geq 0$), $L_0 \leq L$, $\delta_k = \delta$ and $\widetilde{\delta}_k = \widetilde{\delta}$ for all $k \geq 0$, we can show that $A_N \geq \frac{(N+1)^2}{8L}$ (see [17]) and

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

2.4 Universal conditional gradient (Frank–Wolfe) method

Let us show an example of (δ, L)–model conception. We use Algorithm 2 as a proxy method for universal Frank–Wolfe method. In order to construct universal Frank–Wolfe method let us introduce the following constraints to the optimization problem (1):

Algorithm 2 Fast gradient method with oracle using (δ, L)-model

1: **Input:** x_0 is the starting point, $\{\widetilde{\delta}_k\}_{k\geq 0}$, $\{\delta_k\}_{k\geq 0}$ 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 $i_k \geq 0$ 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

$$A_{k+1} = L_{k+1}\alpha_{k+1}^2, \quad A_{k+1} := A_k + \alpha_{k+1}.$$

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

$$\phi_{k+1}(x) = \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}) + V[u_k](x), \quad u_{k+1} := \operatorname*{argmin}_{x\in Q}{}^{\widetilde{\delta}_k}\phi_{k+1}(x).$$

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

5: **end for**
Ensure: x_N

1 The set Q is bounded w.r.t $V[y](x)$:

$$\exists R_Q \in \mathbb{R} : \ \forall x, y \in Q \, V[y](x) \leq R_Q^2.$$

2 The function $f(x)$ has Holder continues 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 [33]):

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

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 [33] we can get the following inequality for A_N:

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

It it shown in [17] 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 (see also section 3, [17]) 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 2:

$$u_{k+1} = \left(\operatorname*{argmin}^{\widetilde{\delta}_k}_{x \in Q} \phi_{k+1}(x) \overset{\text{def}}{=} \operatorname*{argmin}_{x \in Q} \widetilde{\phi}_{k+1}(x) \right).$$

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 1). As in [17] we can show that an error in sense of Definition 2 would not be greater than $2R_Q^2$. Therefore, we can take $\widetilde{\delta}_k = 2R_Q^2$. From Theorem 5 we can get the following inequality:

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

Using (5) 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 $\nu > 0$.

3 The Concept of (δ, L, μ)-model. The Case of Strongly Convex Objective and Strongly Convex Model

Now we consider the case of a strongly convex objective. The following assumption allows us to prove a linear rate of convergence for non-adaptive and adaptive versions of Algorithm 1.

Definition 6. Let function $\psi_\delta(x, y)$ be convex in $x \in Q$ and satisfy $\psi_\delta(x, x) = 0$ for all $x \in Q$. We say that $\psi_\delta(x, y)$ is a (δ, L, μ)-model of the function f at a given point y with respect to $V[y](x)$ iff, for all $x \in Q$, the inequality

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

Note that we allow L to depend on δ. We refer to this case as strongly convex case.

Remark 7. 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:

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

In this subsection we describe a gradient-type method for problems with (δ, L)-model of the objective. This algorithm is a natural extension of gradient method, see [16, 43, 17].

We consider the case of m-strogly convex models ψ. One example of the m-strong convexity of the function $\psi_\delta(x, y)$ appears in composite optimization:

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

where $g(x)$ is μ-strongly convex and smooth function with L-Lipschitz gradient, $h(x)$ is convex function of simple structure. As a $\psi_\delta(x, y)$ function for composite optimization problem, we take

$$\psi_\delta(x, y) = \langle \nabla g(x), y - x \rangle + h(x) - h(y).$$

Notice that, $\psi_\delta(x, y)$ is strongly convex in y when $h(x)$ is strongly convex. An example of such a problem with strongly convex model is the following minimization problem [32]:

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

Another example of an optimization problem that allows for (δ, L, μ)-model with strong convexity of the function $\psi_\delta(x, y$ arises in Y. Nesterov's electoral model [36] [42]. In this model, voters (data points) select a party (cluster) iteratively by minimizing the following function:

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

Algorithm 3 Gradient method with (δ, L)-model of the objective.

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

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

4: **end for**
Ensure: $\bar{x}_N = \frac{1}{N}\sum_{k=0}^{N-1} x_{k+1}$

Thus, we have the following result

Theorem 7. *Assume for f $\psi_\delta(x, y)$ is a m-strongly convex (δ, L, μ)-model w.r.t. $V[y](x)$. Then, after of k iterations of Algorithm 3, we have:*

$$f(y_{k+1}) - f(x_*) \leq (m + L)(x_*) \exp\left((-k + 1)\frac{\mu + m}{L + m}\right) V[x_0] + \delta + \widetilde{\delta}.,$$

$$V[x_{k+1}](x_*) \leq \frac{\delta + \widetilde{\delta}}{m + \mu} + \left(\frac{L - \mu}{L + m}\right)^{k+1} V[x_0](x_*).$$

In other words, if function satisfies right relative strong convexity and relative smoothness, then after performing $O\left(\log(\frac{1}{\varepsilon})\right)$ iterations we can achieve an accuracy of ε accurate to term $O(\delta + \widetilde{\delta})$.

The proof of Theorem 3.2 is in the Appendix B.

Let us consider some adaptive version of Algorithm 3, which is applicable to possibly unknown constant L.

Algorithm 4 Adaptive gradient method with an oracle using the (δ, L, μ)-model

1: **Input:** x_0 is the starting point, $\mu > 0$ $L_0 \geq 2\mu$ and δ.
2: Set $S_0 := 0$
3: **for** $k \geq 0$ **do**
4: Find the smallest $i_k \geq 0$ 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 $L_k \geq 2\mu$ and $L_{k+1} = 2^{i_k}L_k$ for $L_k < 2\mu$,
$\alpha_{k+1} := \frac{1}{L_{k+1}}$, $S_{k+1} := S_k + \alpha_{k+1}$.

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

5: **end for**
Ensure: $\bar{x}_N = \frac{1}{S_N}\sum_{k=0}^{N-1} \frac{x_{k+1}}{L_{k+1}}$

Remark 8. The advantage of Algorithm 4 is that there is no need to know the true value of the Lipschitz constant L. However, this may increase the cost of the iteration due to repeating steps of type (4). At the same time the procedure of choosing L_{k+1} in Algorithm 1 allow us to show that the number of steps of type (4) is less than $2N + \log_2 \frac{2L}{L_0}$.

To obtain the rate of convergence of Algorithm 4 we need to introduce the averaging parameter $\hat{L}$:

$$1 - \frac{\mu}{\hat{L}} = \sqrt[k+1]{\left(1 - \frac{\mu}{L_{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right)\ldots\left(1 - \frac{\mu}{L_1}\right)}. \tag{7}$$

Assume m-strong convexity of the function $\psi_\delta(x, y)$. The following result holds:

Theorem 8. *Assume that m-strongly convex by x functional $\psi_\delta(x, y)$ is a (δ, L, μ)-model w.r.t. $V[y](x)$ for f. Then, after of k iterations of Algorithm 4, we have*

$$f(x_{k+1}) - f(x_*) \leqslant \frac{(2L+m)^2}{(\mu+m)^2}\left(1 - \left(1 - \frac{\mu+m}{2L+m}\right)^{k+1}\right)(2\delta + \widetilde{\delta}) + \\ + (2L+m)\left(1 - \frac{\mu+m}{m+\hat{L}}\right)^{k+1} V[x_0](x_*).$$

For convex ψ_δ $(m = 0)$ we have the following inequalities:

$$V[x_{k+1}](x_*) \leq \frac{2L(2\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + \left(1 - \frac{\mu}{\hat{L}}\right)^{k+1} V[x_0](x_*),$$

$$f(x_{k+1}) - f(x_*) \leqslant \frac{4L^2(2\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + 2L\left(1 - \frac{\mu}{\hat{L}}\right)^{k+1} V[x_0](x_*).$$

To prove Theorem 8 we need the following lemma:

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

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

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 2:

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

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

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

complete the proof. □

Now we ready to prove the Theorem 8.

Proof. After k iterations of Algorithm 4 using lemma 1, we have

$$(L_{k+1}+m)V[x_{k+1}](x) \leq \widetilde{\delta}+\psi_\delta(x, x_k)-\psi_\delta(x_{k+1}, x_k)+L_{k+1}V[x_k](x)-L_{k+1}V[x_k](x_{k+1}). \quad (8)$$

Considering (6) w.r.t. $V[x](y)$ we obtain:

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

or

$$-L_{k+1}V[x_k](x_{k+1}) \leq 2\delta - f(x_{k+1}) + f_\delta(x_k) + \psi_\delta(x_{k+1}, x_k),$$

Now (8) means

$$(L_{k+1} + m)V[x_{k+1}](x) \leq \widetilde{\delta} + 2\delta - f(x_{k+1}) + f_\delta(x_k) + \psi_\delta(x, x_k) + (L_{k+1} - \mu)V[x_k](x).$$

Let $x = x_*$.Therefore we obtain

$$\begin{gathered}V[x_{k+1}](x_*) \leq V[x_{k+1}](x) + \frac{f(x_{k+1} - f(x_*)}{L_{k+1} + m} \leq \frac{2\delta + \widetilde{\delta}}{L_{k+1} + m} + \\ + \frac{L_{k+1} - \mu}{L_{k+1} + m}V[x^k](x_*) \leq (2\delta + \widetilde{\delta})\left(\frac{1}{L_{k+1} + m} + \frac{1}{L_k}\frac{L_{k+1} - \mu}{L_{k+1} + m}\right) + \\ + \frac{L_{k+1} - \mu}{L_{k+1} + m}V[x_{k-1}](x_*) \leq \ldots \leq (2\delta + \widetilde{\delta})\left(\frac{1}{L_{k+1} + m} + \frac{1}{L_k + m}\frac{L_{k+1} - \mu}{L_{k+1} + m} + \ldots + \right. \\ \left. + \frac{1}{L_1}\frac{L_2 - \mu}{L_2 + m}\ldots\frac{L_{k+1} - \mu}{L_{k+1} + m}\right) + \frac{L_{k+1} - \mu}{L_{k+1} + m}\frac{L_k - \mu}{L_k + m}\ldots\frac{L_1 - \mu}{L_1 + m}V[x_0](x_*).\end{gathered}$$

Because $L_0 \leq 2L$, then $L_{k+1} \leq 2L$. In Algorithm 1 we also assume that $L_{k+1} \geq \mu$. So we have:

$$\frac{1}{2L} \leq \frac{1}{L_{k+1}} \leq \frac{1}{\mu}.$$

Then we obtain

$$\frac{L_{k+1}-\mu}{L_{k+1}+m} = 1 - \frac{m+\mu}{L_{k+1}+m} \leq 1 - \frac{m+\mu}{2L+m}.$$

Using the averaging parameter $\hat{L}$

$$1 - \frac{m+\mu}{\hat{L}+m} = \sqrt[k]{\frac{1}{L_1}\frac{L_2-\mu}{L_2+m}\cdots\frac{L_k-\mu}{L_k+m}}$$

we have

$$V[x_{k+1}](x_*) \leq \left(1 - \frac{m+\mu}{m+\hat{L}}\right)^{k+1} V[x_0](x_*) + (2\delta+\widetilde{\delta})\frac{1}{\mu+m}\sum_{i=0}^{k}\left(1-\frac{m+\mu}{2L+m}\right)^i =$$
$$= (1-\frac{m+\mu}{m+\hat{L}})^{k+1}V[x_0](x_*) + \frac{2L+m}{(m+\mu)^2}\left[1-\left(1-\frac{\mu+m}{(2L+m)^2}\right)\right](2\delta+\widetilde{\delta}).$$

So we finally obtain

$$f(x_{k+1}) - f(x_*) \leqslant \frac{(2L+m)^2}{(\mu+m)^2}\left(1-\left(1-\frac{\mu+m}{2L+m}\right)^{k+1}\right)(2\delta+\widetilde{\delta}) +$$
$$+ (2L+m)\left(1-\frac{\mu+m}{m+\hat{L}}\right)^{k+1} V[x_0](x_*).$$

□

Let us consider the case of a strongly convex functional f and show how to accelerate the work of Algorithms 1 and 4 using the restart technique. Let us assume that

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

Note the 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 relative μ-strongly convexity in the following way

Definition 9. Say that the function f is a left relative μ-strongly convex if the following inequality

$$\mu V[x](y) \leq f(x) - f(y) - \psi_\delta(x,y).$$

holds.

Note that concepts of right and left relative strongly convexity from Definitions 6 and 9 are equivalent in the case of assumption from Remark 7 ($V[x](y) \leq C_n\|x-y\|^2$ for each $x, y \in Q$).

Theorem 10. *Let f be a left relative μ-strongly convex function and $\psi_\delta(x,y)$ is a (δ, L)-model w.r.t. $V[y](x)$. Then, using the restarts of Algorithm 1, we obtain the estimate*

$$V[\bar{x}_{N_p}](x_*) \leq \varepsilon + \frac{2\tilde{\delta}}{\mu} + \frac{4\delta}{\mu}$$

for a given $\varepsilon > 0$. The total number for iterations of Algorithm 1 not exceeding

$$N = \left\lceil \log_2 \frac{R^2}{\varepsilon} \right\rceil \cdot \left\lceil \frac{4L}{\mu} \right\rceil. \tag{9}$$

Proof. By Definition 9 and Theorem 4 we have

$$\mu V[\bar{x}_{N_1}](x_*) \leq f(\bar{x}_{N_1}) - f(x_*) \leq \frac{2LV[x_0](x_*)}{N_1} + \widetilde{\delta} + 2\delta.$$

Further, due to the following inequality

$$V[\bar{x}_{N_1}](x_*) \leq \frac{2LV[x_0](x_*)}{\mu N_1} + \frac{\widetilde{\delta}}{\mu} + \frac{2\delta}{\mu} \tag{10}$$

let's choose the smallest number of steps N_1:

$$V[\bar{x}_{N_1}](x_*) \leq \frac{1}{2}V[x_0](x_*) + \frac{\widetilde{\delta}}{\mu} + \frac{2\delta}{\mu}.$$

Similarly, after the 2nd restart (N_2 operations)

$$V[\bar{x}_{N_2}](x_*) \leq \frac{1}{2}V[\bar{x}_{N_1}](x_*) + \frac{\widetilde{\delta}}{\mu} + \frac{2\delta}{\mu} \leq \frac{1}{4}V[x_0](x_*) + \left(\frac{\widetilde{\delta}}{\mu} + \frac{2\delta}{\mu}\right)\left(1 + \frac{1}{2}\right).$$

After the p-th restart (N_p operations)

$$V[\bar{x}_{N_p}](x_*) \leq \frac{1}{2^p}V[x_0](x_*) + \left(\frac{\widetilde{\delta}}{\mu} + \frac{2\delta}{\mu}\right)\left(1 + \frac{1}{2} + ... + \frac{1}{2^{p-1}}\right) <$$

$$< \frac{1}{2^p}V[x_0](x_*) + \frac{2\widetilde{\delta}}{\mu} + \frac{4\delta}{\mu}.$$

Choose p such that

$$\frac{1}{2^p}V[x_0](x_*) \leq \varepsilon$$

After $p = \left\lceil \log_2 \frac{R^2}{\varepsilon} \right\rceil$ restarts we have

$$V[\bar{x}_{N_p}](x_*) \leq \varepsilon + \frac{2\widetilde{\delta}}{\mu} + \frac{4\delta}{\mu}.$$

The number of iterations N_k $(k = \overline{1,p})$ on the k-th restart of Algorithm 1 is estimated from (10):

$$\frac{2L}{\mu N_k} \leq \frac{1}{2}, \;\; N_k \geq \frac{4L}{\mu}.$$

So, we can put $N_k = \left\lceil \frac{4L}{\mu} \right\rceil$ and (9) holds. □

We show that using the restart technique can also accelerate the work of non-adaptive version of Algorithm 4 ($L_{k+1} = L$) for (δ, L)-model $\psi_\delta(x, y)$ w.r.t. norm $\|\cdot\|$ and relative μ-strogly convex function f in sense Definition 9:

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

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

$$f(x_N) - f(x_*) \leqslant \frac{8LV[x_0](x_*)}{(N+1)^2} + \frac{8L\widetilde{\delta}}{N+1} + 2N\delta. \tag{11}$$

Consider the case of relatively μ-strongly convex function f. We will use the restart technique to obtain the method for strongly convex functions. By (11) and Definition 9:

$$\mu V[x_{N_1}](x_*) \leq f(x_{N_1}) - f(x_*) \leq \frac{8LV[x_0](x_*)}{N^2} + \frac{8L\widetilde{\delta}}{N} + 2N\delta. \tag{12}$$

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

$$\frac{8L\widetilde{\delta}}{N_1} + 2N_1\delta \leq \frac{LV[x_0](x_*)}{N_1^2}. \tag{13}$$

We restart method as

$$V[x_{N_1}](x_*) \leq \frac{V[x_0](x_*)}{2}.$$

From (12):

$$\frac{9L}{\mu N_1^2} \leq \frac{1}{2}, \quad N_1 \geq 3\sqrt{2\frac{L}{\mu}}$$

Let's choose

$$N_1 = \left\lceil 3\sqrt{\frac{2L}{\mu}} \right\rceil. \tag{14}$$

Then after N_1 iterations we restart method. Similarly, we restart after N_2 iterations, such that $V[x_{N_2}](x_*) \leq \frac{V[x_{N_1}](x_*)}{2}$. We obtain

$$N_2 = \left\lceil 3\sqrt{\frac{2L}{\mu}} \right\rceil.$$

So, after p-th restart the total number of iterations:

$$M = p \cdot \left\lceil 3\sqrt{\frac{2L}{\mu}} \right\rceil.$$

Now let's consider how many iterations is needed to achieve accuracy $\varepsilon = f(x_{N_p}) - f(x_*)$. From (11) and (14) we take

$$p = \left\lceil \log_2 \frac{\mu R^2}{\varepsilon} \right\rceil$$

and total number of iterations:

$$M = \left\lceil \log_2 \frac{\mu R^2}{\varepsilon} \right\rceil \cdot \left\lceil 3\sqrt{\frac{2L}{\mu}} \right\rceil.$$

We need to choose our errors as $\delta = O(\frac{\varepsilon L}{\mu N_k^3})$ and $\widetilde{\delta} = O(\frac{\varepsilon}{\mu N_k})$ to satisfy (13). Indeed, from (13) using $N_k = \left\lceil 3\sqrt{\frac{2L}{\mu}} \right\rceil$ we can deduce the following inequality:

$$\varepsilon \geq \frac{12\mu}{L}\left(\frac{9}{\sqrt{2}}\delta\left\lceil\sqrt{\frac{L}{\mu}}\right\rceil^3 + \sqrt{2}\widetilde{\delta}\left\lceil\sqrt{\frac{L}{\mu}}\right\rceil\right).$$

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

4 Inexact Model for Variational Inequalities

In this section, we go beyond minimization problems and propose an abstract inexact model counterpart for variational inequalities. Using this model we propose a new 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. According to the lower bounds in [39], this algorithm is optimal for $\nu = 0$ and $\nu = 1$. Based on the model for VI and functions, we extend (δ, L)-model for saddle-point problems (see Definition 16). We are also motivated by mixed variational inequalities [21, 1] and composite saddle-point problems [6].

We consider the problem of finding the solution $x_* \in Q$ for VI in the following abstract form

$$\psi(x, x_*) \geqslant 0 \quad \forall x \in Q \tag{15}$$

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

$$\psi(x, y) + \psi(y, x) \leqslant 0 \quad \forall x, y \in Q, \tag{16}$$

any solution (15) will is a solution of the following inequality

$$\max_{x\in Q} \psi(x_*, x) \leqslant 0 \quad \forall x \in Q. \tag{17}$$

In the general case, we make an assumption about the existence of a solution x_* of the problem (15). 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 (15) and (17) are equivalent, respectively, to a standard strong and weak variational inequality with the operator g.

Example 11. For some operator $g : Q \to \mathbb{R}^n$ and a convex functional $h : Q \to \mathbb{R}^n$ choice

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

leads to a *mixed variational inequality from* [21, 1]

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

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

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

We propose an adaptive proximal method for the problems (15) and (17). We start with a concept of (δ, L)-model for such problems.

Definition 12. We say that functional ψ has (δ, L)-model $\psi_\delta(x, y)$ for some fixed values $L > 0$ at $\delta > 0$ at a given point y w.r.t. $V[y](x)$ if the following properties hold for each $x, y, z \in Q$:

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

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

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

(iv) (*abstract δ-monotonicity*)
$$\psi_\delta(x, y) + \psi_\delta(y, x) \leq \delta; \tag{18}$$

(v) (*generalized relative smoothness*)
$$\psi_\delta(x, y) \leqslant \psi_\delta(x, z) + \psi_\delta(z, y) + LV[z](x) + LV[y](z) + \delta. \tag{19}$$

Remark 9. Similarly to Definition 1 above, in general case, we do not need the (1-SC) assumption in Definition 12 for $V[y](x)$. In some situations we assume that (1-SC) assumption holds (see Examples 13, 14 and Section 5).

Remark 10. In Definition 12 we change 'w.r.t $V[y](x)$' to 'w.r.t. $\|\cdot\|$-norm if we use $\frac{1}{2}\|x-y\|^2$ instead of $V[y](x)$.

Note that for $\delta = 0$ the following analogue of (28) for some fixed $a, b > 0$
$$\psi(x, y) \leqslant \psi(x, z) + \psi(z, y) + a\|z - y\|^2 + b\|x - z\|^2 \quad \forall x, y, z \in Q \tag{20}$$
was introduced in [26]. Condition (20) 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 [33] (see Example 14 below).

One can directly verify that if $\psi_\delta(x, y)$ is $(\delta/3, L)$-model of the function f at a given point y w.r.t. $V[y](x)$ then $\psi_\delta(x, y)$ is (δ, L)-model in the sense of Definition 12 w.r.t. $V[y](x)$.

Let us consider some examples.

Example 13. Variational Inequalities with monotone Lipshitz continuous operator. Consider variational inequality of finding $x \in Q$ such that $\langle g(y), x - y\rangle \leq 0, \forall y \in Q$, the operator $g : Q \to R^n$ is monotone and Lipschitz continuous, i.e. $\|g(x) - g(y)\|_* \leq L\|x - y\|, \ \forall x, y \in Q$. In this case $\psi_\delta(x, y) := \langle g(y), x - y\rangle$ is a (δ, L)-model in a sense of Definition 12 w.r.t. $\|\cdot\|$-norm ($\forall x, y \in Q$).

Example 14. Variational Inequalities with monotone Holder continuous operator. Assume that for 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: $\langle g(z) - g(y), z - x\rangle \leq \|g(z) - g(y)\|_*\|z - x\| \leq$
$$\leq L_\nu\|z - y\|^\nu\|z - x\| \leq \frac{L(\delta)}{2}||z - x||^2 + \frac{L(\delta)}{2}||z - y||^2 + \delta \tag{21}$$

for

$$L(\delta) = \left(\frac{1}{2\delta}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}} \tag{22}$$

and uncontrolled parameter $\delta > 0$. In this case the following function

$$\psi_\delta(x, y) := \langle g(y), x - y\rangle \ \ \forall x, y \in Q.$$

is (δ, L)-model w.r.t. $\|\cdot\|$-norm.

Note that for the previous two examples in Algorithm 5 and Theorem 15 we need $V[z](x)$ to satisfy (1-SC) condition.

Next, we introduce our novel adaptive method (Algorithm 5) for abstract variational inequalities with inexact (δ, L)-model w.r.t. $V[y](x)$. If $V[y](x)$ satisfies (1-SC) condition then we can consider inexact (δ, L)-model w.r.t. $\|\cdot\|$-norm. This method adapts to the local values of L and similarly to [33] allows us to construct universal method for variational inequalities. Applying the following adaptive Algorithm 5 to VI with Holder interpolation (21) for $\delta = \frac{\varepsilon}{2}$ and $L = L\left(\frac{\varepsilon}{2}\right)$ leads us to universal method for VI.

Algorithm 5 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: **for** $k = 0, 1, ...$ **do**

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

$$\psi_\delta(z_{k+1}, z_k) \leq \psi_\delta(z_{k+1}, w_k) + \psi_\delta(w_k, z_k) + L_{k+1}(V[z_k](w_k) + V[w_k](z_{k+1})) + \delta, \tag{23}$$

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

$$w_k = \operatorname*{argmin}_{x\in Q} \{\psi_\delta(x, z_k) + L_{k+1}V[z_k](x)\}.$$
$$z_{k+1} = \operatorname*{argmin}_{x\in Q} \{\psi_\delta(x, w_k) + L_{k+1}V[z_k](x)\}.$$

4: **end for**

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

For a given accuracy ε we can consider the following stopping criterion for Algorithm 5:

$$S_N := \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \geqslant \frac{V[x^0](x_*)}{\varepsilon}.$$

Let us formulate the following result. Note that we consider the case of m-strong convexity of model $\psi_\delta(x, y)$. Clearly, the case $m = 0$ means convexity of $\psi_\delta(x, y) \in X$. Note that for $m > 0$ we can prove that our method converges by argument.

Theorem 15. *Assume that $\psi_\delta(x, y)$ is a m-strongly convex function by x for some $m > 0$. Then for Algorithm 5 the following inequalities hold:*

$$\frac{m}{2}\|\hat{w}_N - x\|^2 - \frac{1}{S_N}\sum_{k=0}^{N-1} \frac{\psi_\delta(x, w_k)}{L_{k+1}} \leq \frac{V[z_0](x)}{S_N} + \delta + 2\widetilde{\delta}.$$

It means that:

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

and for exact solution x_* of considered problem

$$\|\hat{w}_N-x_*\|^2\leq\frac{4LV[z_0](x_*)}{mN}+\frac{4\delta+4\widetilde{\delta}}{m}.$$

Note that the method works no more than

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

iterations.

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

$$m\|u-w_k\|^2+\psi_\delta(w_k,z_k)\leqslant\psi_\delta(u,z_k)+L_{k+1}V[z_k](u)-L_{k+1}V[w_k](u)-L_{k+1}V[z_k](w_k)+\widetilde{\delta}$$

and $m\|u-z_{k+1}\|^2+\psi_\delta(z_{k+1},w_k)\leq$

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

The first inequality means that

$$m\|z_{k+1}-w_k\|^2+\psi_\delta(w_k,z_k)\leqslant\psi_\delta(z_{k+1},z_k)+L_{k+1}V[z_k](z_{k+1})-L_{k+1}V[w_k](z_{k+1})-L_{k+1}V[z_k](w_k)+\widetilde{\delta}.$$

Taking into account (23) and obvious inequality $2(a^2+b^2)\geq(a+b)^2$, we obtain for all $u\in Q$

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

So, the following inequality

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

holds. By virtue of (28) and the choice of $L_0\leqslant 2L$, it is guaranteed that

$$L_{k+1}\leqslant 2L\ \ \forall k=\overline{0,N-1}.$$

and for $u=x_*$ from convexity of function $\varphi(y)=\|y-x_*\|^2$ we have

$$\frac{m}{2}\|\hat{w}_N-u\|^2+\max_{u\in Q}\psi_(\widehat{w}_N,u)\leqslant\frac{m}{2}\|\hat{w}_N-x_*\|^2+\max_{u\in Q}\psi_\delta(\widehat{w}_N,u)+\delta\leqslant$$

$$\leqslant\frac{m}{2}\|\hat{w}_N-x_*\|^2-\frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\psi_\delta(u,w_k)}{L_{k+1}}+2\delta\leqslant$$

$$\leqslant\frac{2L\max_{u\in Q}V[z_0](u)}{N}+3\delta+2\widetilde{\delta}.$$

□

Remark 11. To obtain precision $\varepsilon + 3\delta$ Algorithm 5 works no more than

$$\left\lceil \frac{2L \max_{u \in Q} V[z_0](u)}{\varepsilon} \right\rceil \tag{24}$$

iterations. Note that estimate (24) is optimal for variational inequalities and saddle-point problems [39].

For universal method to obtain precision ε we can choose $\delta = \frac{\varepsilon}{2}$ and $L = L\left(\frac{\varepsilon}{2}\right)$ according to (21) and (22) and the estimate (24) reduces to

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

Thus, the introduced concept of the (δ, L)-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* [21, 1] and *composite saddle-point problems* [6].

Now we extend (δ, L)-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 such that:

$$f(u_*, v) \leqslant f(u_*, v_*) \leqslant f(u, v_*) \tag{25}$$

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 (25) 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)-model for abstract variational inequality (w.r.t. $V[y](x)$ or $\|\cdot\|$).

Definition 16. 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)-model w.r.t. $V[y](x)$ for the saddle-point problem (25) if the following properties hold for each $x, y, z \in Q$:

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

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

(iii) (*abstract δ-monotonicity*)

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

(iv) (*generalized relative smoothness*)

$$\psi_\delta(x, y) \leqslant \psi_\delta(x, z) + \psi_\delta(z, y) + LV[z](x) + LV[y](z) + \delta \tag{26}$$

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

(v)

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

Example 17. The proposed concept of the (δ, L)-model for saddle-point problems is quite applicable, for example, for composite saddle point problems of the form considered in the popular article [6]:

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

for some convex in u and concave in v subdifferentiable functions $\tilde{f}$, as well as convex functions h and φ. In this case, you 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 15 implies

Theorem 18. *If for the saddle problem* (25) *there is a* (δ, L)*-model* $\psi_\delta(x,y)$ *w.r.t.* $V[y](x)$*, 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 inequality is true:

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

If $\psi_\delta(x,y)$ is a m-strongly convex in x for $m > 0$ then for exact solution $x_* = (u_*, v_*)$ we have:

$$\|(\widehat{u}_N, \widehat{v}_N) - (u_*, v_*)\|^2 \leqslant \frac{4L\max_{(u,v)\in Q} V[u_0,v_0](u,v)}{mN} + \frac{4\widetilde{\delta} + 4\delta}{m}.$$

Remark 12. The property of m-strong convexity for $\psi_\delta \in X$ is true for composite saddle point with m-strongly convex h.

5 Modelling for Strongly Monotone VI

In this section similarly with the concept of (δ, L, μ)-model in optimization we consider inexact model for VI with more strong version of (18).

Definition 19. We say that functional ψ has (δ, L, μ)-model $\psi_\delta(x,y)$ at a given point y w.r.t. $V[y](x)$ if the following properties hold for each $x, y, z \in Q$:

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

(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*)

$$\psi_\delta(x,y) + \psi_\delta(y,x) + \mu\|x-y\|^2 \leq \delta; \tag{27}$$

(v) (*generalized relative smoothness*)

$$\psi_\delta(x,y) \leqslant \psi_\delta(x,z) + \psi_\delta(z,y) + LV[z](x) + LV[y](z) + \delta \tag{28}$$

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

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

$$d(x) \leq \frac{\Omega}{2}, \quad \forall x \in Q : \|x\| \leq 1, \tag{29}$$

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 5 restating is applicable for abstract strongly monotone variational inequalities.

Algorithm 6 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 5 after N_p iterations of Algorithm 5 with accuracy $\mu\varepsilon/2$, 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 $R_{p+1}^2 = R_0^2 \cdot 2^{-(p+1)} + \frac{(1-2^{-p})\varepsilon}{2}$.
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** $p > \log_2 \frac{2R_0^2}{\varepsilon}$

Ensure: x_{p+1}.

Theorem 20. *Assume that ψ_δ is a (δ, L, μ)-model for ψ. Also assume that the prox function $d(x)$ satisfies* (29) *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* (17)*. Then, for $p \geq 0$*

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

and the point x_p returned by natural analogue of Algorithm 6 with restarts of Algorithm 5 satisfies $\|x_p - x_*\|^2 \leq \varepsilon$. The total number of iterations of the inner Algorithm 5 does not exceed

$$\left\lceil \frac{2L\Omega}{\mu} \cdot \log_2 \frac{2R_0^2}{\varepsilon} \right\rceil, \tag{30}$$

where Ω is satisfied to (29).

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

$$\|x_p - x_*\|^2 \leq R_0^2 \cdot 2^{-p} + \left(1 - 2^{-p}\right)\left(\frac{\varepsilon}{2} + \frac{2\delta}{\mu} + \frac{2\widetilde{\delta}}{\mu}\right),$$

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 6 is 1-strongly convex w.r.t. the norm $\|\cdot\|/R_p$.

This means that, at each step k of inner Algorithm 5, L_{N_p} changes to $L_{N_p} \cdot R_p^2$. Using the definition of $d_p(\cdot)$ and (29), 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 15, taking $u = x_*$, we obtain

$$-\frac{1}{S_{N_p}} \sum_{k=0}^{N_p-1} \frac{\psi_\delta(x_*, w_k)}{L_{k+1}} + 2\delta \leq \frac{R_p^2 V_p[x_p](x_*)}{S_{N_p}} + \frac{\mu\varepsilon}{4} \leq \frac{\Omega R_p^2}{2S_{N_p}} + \frac{\mu\varepsilon}{4} + 2\widetilde{\delta}.$$

Since the operator ψ is continuous and abstract monotone, we can assume that the solution to weak VI (15) is also a strong solution and

$$-\psi(w_k, x_*) \leq 0, \quad k = 0, ..., N_p - 1$$

and by Definition 19 (i)

$$-\psi_\delta(\omega_k, x_*) \leq \delta, \quad k = 0, \ldots, N_p - 1$$

This and (27) gives, that for each $k = 0, ..., N_p - 1$,

$$-\psi_\delta(x_*, w_k) \geq \delta - \psi_\delta(x_*, w_k) - \psi_\delta(w_k, x_*) \geq \mu\|w_k - x_*\|^2.$$
$$-\psi_\delta(x_*, \omega_k) \geq \delta - \psi_\delta(x_*, \omega_k) - \psi_\delta(\omega_k, x_*) \geq \mu\|\omega_k - x_*\|^2$$

Thus, by convexity of the squared norm, we obtain

$$\mu\|x_{p+1} - x_*\|^2 = \mu \left\| \frac{1}{S_{N_p}} \sum_{k=0}^{N_p-1} \frac{w_k}{L_{k+1}} - x_* \right\|^2 \leq \frac{\mu}{S_{N_p}} \frac{1}{L_{k+1}} \sum_{k=0}^{N_p-1} \|w_k - x_*\|^2$$
$$\leq -\frac{1}{S_{N_p}} \sum_{k=0}^{N_p-1} \frac{\psi_\delta(x_*, w_k)}{L_{k+1}} \leq \frac{\Omega R_p^2}{2S_{N_p}} + \frac{\mu\varepsilon}{4} + 2\delta + 2\widetilde{\delta}.$$

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

$$\|x_{p+1} - x_*\|^2 \leq \frac{R_p^2}{2} + \frac{\varepsilon}{4} + \frac{2\delta + 2\widetilde{\delta}}{\mu} = \frac{1}{2}\left(R_0^2 \cdot 2^{-p} + \frac{(1 - 2^{-p})\varepsilon}{2}\right) + \frac{\varepsilon}{4} + \frac{2\delta + 2\widetilde{\delta}}{\mu} =$$
$$= R_0^2 \cdot 2^{-(p+1)} + (1 - 2^{-p})\varepsilon\left(\frac{\varepsilon}{2} + \frac{2\delta + 2\widetilde{\delta}}{\mu}\right),$$

which finishes the induction proof. □

Remark 13. If for some $m > 0$ $\psi_\delta(x, y)$ is a m-strong convex functional in x then (31) can be exchanged by

$$\left\lceil \frac{2L\Omega}{m+\mu} \cdot \log_2 \frac{2R_0^2}{\varepsilon} \right\rceil. \tag{31}$$

References

[1] T. Q. Bao and P. Q. Khanh. Some algorithms for solving mixed variational inequalities. *Acta Mathematica Vietnamica*, 31(1):77–98, 2006.

[2] 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(2):330–348, 2016.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

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

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

[6] 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(1):120–145, 2011.

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

[8] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

[9] 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.

[10] P. Dvurechensky, A. Gasnikov, and D. Kamzolov. Universal intermediate gradient method for convex problems with inexact oracle. *arXiv:1712.06036*, 2017.

[11] P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376, 2018. arXiv:1802.04367.

[12] P. Dvurechensky, A. Gasnikov, F. Stonyakin, and A. Titov. Adaptive proximal method for variational inequalities. *arXiv preprint arXiv:1804.02579*, 2018.

[13] 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.

[14] 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.

[15] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

[16] A. Gasnikov. Universal gradient descent. *arXiv preprint arXiv:1711.00394*, 2017.

[17] 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(7):1085–1097, 2019.

[18] 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(5):836–841, May 2019.

[19] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *// Mathematical Programming*, 152(1-2):75–112, 2015.

[20] M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[21] I. Konnov and R. Salahutdin. Two-level iterative method for non-stationary mixed variational inequalities. *Izvestija vysshih uchebnyh zavedenij. Matematika*, 61(10):50–61, 2017.

[22] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, Jun 2012. Firs appeared in June 2008.

[23] G. Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. *Mathematical Programming*, 149(1-2):1–45, 2015.

[24] H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[25] J. Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791, 2013.

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

[27] 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(1):229–251, 2004.

[28] A. Nemirovski, S. Onn, and U. G. Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

[29] A. Nemirovskii and Y. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21 – 30, 1985.

[30] Y. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

[31] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, Aug 2009. First appeared in 2005 as CORE discussion paper 2005/67.

[32] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013. First appeared in 2007 as CORE discussion paper 2007/76.

[33] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

[34] Y. Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Math. Program.*, 171(1-2):311–330, 2018.

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

[36] Y. Nesterov. Soft clustering by convex electoral model. CORE Discussion Papers 2018001, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Jan. 2018.

[37] Y. Nesterov and B. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

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

[39] 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.

[40] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[41] B. Polyak. *Introduction to Optimization*. New York, Optimization Software, 1987.

[42] F. S. Stonyakin, M. S. Alkousa, A. A. Titov, and V. V. Piskunova. On some methods for strongly convex optimization problems with one functional constraint. In *MOTOR*, 2019.

[43] 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 M. Khachay, Y. Kochetov, and P. Pardalos, editors, *Mathematical Optimization Theory and Operations Research*, pages 97–114, Cham, 2019. Springer International Publishing. arXiv:1902.09001.

[44] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, MIT, 2008.

A Model examples

In this section we present more examples of a (δ, L)-model of objective f.

Example 21. Convex optimization problem with Lipschitz continuous gradient, [30]

If convex function f has Lipschitz continuous gradient:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\|, \ \ \forall x, y \in Q.$$

then

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

In this case

$$\psi_\delta(x, y) := \langle \nabla f(y), x - y \rangle \ \ \forall x, y \in Q$$

is $(0, L)$-model of f with $f_\delta(y) = f(y)$ at a given point y w.r.t. $\|\cdot\|$-norm.

Example 22. Composite optimization, [3, 32]

Let us consider composite convex optimization problem:

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

where g is a smooth convex function and the gradient of g is Lipschitz continuous with parameter L. Function h is a simple convex function. One can show

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

Therefore

$$\psi_\delta(x, y) = \langle \nabla g(y), x - y \rangle + h(x) - h(y),$$

is $(0, L)$-model of f with $f_\delta(y) = f(y)$ at a given point y w.r.t. $\|\cdot\|$-norm.

Example 23. Superposition of functions, [29]

Let us consider the following optimization problem [23]:

$$f(x) := g(g_1(x), \ldots, g_m(x)) \to \min_{x \in Q}$$

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. From these assumptions we have ([5, 23]) that function $f(x)$ is also convex function and the following inequality holds (see [23]):

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

Also

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

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 w.r.t. $\|\cdot\|$-norm.

Example 24. Proximal method, [7]

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 w.r.t $V[y](x)$, see Definition 1 and Remark 1. Gradient method (see[1] Algorithm 1) with the proposed model is equivalent to the proximal method with general Bregman divergence instead of Euclidean one [40]. In particular, based on this model (with Bregman divergence to be Kullback–Leibler divergence) and Algorithm 1 we propose in [43] proximal Sinkhorn's algorithm for Wasserstein distance calculation problem.

Example 25. Min-min problem

Consider optimization problem:

$$f(x) := \min_{z \in Q} F(z, x) \to \min_{x \in \mathbb{R}^n}.$$

Set Q is convex and bounded. Function F is smooth and convex w.r.t. all variables. Moreover,

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

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 $(6\delta, 2L)$-model of f with $f_\delta(y) = F(\widetilde{z}_\delta(y), y) - 2\delta$ at a given point y w.r.t 2-norm.

Example 26. Saddle point problem, [8]

Let us consider

$$f(x) = \max_{z \in Q} [\langle x, b - Az \rangle - \phi(z)] \to \min_{x \in \mathbb{R}^n},$$

where $\phi(z)$ is a μ-strong convex function w.r.t. p-norm ($1 \leq p \leq 2$). Then f is a smooth convex function and the gradient of f is Lipschitz continuous with parameter

$$L = \frac{1}{\mu} \max_{\|z\|_p \leq 1} \|Az\|_2^2.$$

If $z_\delta(y) \in Q$ is a solution of auxiliary max-problem in the following sense

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

then

$$\psi_\delta(x, y) = \langle b - Az_\delta(y), x - y \rangle$$

is $(\delta, 2L)$-model of f with

$$f_\delta(y) = \langle y, b - Az_\delta(y) \rangle - \phi(z_\delta(y))$$

at the point y w.r.t 2-norm.

[1] To say more precisely if we deal with proximal model (see also Remark **??** and Examples **??**, **??**) it is worth to use non adaptive algorithm, with fixed L.

Example 27. Augmented Lagrangians, [8]

Let us consider

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

and it's dual problem

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

If $z_\delta(y)$ is a solution of auxiliary max-problem in the following sense

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

then

$$\psi_\delta(x, y) = \langle b - Az_\delta(y), x - y\rangle$$

is (δ, μ^{-1})-model of f with

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

at the point y w.r.t 2-norm.

Example 28. Moreau envelope of target function, [8]

Let us consider optimization problem:

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

Assume that function f is a convex function and

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

Then

$$\psi_\delta(x, y) = \langle L(y - z_L(y)), x - y\rangle$$

is (δ, L)-model of f with

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

at the point y w.r.t 2-norm.

B Analysis of Algorithm 3 in the case of (δ, L, μ)-model

Theorem 29. *Let $\psi_\delta(x, y)$ be a (δ, L, μ)-model for f w.r.t. $V[y](x)$ and $y_k = \mathrm{argmin}_{i=1,\dots,k}(f(x_i))$. Then, after k iterations of Algorithm 3, we have*

$$V[x_{k+1}](x_*) \leq \frac{\delta + \widetilde{\delta}}{\mu} + \left(1 - \frac{\mu}{L}\right)^{k+1} V[x_0](x_*).$$

and

$$f(y_{k+1}) - f(x_*) \leq (m + L) \exp\left(-\frac{m+\mu}{m+L}(k+1)\right) V(x_*, x_0) + \delta + \widetilde{\delta}.$$

Proof. Clearly, $f(x_*) \leq f(x_{k+1})$ and

$$(L+m)V[x_{k+1}](x_*) \leq \widetilde{\delta} + \delta + (L-\mu)V[x_k](x_*),$$

i.e.

$$V[x_{k+1}](x_*) \leq \frac{1}{L+m}(\delta + \widetilde{\delta}) + \left(\frac{L-\mu}{L+m}\right) V[x_k](x_*).$$

Further,

$$V[x_{k+1}](x_*) \leq \frac{1}{L+m}(\delta + \widetilde{\delta}) + \left(\frac{L-\mu}{L+m}\right)\left(\frac{1}{L+m}(\delta + \widetilde{\delta}) + \left(\frac{L-\mu}{L+m}\right) V[x_{k-1}](x_*)\right) \leq \dots \leq$$
$$\leq \frac{1}{L+m}(\widetilde{\delta} + \delta)\left(1 + \left(\frac{L-\mu}{L+m}\right) + \dots + \left(\frac{L-\mu}{L+m}\right)^k\right) + \left(\frac{L-\mu}{L+m}\right)^{k+1} V[x_0](x_*).$$

Therefore, taking into account the following fact

$$\sum_{i=0}^{k}\left(\frac{L-\mu}{L+m}\right)^i < \frac{1}{1 - \left(\frac{L-\mu}{L+m}\right)} = \frac{L+m}{\mu+m},$$

we obtain

$$V[x_{k+1}](x_*) \leq \frac{\delta + \widetilde{\delta}}{m+\mu} + \left(\frac{L-\mu}{L+m}\right)^{k+1} V[x_0](x_*).$$

Now we consider the question on convergence by function:

$$V[x_{k+1}](x_*) \leq \left(f(x_*) - f(x_{k+1}) + \delta + \widetilde{\delta}\right)\frac{1}{L+m} + \left(\frac{L-\mu}{L+m}\right) V[x_k](x_*) \leq$$
$$\leq \left(f(x_*) - f(x_{k+1}) + \delta + \widetilde{\delta}\right)\frac{1}{L+m} +$$
$$+ \left(\frac{L-\mu}{L+m}\right)\left(\left(f(x_*) - f(x_k) + \delta + \widetilde{\delta}\right)\frac{1}{L+m} + \left(\frac{L-\mu}{L+m}\right) V[x_{k-1}](x_*)\right) \leq$$
$$\leq \dots \leq \left(\frac{L-\mu}{L+m}\right)^{k+1} V[x_0](x_*) + \frac{1}{L+m}\sum_{i=0}^{k}\left(\frac{L-\mu}{L+m}\right)^i \left(f(x_*) - f\left(x_{k+1-i}\right) + \delta + \widetilde{\delta}\right).$$

Therefore, we have

$$\frac{1}{L+m}\sum_{i=0}^{k}\left(\frac{L-\mu}{L+m}\right)^i (f(x_{k+1-i}) - f(x_*)) \leq \left(\frac{L-\mu}{L+m}\right)^{k+1} V[x_0](x_*) + \frac{1}{L}\sum_{i=0}^{k}\left(\frac{-mu}{L+m}\right)^i (\delta + \widetilde{\delta}).$$

Denote by $y_k = \operatorname{argmin}_{i=1,\dots,k}(f(x_i))$. Then, taking into account

$$\frac{1}{m+L}\sum_{i=0}^{k}\left(\frac{L-\mu}{m+L}\right)^i = \frac{1}{m+L}\frac{\left(\left(\frac{L-\mu}{m+L}\right)^{k+1}-1\right)}{\frac{L-\mu}{m+L}-1} = \frac{1-\left(\frac{L-\mu}{m+L}\right)^{k+1}}{m+\mu},$$

we obtain

$$f(y_{k+1}) - f(x_*) \leqslant (m+\mu)\frac{\left(\frac{L-\mu}{m+L}\right)^{k+1}}{1-\left(\frac{L-\mu}{m+L}\right)^{k+1}}V(x_*,x_0) + \delta + \widetilde{\delta} \leqslant$$
$$\leqslant (m+L)\exp\left(-\frac{m+\mu}{m+L}(k+1)\right)V(x_*,x_0) + \delta + \widetilde{\delta}.$$

□

C Some Numerical Tests for Algorithms 1 and 4

We consider two numerical examples for Algorithms 1 and 4 for minimizing μ-strongly convex objective function of N variables on a unit ball $B_1(0)$ with center at zero with respect to the standard Euclidean norm. It is clear that such functions admit (δ, L, μ)-model of the standard form $\psi_\delta(x,y) = \langle \nabla f(y), x-y\rangle$ for the case of Lipschitz-continuous gradient ∇f. In the first of the considered examples, it is easy to estimate L and μ, and the ratio $\frac{\mu}{L}$ is not very small, which ensures a completely acceptable rate of convergence of the non-adaptive method (see Table 1 below). In the second example, the objective is ill-conditioned meaning that the ratio $\frac{\mu}{L}$ so small that the computer considers the value $1-\frac{\mu}{L}$ to be equal to 1 and Theorem 29 for the non-adaptive algorithm does not allow to estimate the rate of convergence at all. In this case, the use of adaptive Algorithm 4 leads to noticeable results (see the Table 2 below).

Example 30. Given a matrix A of size 2000×2000 and a vector $b \in \mathbb{R}^{2000}$ with coordinates represented by random integers from the interval $[-1000, 1000]$. The matrix A is a diagonal matrix in which the main diagonal is represented by random integers from the interval $[1, 1000]$, as well as 100 randomly selected elements of this matrix are replaced by integers from the interval $[1, 1000]$.

Consider the problem of solving the matrix equation, which, in the case of solvability, is equivalent to the problem of minimizing the convex functional $f(x) = \|Ax-b\|_2^2$. This function is μ-strongly convex and has L -Lipschitz gradient, where μ is the smallest positive eigenvalue of the matrix A^TA, L is the largest eigenvalue of A^TA (A^T is the matrix transposed to A). Start point $x^0 = (0.2, \dots, 0.2)$ selected. The values $\left(1-\frac{\mu}{L}\right)^k$ and (7) are compared, which determine the quality of the solution for the 4 algorithm and its non-adaptive version. The results are presented in the table ref tab1. As you can see, with the same number of iterations and comparable time costs, the adaptive method guarantees a slightly better solution quality.

Average results (for different matrices) of 10 experiments of the comparison of the work of algorithm 4 and its non-adaptive version are presented in the comparative Table 1, where K is the number of iterations of these algorithms. Time presented in seconds.

Example 31. We consider operator

$$g(x_1, x_2, \dots, x_n) = \left(e^{x_1+\frac{x_2}{10e^3}}, e^{x_2+\frac{x_3}{10e^3}}, \dots, e^{x_n+\frac{x_1}{10e^3}}\right),$$

Table 1: Results for Example 30.

	Adaptive		Non-adaptive	
K	**Time, s**	**Estimate**	**Time, s**	**Estimate**
100	5.2	0.99996	28.9	0.33332
200	10.2	0.99993	52.3	0.33332
300	15.5	0.99989	75.4	0.33332
400	20.7	0.99986	98.9	0.33332
500	25.9	0.99982	122.4	0.33332

Table 2: Results for Example 31.

	Algorithm 5		Algorithm 6	
ε	**Time, s**	**Iterations**	**Time, s**	**Iterations**
$1/2$	209.0	3	386.0	6
$1/4$	205.7	3	454.0	7
$1/6$	278.3	4	516.0	8
$1/8$	272.3	4	529.0	8
$1/10$	335.7	5	597.0	9

initial point $x^0 = (x_1, x_2, \ldots, x_n) \in Q$ ($Q = B_1(0) = \{x \in \mathbb{R}^n \mid \|x\|_2 \leqslant 1\}$, $x_1, x_2, \ldots, x_n$ — random numbers from the interval $(0, 1)$), $L_0 = \frac{\|g(1,0,0,\ldots,0) - g(0,1,0,\ldots,0)\|}{\sqrt{2}}$ and $\mu = \frac{9}{10}\exp(-\sqrt{2})$ for standard Euclidean norm in $\mathbb{R}^n$. Average results (for different initial points) of 10 experiments of the comparison of the work of algorithms 5 and 6 for $n = 10000$ are presented in the comparative tables 2 and 3. Time presented in seconds.

Thus, with $\varepsilon \leqslant \frac{1}{10}$ the algorithm 5 works faster than algorithm 5, however, for higher accuracy, the situation changes (see table 3). Experiments were performed using CPython 3.7 software on a computer with a 3-core AMD Athlon II X3 450 processor with a clock frequency of 3.2 GHz. The computer's RAM was 8 GB.

D Analysis of Algorithm in the case of $\psi_\delta(x, y)$ is m-strongly convex function

Let us denote

$$q_k \overset{\text{def}}{=} \frac{L_k - \mu}{L_k + m} \leq q \overset{\text{def}}{=} \frac{2L - \mu}{2L + m}$$

Table 3: Results for Example 31.

	Algorithm 5		Algorithm 6	
ε	**Time, s**	**Iterations**	**Time, s**	**Iterations**
$5 \cdot 10^{-4}$	923	13	1318	16
10^{-4}	2501	26	1716	19
$5 \cdot 10^{-5}$	$>$ 3600	—	2244	20
10^{-5}	$>$ 3600	—	2456	22

and

$$Q_j^k \overset{\text{def}}{=} \prod_{i=j}^{k} q_i.$$

for all $k \geq 0$.

Theorem 32. *We denote* $y_{k+1} = \operatorname{argmin}_{i=1,\dots,k+1} f(x_i)$.

1

$$\begin{aligned}\frac{1}{2}\|x_{k+1} - x_*\|^2 &\leq V[x_{k+1}](x_*) \\ &\leq Q_1^{k+1} V[x_0](x_*) + (\widetilde{\delta} + 2\delta)\sum_{i=1}^{k+1} \frac{Q_{i+1}^{k+1}}{L_i + m}\end{aligned}$$

2

$$\begin{aligned}f(y_{k+1}) - f(x_*) &\leq \frac{Q_1^{k+1}}{\sum_{i=1}^{k+1} \frac{Q_{i+1}^{k+1}}{L_i+m}} V[x_0](x_*) + \widetilde{\delta} + 2\delta \\ &\leq \min\left[(L_{k+1}+m)Q_1^{k+1}V[x_0](x_*), \frac{1}{\sum_{i=1}^{k+1}\frac{1}{L_i+m}} V[x_0](x_*)\right] + \widetilde{\delta} + 2\delta \\ &\leq \min\left[(2L+m)q^{k+1}V[x_0](x_*), \frac{2L+m}{k+1} V[x_0](x_*)\right] + \widetilde{\delta} + 2\delta.\end{aligned}$$

3

$$\begin{aligned}f(\bar{x}_N) - f(x_*) + \frac{m}{2}\|\bar{x}_N - x_*\|^2 &\leq \frac{V[x_0](x_*)}{A_N} + \widetilde{\delta} + 2\delta \\ &\leq \frac{2LV[x_0](x_*)}{N} + \widetilde{\delta} + 2\delta.\end{aligned}$$

and for case when $m > 0$*:*

$$\begin{aligned}\frac{1}{2}\|\bar{x}_N - x_*\|^2 &\leq \frac{V[x_0](x_*)}{mA_N} + \frac{\widetilde{\delta} + 2\delta}{m} \\ &\leq \frac{2LV[x_0](x_*)}{mN} + \frac{\widetilde{\delta} + 2\delta}{m}\end{aligned}$$

Proof. 1

$$\begin{aligned}f(x_{k+1}) &\leq f_\delta(x_{k+1}) + \delta \\ &\leq f_\delta(x_k) + \psi_\delta(x_{k+1}, x_k) + L_{k+1}V[x_k](x_{k+1}) + 2\delta \\ &\leq f_\delta(x_k) + \psi_\delta(x, x_k) + L_{k+1}V[x_k](x) - (L_{k+1}+m)\,V[x_{k+1}](x) + \widetilde{\delta} + 2\delta \\ &\leq f(x) + (L_{k+1} - \mu)V[x_k](x) - (L_{k+1}+m)\,V[x_{k+1}](x) + \widetilde{\delta} + 2\delta.\end{aligned}$$

Let us take $x = x_*$, then $f(x_*) \leq f(x_{k+1})$ and

$$(L_{k+1}+m)\,V[x_{k+1}](x_*) \leq (L_{k+1} - \mu)V[x_k](x_*) + \widetilde{\delta} + 2\delta.$$

Thus, we have that

$$V[x_{k+1}](x_*) \leq q_{k+1}V[x_k](x_*) + \frac{\widetilde{\delta} + 2\delta}{L_{k+1} + m}$$
$$\leq Q_1^{k+1}V[x_0](x_*) + (\widetilde{\delta} + 2\delta)\sum_{i=1}^{k+1}\frac{Q_{i+1}^{k+1}}{L_i + m}.$$

2

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

Using that $V[x_{k+1}](x_*) \geq 0$ and y_{k+1} definition we have

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

Let us divide inequality by $\sum_{i=1}^{k+1}\frac{Q_{i+1}^{k+1}}{L_i+m}$:

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

Notice that $\sum_{i=1}^{k+1}\frac{Q_{i+1}^{k+1}}{L_i+m} \geq \frac{1}{L_{k+1}+m}$ and $Q_i^{k+1} \geq Q_1^{k+1}$ for all $i \geq 1$, this gives us the following ineqaulity

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

3

$$f(x_{k+1}) - f(x_*) + (L_{k+1} + m)\,V[x_{k+1}](x_*) \leq (L_{k+1} - \mu)V[x_k](x_*) + \widetilde{\delta} + 2\delta$$
$$\leq L_{k+1}V[x_k](x_*) + \widetilde{\delta} + 2\delta.$$

After dividing both parts by L_{k+1} we have:

$$\frac{1}{L_{k+1}}(f(x_{k+1}) - f(x_*)) + \left(1 + \frac{m}{L_{k+1}}\right)V[x_{k+1}](x_*) + \frac{\widetilde{\delta} + 2\delta}{L_{k+1}} \leq V[x_k](x_*) + \frac{\widetilde{\delta} + 2\delta}{L_{k+1}}.$$

Let us telescope the last inequality for k from 0 to N and use 1–strong convexity of V:

$$\sum_{k=0}^{N}\frac{1}{L_{k+1}}(f(x_{k+1}) - f(x_*)) + \sum_{k=0}^{N}\frac{m}{2L_{k+1}}\|x_i - x_*\|^2$$
$$\leq V[x_0](x_*) - V[x_N](x_*) + (\widetilde{\delta} + 2\delta)A_N.$$

Now we divide both parts by A_N. Using the fact that $V[x_N](x_*) \geq 0$ and convexity we can get:

$$f(\bar{x}_N) - f(x_*) + \frac{m}{2}\|\bar{x}_N - x_*\|^2 \leq \frac{V[x_0](x_*)}{A_N} + \widetilde{\delta} + 2\delta.$$

□