Moscow Institute of Physics and Technology

Bachelor Thesis

Stopping Rules in Mirror Descent Algorithm

Author:
Vo Thi Thu Ha

Supervisor:
Fedor Stonyakin

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Discrete Mathematics Direction
Department of Applied Mathematics and Informatics

Abstract

In this thesis, we study the Mirror Descent method applied to optimization problems with non-Lipschitz smooth objective functions. This method involves a distance generating function that satisfies the condition of relative smoothness. At each iteration step, this function is utilized to compute the Bregman divergence, which is then used to update the iterative process. Beyond the traditional Mirror Descent algorithm, we introduce an adaptive method that can be applied when the value L in the class of relatively smooth functions is unknown. We also provide a proof of the linear convergence rate and a new stopping rule for this algorithm. To validate our theoretical results, we conduct numerical experiments to assess the effectiveness of these methods in solving the well-known Quadratic Inverse Problems. Subsequently, we compare the performance of both adaptive and non-adaptive methods, highlighting their respective advantages and disadvantages across various circumstances.

Contents

1 Introduction

First-order methods, also known as gradient methods, play a crucial role in optimization due to their wide range of applications across various fields such as signal recovery, machine learning and image reconstruction. For example, Stochastic Gradient Descent (SGD) is widely used for training machine learning and deep learning models, see [24, 6, 13], while Proximal Gradient Descent has many applications in image denoising, sparse recovery and matrix completion, see [17, 4, 16, 7]. As their name suggests, these methods utilize the gradients or subgradients of the functions involved in the optimization model. Their popularity stems from their ability to handle large and complex optimization problems effectively while maintaining simplicity in implementation.

A common way to approach these problems is by reformulating them as classical optimization problems. Specifically, consider $Q \subset E$ be a closed convex subset, and let f be a differentiable function. Here we need to find

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

There are many gradient methods introduced to solve (1), with a common assumption being that the target function always has the Lipschitz continuous gradient (also known as L-smoothness), see [18, 22, 2]. This condition means that there exists a positive value L such that

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

where $\|.\|$ denotes the norm over E and $\|.\|_*$ denotes the dual norm on E^*.

Here the dual norm is defined as follows: Let E be a normed vector space with norm $\|.\|$ and E^* be its continuous dual space, then

$$\|x\|_* = \max_{\|y\| \leq 1} \langle y, x \rangle$$

However, the condition for L-smoothness is often strict and difficult to satisfy in many real-world problems. For example, [1, 14] highlight several optimization problems where the objective functions do not meet this condition. As a result, there is growing interest in developing optimization methods that do not require Lipschitz smoothness.

In recent years, many researchers have introduced a new concept for first-order methods called relative smoothness. This condition corresponds to an important inequality with the Bregman divergence, which we will explore further in Section 2,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + LV(y, x), \quad \forall x, y \in \operatorname{int} Q$$

This property is satisfied by a large number of problems, see [1, 14]. In [1], the authors derive a new Descent Lemma without Lipschitz gradient continuity, which is used to define the relatively smooth functions. This work is developed using composite optimization problems, which involve minimizing the sum of two non-convex functions: a possibly extended-valued function g and a differentiable function f, see [1, 17]

$$\inf \{f(x) + g(x) : x \in \overline{Q}\}, \tag{2}$$

where $\overline{Q}$ denotes the closure of Q. Moreover, this closure is assumed to be a convex, nonempty and open set.

The work of [14] also introduces the definition of relative smoothness but extends it by ignoring the smoothness and convexity of the generating distance function (also known as the kernel function).

The main content of this thesis is motivated by these results, and the outline of the thesis is as follows.

- In Section 2, we study the concept of relative smoothness and provide the main properties and definitions based on works [1, 14]. Moreover, we study the Mirror Descent method on this class of problems and derive an important theorem for the Bregman divergence.
- In Section 3, we propose several variants of the Mirror Descent algorithm, including constant step-size and adaptive step-size, considering the possibility of inexactness. Additionally, we provide theoretical estimates for these approaches.
- In Section 4, we investigate the convergence of the Mirror Descent method and prove that its convergence rate is linear. Moreover, the results from the Polyak-Lojasiewicz (PL) condition also create a new stopping criterion for this problem.
- Finally, to demonstrate the potential of our approach, in Section 5, we conduct numerical experiments on Quadratic Inverse Problems, comparing the results between adaptive and non-adaptive methods, computing the required number of iterations to meet the early stopping criterion, and comparing theoretical estimates with actual values of the Bregman divergence.

2 Relative Smoothness and the Mirror Descent Algorithm

In this section, we introduce the definitions of the distance generating function and Bregman divergence, alongside the definition of relative smoothness, which are fundamental throughout this thesis. These concepts are motivated by the works of Bauschke et al. [1] and Lu et al. [14]. Additionally, we present the Mirror Descent method and derive a theorem analogous to that of the norm of the gradient, but formulated for the Bregman divergence.

2.1 Bregman Divergence and Relative Smoothness

Consider a closed convex subset $Q \subset E$ and a distance generating function (also known as a prox function or a kernel function) $d : Q \to \mathbb{R}$ which is continuously differentiable and convex. Then, for all $x, y \in Q$, we define the Bregman divergence (or Bregman distance) as follows

$$V(y, x) := d(y) - d(x) - \langle \nabla d(x), y - x \rangle \tag{3}$$

The application of Bregman divergence and the distance generating function has been extensively discussed in previous research, see [9, 28]. As we can see, the Bregman divergence $V(y, x)$ is non-symmetric and convex with respect to y. Moreover, due to the convexity of d, it is also non-negative. Therefore, $V(y, x)$ can be understood as a generalized measure of distance between points in the subset Q. In [14], it is clarified that the generating function d does not need to be strictly convex and essentially smooth. The crucial requirement is that d satisfies the condition of relative smoothness with respect to the objective function.

Notice that if we choose the distance generating function $d(x) = \frac{1}{2}\|x\|^2$, we obtain the Bregman divergence $V(y, x) = \frac{1}{2}\|y - x\|^2$, which is called the Euclidean divergence.

We also recall the definition of relative smoothness from [1, 14], here our objective function f is not necessarily convex.

Definition 2.1 (Relative smoothness)**.** *We say that function f is smooth relative to d on* $\operatorname{int} Q$ *(here* $\operatorname{int} Q$ *denotes the interior of Q) if $\exists L > 0$ such that*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + LV(y, x), \quad \forall x, y \in \operatorname{int} Q \tag{4}$$

for a gradient (or arbitrary subgradient) ∇f of the function f.

Regarding the work in [1, 14], this definition is also equivalent to $Ld - f$ being a convex function on $\operatorname{int} Q$ (which often makes it easier to determine the value of L that satisfies relative smoothness). Since $V(y, x) \geq 0$, if L satisfies (4), then for all $L' > L$, it also satisfies (4).

Moreover, if we define the distance generating function $d(x) = \frac{1}{2}\|x\|^2$, we obtain the Bregman divergence $V(y, x) = \frac{1}{2}\|y - x\|^2$. Then, the condition for relative smoothness becomes

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

which is exactly the well-known Descent Lemma for Lipschitz smooth functions.

2.2 Mirror Descent Algorithm

Here, we recall the Mirror Descent algorithm (also known as the Bregman Gradient method), which uses the Bregman divergence and follows the iteration rule

$$x_{k+1} \leftarrow \arg\min_{x \in Q} \{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + LV(x, x_k) \} \tag{5}$$

Now, our concern is how to solve the complex problem (5). In the works of [5, 14], it is shown that problem (5) can be efficiently solved by converting the initial problem into

$$x_{k+1} \leftarrow \arg\min_{x \in Q} \{ \langle p, x \rangle + d(x) \}$$

where $p := \frac{1}{L} \nabla f(x_k) - \nabla d(x_k)$. The intuition comes from the fact that the function in the curly bracket is convex with respect to x. Therefore, by the optimality condition, we obtain

$$\nabla f(x_k) + L \nabla d(x_{k+1}) - L \nabla d(x_k) = 0$$

This result is crucial for the numerical part of our analysis. Notice that if we consider the distance generating function $d(x) = \frac{1}{2}\|x\|^2$, then $\nabla d(x) = x$, the above equation becomes

$$\nabla f(x_k) + L x_{k+1} - L x_k = 0$$

equivalent to

$$x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$$

which is the formula of the Gradient Descent.

Let us formulate a theorem about the upper bound of the minimum value of the Bregman divergence for this algorithm.

Theorem 2.1. *Consider the following optimization problem*

$$\min_{x \in Q} f(x)$$

Assume that f is smooth relative to a convex function d with parameter L. Let us consider the Mirror Descent algorithm

$$x_{k+1} \leftarrow \arg\min_{x \in Q} \{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + LV(x, x_k) \}$$

Then, for x_0 as a starting point and x_* as one of the exact solutions of the problem (1)*, the following inequality holds*

$$\min_{k=0,\ldots,N-1} V(x_k, x_{k+1}) \leq \frac{f(x_0) - f(x_*)}{NL} \tag{6}$$

Notice that the left-hand side of Theorem 2.1 is analogous to the norm of the gradient, see [23, Theorem 1.1]. In order to prove it, we need to revisit and utilize the Three-Point Property [14, Lemma 3.1].

Lemma 2.1 (Three-point Property)**.** *Let $\phi(x)$ be a convex function, let $V(\cdot,\cdot)$ be the Bregman divergence for $d(\cdot)$. For a given vector z, let*

$$z^+ := \arg\min_{x\in Q}\{\phi(x) + V(x,z)\}$$

Then

$$\phi(x) + V(x,z) \geq \phi(z^+) + V(z^+,z) + V(x,z^+), \quad \forall x \in Q$$

Proof of Theorem 2.1. Let consider

$$z^+ \leftarrow \arg\min_{x\in Q}\left\{\frac{\langle\nabla f(x_k), x\rangle}{L} + V(x,z)\right\}$$

then apply Lemma 2.1 with $x = z = x_k$ and $z^+ = x_{k+1}$, we obtain

$$\frac{\langle\nabla f(x_k), x_k\rangle}{L} + V(x_k,x_k) \geq \frac{\langle\nabla f(x_k), x_{k+1}\rangle}{L} + V(x_{k+1},x_k) + V(x_k,x_{k+1})$$

From this, we get

$$\langle\nabla f(x_k), x_{k+1} - x_k\rangle \leq -LV(x_{k+1},x_k) - LV(x_k,x_{k+1}) \tag{7}$$

Take into account (4), we receive

$$f(x_{k+1}) - f(x_k) \leq -LV(x_k,x_{k+1}) \tag{8}$$

Let $f^* = f(x_*)$ be the value of the function f at one of the exact solution x_*, then

$$\begin{aligned} f^* - f(x_0) &\leq f(x_N) - f(x_0) \\ &\leq -L\sum_{k=0}^{N-1} V(x_k,x_{k+1}) \\ &\leq -NL\min_{k=0,\ldots,N-1} V(x_k,x_{k+1}) \end{aligned}$$

i.e

$$\min_{k=0,\ldots,N-1} V(x_k,x_{k+1}) \leq \frac{f(x_0) - f(x_*)}{NL}$$

□

From the result of this theorem, it is evident that when the number of iterations is sufficiently large, we can achieve a small value of $V(x_k,x_{k+1})$. Hence, the sequence of $V(x_k,x_{k+1})$ will converge to 0, ensuring that the method attains to the optimal solution, even though the function f is non-convex.

3 Variants of the Mirror Descent Algorithm

In this section, we try to review some variants of the Mirror Descent algorithm. Specially, we are also interested in the notation of inexact gradient, which is denoted by $\nabla_\Delta f(x)$ (this value is the approximate value of $\nabla f(x)$ at point x). Moreover, $\forall y \in \operatorname{int} Q$, this inequality needs to be held

$$f(y) \leq f(x) + \langle \nabla_\Delta f(x), y - x \rangle + LV(y, x) + \Delta \tag{9}$$

for a small value of $\Delta > 0$. So, $\nabla_\Delta f$ is an analogue of inexact oracle for relative smooth problems (see e.g. [27, 26]). For example, if for each x, $\|\nabla_\Delta f(x) - \nabla f(x)\|_* \leq \delta$ for some $\delta > 0$ and bounded Q, we have $\forall x, y \in int\ Q$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + LV(y, x) \leq f(x) + \langle \nabla_\Delta f(x), y - x \rangle + LV(y, x) + \delta \cdot diam(Q).$$

3.1 The Mirror Descent Algorithm with a Constant Step-size

Let us now consider the first variant of the Mirror Descent algorithm where the inexact gradient $\Delta > 0$ and $L > 0$ are known. Moreover, the value of L must satisfy the condition of relative smoothness (Definition 2.1). Then, the variant of the Mirror Descent algorithm is generated as follows

Algorithm 1 Mirror Descent with a Constant Step-size

1: Input: Initial point $x_0 \in \operatorname{int} Q$, L
2: **for** $k = 0, 1, \ldots$ **do**
Compute $\nabla_\Delta f(x_k)$:

$$x_{k+1} \leftarrow \arg\min_{x \in Q} \{f(x_k) + \langle \nabla_\Delta f(x_k), x - x_k \rangle + LV(x, x_k)\}$$

3: **end for**

Notice that, the Mirror Descent method has the following update form

$$x_{k+1} \leftarrow \arg\min_{x \in Q} \{f(x_k) + \langle \nabla_\Delta f(x_k), x - x_k \rangle + LV(x, x_k)\} \tag{10}$$

which can be used to address the minimization problem of the function f with an inexact gradient. Using the same result obtained from the proof of Theorem 2.1, we receive

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla_\Delta f(x_k), x_{k+1} - x_k \rangle + LV(x_{k+1}, x_k) + \Delta \\ &\leq f(x_k) - LV(x_k, x_{k+1}) + \Delta \end{aligned}$$

i.e

$$f(x_{k+1}) - f(x_k) \leq -LV(x_k, x_{k+1}) + \Delta \tag{11}$$

Let $f^* = f(x_*)$ be the value of the function f at one of the exact solution x_*, then

$$\begin{aligned} f^* - f(x_0) &\leq f(x_N) - f(x_0) \\ &\leq \sum_{k=0}^{N-1} (-LV(x_k, x_{k+1}) + \Delta) \\ &\leq -NL \min_{k=0,\dots,N-1} V(x_k, x_{k+1}) + N\Delta \end{aligned}$$

Finally, we obtain

$$\min_{k=0,\dots,N-1} V(x_k, x_{k+1}) \leq \frac{f(x_0) - f(x_*)}{NL} + \frac{\Delta}{L} \tag{12}$$

It is crucial to choose an appropriate kernel function d. According to [14, Section 1.3], f needs to be smooth relative to d with parameter L, and the problem (10) always has a solution which can be computed efficiently. This requires that the function d is easy to compute and suitable for numerical computations.

3.2 The Mirror Descent Algorithm with an Adaptive Step-size

Let us consider the modified version of Algorithm 1 when the value of L is not known in advance. Therefore, at each iteration, we update the step-size to ensure it satisfies the condition of relative smoothness.

Algorithm 2 Mirror Descent with an Adaptive Step-size

1: Input: Initial point $x_0 \in \operatorname{int} Q$, $L_0 > 0$, $\Delta > 0$
2: Set $L_{k+1} := \frac{L_k}{2}$
3: Find

$$x_{k+1} \leftarrow \arg\min_{x \in Q} \{f(x_k) + \langle \nabla_\Delta f(x_k), x - x_k \rangle + L_{k+1} V(x, x_k)\}$$

4: If

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla_\Delta f(x_k), x_{k+1} - x_k \rangle + L_{k+1} V(x_{k+1}, x_k) + \Delta$$

then $k := k + 1$ and go to Step 2. Otherwise, $L_{k+1} := 2L_{k+1}$ and go to Step 3.

From (11), we can easily obtain the similar result with an adaptive step-size

$$f(x_{k+1}) - f(x_k) \leq -L_{k+1} V(x_k, x_{k+1}) + \Delta \tag{13}$$

Let $f^* = f(x_*)$ be the value of the function f at one of the exact solution x_*, then

$$\begin{aligned} f^* - f(x_0) &\leq f(x_N) - f(x_0) \\ &\leq \sum_{k=0}^{N-1} (-L_{k+1} V(x_k, x_{k+1}) + \Delta) \\ &\leq - \min_{k=0,\dots,N-1} V(x_k, x_{k+1}) \sum_{k=0}^{N-1} L_{k+1} + N\Delta \end{aligned}$$

This result is equivalent to

$$\min_{k=0,\dots,N-1} V(x_k, x_{k+1}) \leq \frac{f(x_0) - f(x_*)}{\sum_{k=0}^{N-1} L_{k+1}} + \frac{N\Delta}{\sum_{k=0}^{N-1} L_{k+1}} \tag{14}$$

It is evident that the adaptive method has one more step that can be repeated many times until it satisfies the condition of relative smoothness and inexact gradient. Therefore, to evaluate the effectiveness of this method regarding running time, we need to estimate the elementary steps that are needed in this algorithm.

Assume that at iteration $k+1$, we need to spend l_{k+1} steps to satisfy the inequality

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla_\Delta f(x_k), x_{k+1} - x_k \rangle + L_{k+1} V(x_{k+1}, x_k) + \Delta \tag{15}$$

Since at the beginning $L_{k+1} = \frac{L_k}{2}$, we can observe that

$$2^{l_{k+1}-2} = \frac{L_{k+1}}{L_k}$$

In order to obtain the sum of l_{k+1}, which represents exactly the number of additional steps, we need to compute this product across all iterations

$$\prod_{k=0}^{N-1} 2^{l_{k+1}-2} = \prod_{k=0}^{N-1} \frac{L_{k+1}}{L_k}$$
$$\prod_{k=0}^{N-1} 2^{l_{k+1}-2} = \frac{L_N}{L_0}$$

Take the logarithm on both sides

$$\sum_{k=0}^{N-1} (l_{k+1} - 2) = \log \frac{L_N}{L_0}$$
$$\sum_{k=0}^{N-1} l_{k+1} = \log \frac{L_N}{L_0} + 2N$$

Since f is smooth relative to d with parameter L, we can conclude that $L_N \leq 2L$ (in other words, L_N is constant with respect to N). This implies that $\log \frac{L_N}{L_0}$ is independent of N. Therefore, the number of elementary steps is insignificant, and indeed equal to $O(N)$. Additionally, on average, one iteration of the adaptive method (Algorithm 2) is roughly equivalent to about two iterations of the non-adaptive method (Algorithm 1).

3.3 The Mirror Descent Algorithm with an Adaptive Step-size and Inexactness

Consider the following modification of Algorithm 2, which incorporates the inexactness δ_k.

Algorithm 3 Mirror Descent with an Adaptive Step-size and Inexactness

1: Input: Initial point $x_0 \in \operatorname{int} Q$, $L_0 > 0$, $\delta_0 > 0$
2: $L_{k+1} := \frac{L_k}{2}, \delta_{k+1} := \frac{\delta_k}{2}$
3: Find
$$x_{k+1} \leftarrow \arg\min_{x \in Q} \{f(x_k) + \langle \nabla_{\delta_{k+1}} f(x_k), x - x_k \rangle + L_{k+1} V(x, x_k)\}$$
4: If
$$f(x_{k+1}) \leq f(x_k) + \langle \nabla_{\delta_{k+1}} f(x_k), x_{k+1} - x_k \rangle + L_{k+1} V(x_{k+1}, x_k) + \delta_{k+1}$$
then $k := k + 1$ and go to Step 2. Otherwise, $L_{k+1} := 2L_{k+1}, \delta_{k+1} := 2\delta_k$ and go to Step 3.

Using the result from (11), we obtain the similar inequality with an adaptive step-size and inexactness

$$f(x_{k+1}) - f(x_k) \leq -L_{k+1} V(x_k, x_{k+1}) + \delta_{k+1} \tag{16}$$

Let $f^* = f(x_*)$ be the value of the function f at one of the exact solution x_*, then

$$\begin{aligned} f^* - f(x_0) &\leq f(x_N) - f(x_0) \\ &\leq \sum_{k=0}^{N-1} (-L_{k+1} V(x_k, x_{k+1}) + \delta_{k+1}) \\ &\leq - \min_{k=0,\ldots,N-1} V(x_k, x_{k+1}) \sum_{k=0}^{N-1} L_{k+1} + \sum_{k=0}^{N-1} \delta_{k+1} \end{aligned}$$

Finally we get this result

$$\min_{k=0,\ldots,N-1} V(x_k, x_{k+1}) \leq \frac{f(x_0) - f(x_*)}{\sum_{k=0}^{N-1} L_{k+1}} + \frac{\sum_{k=0}^{N-1} \delta_{k+1}}{\sum_{k=0}^{N-1} L_{k+1}} \tag{17}$$

4 Convergence Rate and a New Stopping Rule for the Mirror Descent Algorithm

To guarantee the convergence of the non-convex objective function f to a global minimum, we also study the Polyak-Lojasiewicz (PL) condition, see [21].

Consider a function $h : Q \to \mathbb{R}$ and assume that a minimizer x^* exists, though it may not be unique. We say that h satisfies the μ-PL inequality, also known as the Polyak-Lojasiewicz condition, if for some $\mu > 0$,

$$\frac{1}{2}\|\nabla h(x)\|_2^2 \geq \mu(h(x) - h(x^*)) \tag{18}$$

We aim to show that the PL condition is sufficient to ensure the linear convergence of the Mirror Descent algorithm. Returning to the problem

$$x_L \leftarrow \arg\min_{y \in Q}\{\langle \nabla f(x), y - x\rangle + LV(y, x)\} \tag{19}$$

Since $\langle \nabla f(x), y - x\rangle$ is linear with respect to y, it is convex. Additionally, $V(y, x)$ is convex by the way we define the kernel function d. Therefore, their sum is also convex, which allows us to apply the first-order global optimality condition

$$\begin{aligned}
\frac{d}{dy}\left(\langle \nabla f(x), y - x\rangle + LV(y, x)\right)\Big|_{y=x_L} &= 0 \\
\nabla f(x) + L\left(\nabla d(y) - \nabla d(x)\right)\Big|_{y=x_L} &= 0 \\
\nabla f(x) + L\left(\nabla d(x_L) - \nabla d(x)\right) &= 0
\end{aligned}$$

which means

$$\nabla d(x_L) = \nabla d(x) - \frac{1}{L}\nabla f(x) \tag{20}$$

Now, our task is to obtain an analogous version of PL inequality for our problem. First, let us assume that $d(x)$ is l-smooth, then by the characterization of l-smoothness [2, Theorem 5.8], $\forall x, y \in Q$

$$d(y) \geq d(x) + \langle \nabla d(x), y - x\rangle + \frac{1}{2l}\|\nabla d(y) - \nabla d(x)\|_*^2$$

i.e

$$V(y, x) \geq \frac{1}{2l}\|\nabla d(y) - \nabla d(x)\|_*^2$$

Hence $\forall x, x_L \in Q$

$$V(x, x_L) \geq \frac{1}{2l}\|\nabla d(x) - \nabla d(x_L)\|_*^2 \overset{(20)}{=} \frac{1}{2lL^2}\|\nabla f(x)\|_*^2 \tag{21}$$

Assume that f satisfies the PL condition (18) with constant μl, then

$$f(x) - f^* \leq \frac{1}{2\mu l}\|\nabla f(x)\|_*^2 \leq \frac{L^2}{\mu}V(x, x_L) \tag{22}$$

We consider the case with a constant step-size and an inexact gradient. Let f be L-relative smooth, we examine the Mirror Descent algorithm with inexactness

$$x_{k+1} = \arg\min_{x \in Q} \{\langle \nabla_\Delta f(x_k), x - x_k \rangle + LV(x, x_k)\}$$

At each iteration, we aim to satisfy the inequality

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla_\Delta f(x_k), x_{k+1} - x_k \rangle + LV(x_{k+1}, x_k) + \Delta \tag{23}$$

Once again, we recall the result from Theorem 2.1

$$f(x_k) - f(x_{k+1}) \geq LV(x_{k+1}, x_k) - \Delta$$

Then (22) guarantees that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq L\frac{\mu}{L^2}(f(x_k) - f^*) - \Delta \\ &\geq \frac{\mu}{L}(f(x_k) - f^*) - \Delta \quad (\text{here, } \mu \leq L) \end{aligned}$$

Consequently, we obtain the result

$$\begin{aligned} f(x_{k+1}) - f^* &\leq (f(x_k) - f^*)\left(1 - \frac{\mu}{L}\right) + \Delta \\ &\leq (f(x_0) - f^*)\left(1 - \frac{\mu}{L}\right)^{k+1} + \Delta\left(1 + \left(1 - \frac{\mu}{L}\right) + \ldots + \left(1 - \frac{\mu}{L}\right)^k\right) \\ &\leq (f(x_0) - f^*)\left(1 - \frac{\mu}{L}\right)^{k+1} + \frac{\Delta}{1 - (1 - \frac{\mu}{L})} \\ &\leq (f(x_0) - f^*)\left(1 - \frac{\mu}{L}\right)^{k+1} + \frac{\Delta L}{\mu} \end{aligned}$$

This result shows that if the L-smooth relative function f satisfies the analogous PL condition (22), then for the Mirror Descent algorithm that satisfies the inequality (23), the convergence rate is linear. Moreover, since $\mu \leq L$, at iteration k, the value of the rate $\left(1 - \frac{\mu}{L}\right)^{k+1}$ becomes sufficiently small to be negligible, guaranteeing that the algorithm will converge to the global optimum.

Moreover, it is important to notice that the result (22) can be written as

$$f(x_k) - f^* \leq \frac{L^2}{\mu} V(x_k, x_{k+1}) \tag{24}$$

Since the values of L and μ are constant, when $V(x_k, x_{k+1})$ is sufficiently small, the sequence generated by the Mirror Descent algorithm attains its minimal solution. Hence, this condition can be considered as a new stopping rule for the algorithm.

5 Numerical Experiments

In this section, to examine the effectiveness of our proposed approach, we use the class of Quadratic Inverse Problems (QIP) (see [5, 3]). Quadratic Inverse Problems represent a significant area of study due to their widespread applications in various scientific and engineering fields, such as signal processing, phase retrieval and unsupervised learning (see [20, 8, 19]).

First, we will explore the theoretical foundation of Quadratic Inverse Problems (QIP), providing a comprehensive overview of their formulation. Next, we will conduct a comparative analysis of the performance of Algorithm 1 and Algorithm 2.

5.1 Quadratic Inverse Problems

Let us outline the Quadratic Inverse Problems. Given m symmetric matrices $A_1, ..., A_m \in \mathbb{R}^{n\times n}$, and a vector $c \in R^m$ (noisy measurements), our objective is to find $x \in \mathbb{R}^n$, that satisfies the following system

$$x^T A_i x \simeq c_i, \quad i = 1, 2, ..., m \tag{25}$$

As demonstrated, we can rewrite this problem in the form of (1) with

$$f(x) = \sum_{i=1}^{m} (x^T A_i x - c_i)^2$$

and the function f is nonconvex.

Using the least squares method to measure the error, we can reformulate the problem as a non-convex optimization problem, specifically the Sparse Quadratic Inverse Problem (SQIP). Here, we denote $Q \equiv \mathbb{R}^n$. Additionally, we include the regularizer g with the corresponding parameter θ.

$$\text{(SQIP)} \quad \min_{x\in Q}\{f(x) + \theta g(x)\}, \tag{26}$$

where $f(x) : Q \to (-\infty, +\infty]$ and $f(x) := \frac{1}{4}\sum_{i=1}^{m}(x^T A_i x - c_i)^2$, and is continuously differentiable on $\mathbb{R}^n$, while the regularizer $g(x)$ is added with some parameter θ. Here matrices $A_i \in \mathbb{R}^{n\times n}$ are symmetric and sparse, meaning that they are comprised of mostly zero values (see, [3]).

5.2 Solving Quadratic Inverse Problems with Proposed Algorithms

One significant application of relative smoothness is solving the Quadratic Inverse Problem (QIP) (26). Here, we consider the function $f(x)$ where $x \in Q$, $A_i \in \mathbb{R}^{n\times n}$. If we do not consider the regularizer $g(x)$ and set $\theta = 0$, (26) becomes the optimization problem

$$\min_{x\in Q} f(x)$$

The next step is to find a kernel function d and the value of L such that our objective function f is L-smooth relative to d. In the work of [5], it is shown that the appropriate distance generating function is $d = \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2$.

To prove that f is smooth relative to d, we need to refer to an important proposition about the equivalence of the L-relative smoothness condition in [14, Proposition 1.1] (or we can refer to the proof of [1, Lemma 1]). Instead of proving the inequality (4) for arbitrary values of $x, y \in Q$, the task becomes proving that $Ld - f$ is convex on Q, which is a more manageable task. In [5, Lemma 5.1], it is shown that if $L \geq \sum_{k=1}^{m}(3\|A_i\|^2 + \|A_i\||c_i|)$, then $f(x)$ is L-smooth relative to the function d. Hence, we can use this problem for our numerical experiments with the Mirror Descent algorithm.

It is important to notice that each iteration of Algorithm 1 requires the ability to solve the subproblem

$$x_{k+1} \leftarrow \arg\min_{x \in Q}\{\langle \nabla_\Delta f(x_k), x - x_k\rangle + LV(x, x_k)\}$$

The original of this problem is solved in [5, Section 5.2]. If we denote

$$p = p(x_k) := \frac{1}{L}\nabla_\Delta f(x_k) - \nabla d(x_k) \tag{27}$$

then

$$\begin{aligned}
x_{k+1} &= \arg\min_{x \in Q}\{\langle \nabla_\Delta f(x_k), x - x_k\rangle + LV(x, x_k)\} \\
&= \arg\min_{x \in Q}\left\{\langle \frac{1}{L}\nabla_\Delta f(x_k), x - x_k\rangle + d(x) - d(x_k) - \langle \nabla d(x_k), x - x_k\rangle\right\} \\
&= \arg\min_{x \in Q}\left\{\langle \frac{1}{L}\nabla_\Delta f(x_k) - \nabla d(x_k), x - x_k\rangle + d(x)\right\} \\
&= \arg\min_{x \in Q}\left\{\langle p, x\rangle + \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2\right\}
\end{aligned}$$

According to the first-order global optimality condition, it is necessary that (for convenience in writing the indices, let us denote that $u := x_{k+1}$)

$$u_i(1 + \|u\|_2^2) + p_i = 0, \quad i = 1, 2, ..., n \tag{28}$$

In order to solve the problem which is related to sparse matrices, we will now recall the operator soft-thresholding. For any $y \in \mathbb{R}^n$

$$S(\tau, y) = \arg\min_{x \in Q}\left\{\frac{1}{2}\|x - y\|^2 + \tau\|x\|_1\right\} = \operatorname{sgn}(y)(|y| - \tau)_+ = \operatorname{sgn}(y)\max\{|y| - \tau, 0\}$$

Here the problem of minimizing $\frac{1}{2}\|x - y\|^2 + \tau\|x\|_1$ is called Lasso Regression (see, [25]). Using the result of [5, Proposition 5.1], in the case when $\theta = 0$, let

$$v(x) := S\left(\frac{\theta}{L}, p\right)$$

then

$$v(x) = S(0,p) = \arg\min_{x \in Q} \left\{ \frac{1}{2}\|x-p\|_2^2 \right\} = \operatorname{sgn}(p)\max\{|p|,0\} = p$$

Hence, the value of u from (28) can be obtained as follows

$$u = -t^* v(x) = -t^* S\left(\frac{\theta}{L}, p\right) = -t^* p, \tag{29}$$

where t^* is the unique positive solution of

$$t^3\|v(x)\|_2^2 + t - 1 = 0, \tag{30}$$

or equivalent to

$$t^3\|p\|_2^2 + t - 1 = 0 \tag{31}$$

Notice that if we use this approach, we can reduce the main step of the Mirror Descent algorithm to the expression (29) by first solving the cubic equation (31). In this case, $p = \frac{1}{L}\nabla_\Delta f(x_k) - \nabla d(x_k)$ can be easily computed at each iteration. Additionally, this type of cubic equation is known as a depressed cubic equation and can be solved using Cardano's Method (see Appendix A).

We also apply the same method to solve the subproblem in Algorithm 2

$$x_{k+1} \leftarrow \arg\min_{x \in Q}\{f(x_k) + \langle \nabla_\Delta f(x_k), x - x_k\rangle + L_{k+1}V(x, x_k)\}$$

We then implement Algorithm 1 and Algorithm 2 with $\Delta = 0.1$. The starting point $x_0 \in \mathbb{R}^n$, the vector $c \in \mathbb{R}^m$. Additionally, we initialize m symmetric matrices $A_1, \ldots, A_m \in \mathbb{R}^{n\times n}$, where each matrix has mostly zero values except for n values generated from a standard normal distribution. For Algorithm 1, we choose the value $L = \sum_{k=1}^m (3\|A_i\|^2 + \|A_i\||c_i|)$ since it was proven that f is L-smooth relative to $d = \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2$. However, because this value of L is very large and would require many iterations for convergence, we divide it by 500 to ensure a faster convergence while still guaranteeing the problem's convergence. For Algorithm 2, we set the initial value of L_0 to 10.

We applied the same problem and input to both algorithms. Figure 1 presents a comparison of the results, illustrating the value of the objective function f at the output point of each algorithm. As we can see, the adaptive algorithm (Algorithm 2) performs better in this case, requiring fewer iterations to reach the optimal solution, while the non-adaptive algorithm (Algorithm 1) takes more iterations.

Regarding the running time in Figure 2, Algorithm 1 runs faster. This is understandable since each iteration of Algorithm 2 requires more operations to satisfy the inequality (15). Moreover, the running time of Algorithm 2 is linear with respect to the number of iterations. This is evident as in Section 3.2, we prove that the number of elementary steps in the adaptive algorithm is $O(N)$.

However, if we look at both Figure 1 and Figure 2, in the first 20 iterations, Algorithm 2 appears to attain the optimal solution. When we compare the time it takes for Algorithm

2 to reach this value and the time it takes for Algorithm 1 to obtain the same value, the adaptive algorithm (Algorithm 2) works more efficiently. In conclusion, the adaptive algorithm performs better and is suitable for this class of problems when the quality of final result is not exceedingly high.

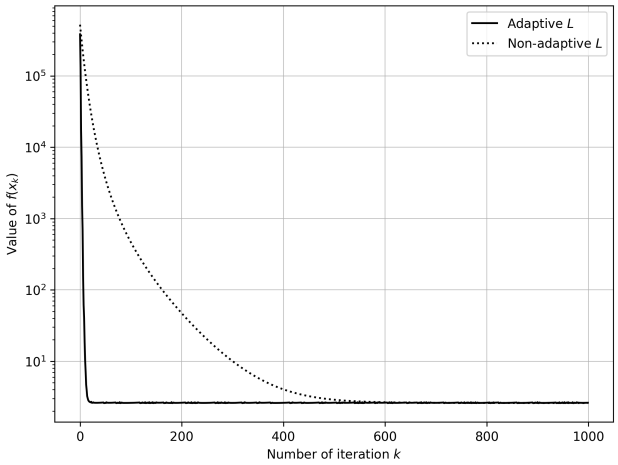


Figure 1: The result of the non-adaptive Mirror Descent algorithm (Algorithm 1) and the adaptive Mirror Descent algorithm (Algorithm 2) with respect to the value of f at the point x_k for QIP with $n = 1000$ and $m = 10$.

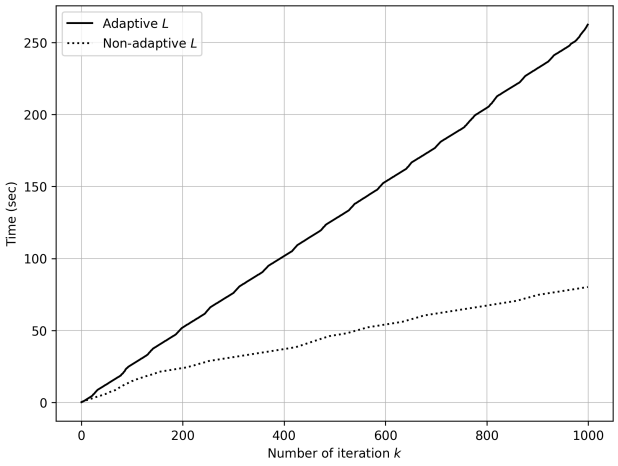


Figure 2: The running time of the non-adaptive Mirror Descent algorithm (Algorithm 1) and the adaptive Mirror Descent algorithm (Algorithm 2) for QIP with $n = 1000$ and $m = 10$.

We are also interested in the performance of the value of $\min V(x_k, x_{k+1})$ in theoretical estimates. As we can see from Figure 3, their actual values seem to converge to 0 more quickly than the upper bound obtained from (12) and (14).

Moreover, we also evaluate the performance of two algorithms based on the new stopping criterion (24), which is formulated from the Polyak-Lojasiewicz condition. This stopping rule

specifies that when the Bregman distance between two consecutive points is sufficiently small, we can terminate our algorithm with assured quality.

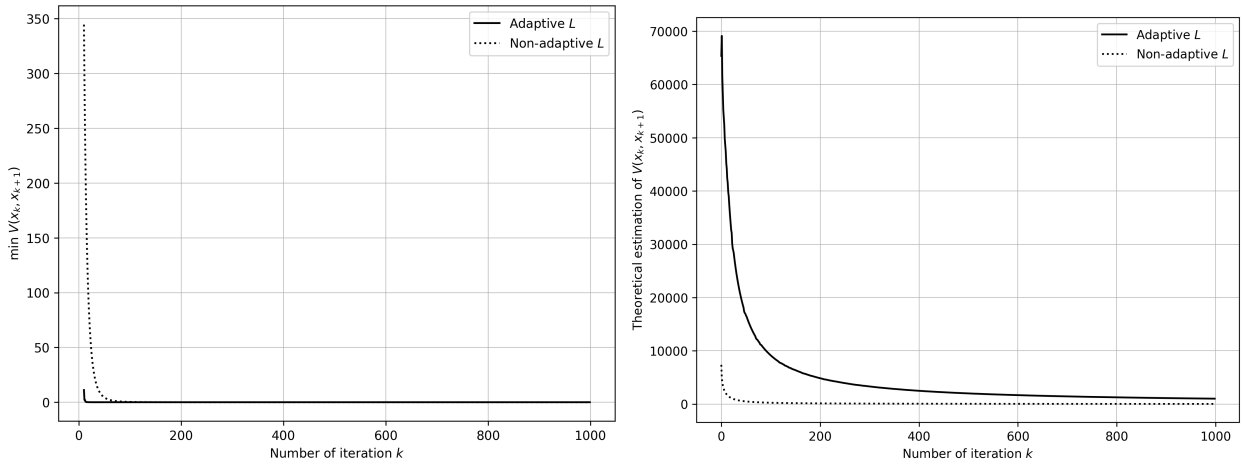
In particular, Table 1 shows the number of iterations required to obtain the value of $V(x_k, x_{k+1})$. Instead of considering the number of iterations to achieve

$$\min_{k=0,\dots,N-1} V(x_k, x_{k+1}) \leq \varepsilon$$

we can simply rewrite it as

$$V(x_N, x_{N+1}) \leq \varepsilon \tag{32}$$

which is the main stopping criterion that we are interested.



(a) (b)

Figure 3: The theoretical estimates of (12) and (14) and the actual values $\min V(x_k, x_{k+1})$ for Algorithm 1 and Algorithm 2.

	Constant L		Adaptive L	
ε, Δ	N	Time (sec)	N	Time (sec)
10^{-1}	133	5.7	19	3.79
10^{-2}	229	10.62	23	6.20
10^{-3}	356	11.98	29	6.81
10^{-4}	497	15.38	42	12.07
10^{-5}	645	19.50	59	16.51
10^{-6}	840	26.26	105	29.60
10^{-7}	1124	33.27	206	54.27

Table 1: The number of iterations required to satisfy the condition (32).

The results from Table 1 show that the adaptive algorithm (Algorithm 2) performs better with fewer iterations and less time when the value of ε used to stop the Mirror Descent algorithm is relatively large. However, as this value becomes smaller, Algorithm 1 requires less time to satisfy the criteria.

6 Conclusion

We studied the Mirror Descent method for the class of relatively smooth functions, which do not require the typical Lipschitz smoothness condition. We introduced the concept of relative smoothness, which is associated with an appropriate distance generating function, and demonstrated the solvability of this class of problems.

We explored several variants of the Mirror Descent algorithm in the presence of gradient noise. In addition to employing a traditional constant step-size, we introduced two adaptive step-size algorithms. We provided theoretical estimates for the upper bounds of the minimum Bregman distance at each iteration and demonstrated that the linear convergence rate is maintained even with inexact gradients. Furthermore, a new stopping rule is proposed within this algorithm to ensure the quality of the termination point.

We conducted numerical experiments on Quadratic Inverse Problems to validate our theoretical results. The experiments highlighted the practical advantages and disadvantages of adaptive versus non-adaptive methods. Our findings indicated that while the adaptive algorithm generally required fewer iterations to converge, the non-adaptive algorithm was faster per iteration. Moreover, if a lower quality of the terminating point is acceptable, the adaptive algorithm performs better with less time and fewer iterations.

In future work, we are interested in expanding our research to explore the Stochastic Gradient Descent (SGD) algorithm on the class of relatively smooth functions.

A Using Cardano's Method to Solve the Depressed Cubic Equation

To solve the problem of finding the root of the depressed cubic equation when $a > 0$, we begin by expressing x in a specific form. This initial transformation is crucial as it allows us to convert the original cubic equation into a simpler quadratic equation. Once we have the quadratic equation, we can apply standard methods for solving quadratic equations to determine the value of x.

$$ax^3 + x - 1 = 0 \tag{33}$$

$$\Leftrightarrow \quad x^3 + \frac{x}{a} - \frac{1}{a} = 0 \tag{34}$$

Let $x = u - v$, then

$$\begin{aligned} x^3 &= (u-v)^3 \\ &= u^3 - 3u^2v + 3uv^2 - v^3 \\ &= -3uv(u-v) + u^3 - v^3 \\ &= -3uvx + u^3 - v^3 \end{aligned}$$

Put the left hand-side to the right, we get

$$x^3 + 3uvx - (u^3 - v^3) = 0$$

Compare it with (34), we obtain

$$3uv = \frac{1}{a} \Leftrightarrow u = \frac{1}{3av} \tag{35}$$

$$u^3 - v^3 = \frac{1}{a} \tag{36}$$

Substitute (35) to (36)

$$\begin{aligned} &\frac{1}{27a^3v^3} - v^3 = \frac{1}{a} \\ \Leftrightarrow \quad &\frac{1}{27a^3} - v^6 = \frac{v^3}{a} \\ \Leftrightarrow \quad &v^6 + \frac{v^3}{a} - \frac{1}{27a^3} = 0 \end{aligned} \tag{37}$$

We obtain the quadratic equation (37). Solving this quadratic equation, we obtain two values of v, but since $a > 0$, we are only interested in the value of v such that $v < u$. Knowing u, v, we can finally obtain the only positive root of (33)

$$x = \sqrt[3]{\frac{\frac{1}{a} + \sqrt{\frac{1}{a^2} + \frac{4}{27a^3}}}{2}} - \sqrt[3]{\frac{\frac{-1}{a} + \sqrt{\frac{1}{a^2} + \frac{4}{27a^3}}}{2}}$$

References

[1] 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, 2017.

[2] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[3] A. Beck and Y. C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[4] 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.

[5] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[6] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

[7] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

[8] E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[9] Y. Censor and S. A. Zenios. Proximal minimization algorithm with d-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464, 1992.

[10] M. Danilova, P. Dvurechensky, A. Gasnikov, E. Gorbunov, S. Guminov, D. Kamzolov, and I. Shibaev. Recent theoretical advances in non-convex optimization. arxiv. *arXiv preprint arXiv:2012.06188*, 2020.

[11] D. L. Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627, 1995.

[12] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] 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.

[15] A. S. Nemirovskij and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.

[16] B. S. B. J. Nesta. A fast and accurate first-order method for sparse recovery, 2009.

[17] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

[18] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[19] Y. Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 186:157–183, 2021.

[20] Y. Nesterov. Superfast second-order methods for unconstrained convex optimization. *Journal of Optimization Theory and Applications*, 191:1–30, 2021.

[21] B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[22] B. T. Polyak. *Introduction to optimization.* Optimization Software, Inc., New York, 1987.

[23] B. T. Polyak, I. A. Kuruzov, and F. S. Stonyakin. Stopping rules for gradient methods for non-convex problems with additive noise in gradient. *arXiv preprint arXiv:2205.07544*, 2022.

[24] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[25] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

[26] F. Stonyakin, A. Titov, M. Alkousa, O. Savchuk, and A. Gasnikov. Adaptive algorithms for relatively lipschitz continuous convex optimization problems. *arXiv preprint arXiv:2107.05765*, 2021.

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

[28] M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):670–690, 1992.