

Universal Intermediate Gradient Method for Convex Problems with Inexact Oracle

Dmitry Kamzolov ^a Pavel Dvurechensky^b and Alexander Gasnikov^{a,c}.

^aMoscow Institute of Physics and Technology, Moscow, Russia;

^bWeierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany;

^cInstitute for Information Transmission Problems RAS, Moscow, Russia.

October 22, 2019

Abstract

In this paper, we propose new first-order methods for minimization of a convex function on a simple convex set. We assume that the objective function is a composite function given as a sum of a simple convex function and a convex function with inexact Hölder-continuous subgradient. We propose Universal Intermediate Gradient Method. Our method enjoys both the universality and intermediateness properties. Following the ideas of Y. Nesterov (Math.Program. 152: 381-404, 2015) on Universal Gradient Methods, our method does not require any information about the Hölder parameter and constant and adjusts itself automatically to the local level of smoothness. On the other hand, in the spirit of the Intermediate Gradient Method proposed by O. Devolder, F.Glineur and Y. Nesterov (CORE Discussion Paper 2013/17, 2013), our method is intermediate in the sense that it interpolates between Universal Gradient Method and Universal Fast Gradient Method. This allows to balance the rate of convergence of the method and rate of the oracle error accumulation. Under additional assumption of strong convexity of the objective, we show how the restart technique can be used to obtain an algorithm with faster rate of convergence.

1 Introduction

In this paper, we consider first-order methods for minimization of a convex function over a simple convex set. The renaissance of such methods started more than ten years ago and was mostly motivated by large-scale problems in data analysis, imaging and machine learning. Simple black-box oriented methods like Mirror Descent [21] or Fast Gradient Method [23], which were known in the 1980s, got a new life.

For a long time algorithms and their analysis were, mostly, separate for two main classes of problems. The first class, with optimal method being Mirror Descent, is the class of non-smooth convex functions with bounded subgradients. The second is the class of smooth

convex functions with Lipschitz-continuous gradient, and the optimal method for this class is Fast Gradient Method. An intermediate class of problems with Hölder-continuous subgradient was also considered and optimal methods for this class were proposed in [20]. However, these methods require to know the Hölder constant. In 2013, Nesterov proposed a Universal Fast Gradient Method [22] which is free of this drawback and is uniformly optimal for the class of convex problems with Hölder-continuous subgradient in terms of black-box information theoretic lower bounds [21]. In 2012, Lan proposed a Fast gradient method with one prox-mapping for stochastic optimization problems [19]. In 2016, Gasnikov and Nesterov proposed a Universal Triangle Method [15], which possesses all the properties of Universal Fast Gradient Method, but uses only one proximal mapping instead of two, as opposed to the previous version. We also mention the work [16], where the authors introduce a method which is uniformly optimal for convex and non-convex problems with Hölder-continuous subgradient, and the work [26], in which a universal primal-dual method is proposed to solve linearly constrained convex problems.

Another line of research [4–7, 11] studies first-order methods with inexact oracle. The considered inexactness can be of deterministic or stochastic nature, it can be connected to inexact calculation of the subgradient or to inexact solution of some auxiliary problem. As it was shown in [6], gradient descent has slower rate of convergence, but does not accumulate the error of the oracle. On the opposite, Fast Gradient Method has faster convergence rate, but accumulates the error linearly with the iteration counter. Later, in [5] an Intermediate Gradient Method was proposed. The main feature of this method is that, depending on the choice of a hyperparameter, it interpolates between Gradient Method and Fast Gradient Method to exploit the trade-off between the rate of convergence and the rate of error accumulation.

In this paper, we join the above two lines of research and present Universal Intermediate Gradient Method (UIGM) for problems with deterministic inexact oracle. Our method enjoys both the universality with respect to smoothness of the problem and interpolates between Universal Gradient Method and Universal Fast Gradient Method, thus, allowing to balance the rate of convergence of the method and rate of the error accumulation. We consider a composite convex optimization problem on a simple set with convex objective, which has inexact Hölder-continuous subgradient, propose a method to solve it, and prove the theorem on its convergence rate. The obtained rate of convergence is uniformly optimal for the considered class of problems. This method can be used in different applications such as transport modeling [1, 12], inverse problems [13] and others.

We also consider the same problem under additional assumption of strong convexity of the objective function and show how the restart technique [10, 14, 18, 20, 21, 23, 25] can be applied to obtain a faster convergence rate of UIGM. The obtained rate of convergence is again optimal for the class of strongly convex functions with Hölder-continuous subgradient.

The rest of the paper is organized as follows. In Sect. 2, we state the problem. After that, in Sect. 3, we present Universal Intermediate Gradient Method and prove a convergence rate theorem with general choice of controlling sequence of coefficients. In Sect. 4, we analyze particular choice of controlling sequence of coefficients and prove a convergence rate theorem

under this choice of coefficients. In Sect. 5, we present UIGM for strongly convex functions and prove convergence rate theorem under this additional assumption. In Sect. 6, we introduce another choice of coefficients that don't need any additional information. In Sect. 7, we present numerical experiments for our method.

2 Problem Statement and Preliminaries

In what follows, we work in a finite-dimensional linear vector space E . Its dual space, the space of all linear functions on E , is denoted by E^* . Relative interior of Q is denoted as $\text{rint } Q$. For $x \in E$ and $s \in E^*$, we denote by $\langle s, x \rangle$ the value of a linear function s at x . For the (primal) space E , we introduce a norm $\|\cdot\|_E$. Then the dual norm is defined in the standard way:

$$\|s\|_{E,*} = \max_{x \in E} \{\langle s, x \rangle : \|x\|_E \leq 1\}.$$

Finally, for a convex function $f : \mathbf{dom } f \rightarrow R$ with $\mathbf{dom } f \subseteq E$ we denote by $\nabla f(x) \in E^*$ one of its subgradients.

We consider the following convex composite optimization problem [24]:

$$\min_{x \in Q} \left[F(x) \stackrel{\text{def}}{=} f(x) + h(x) \right], \quad (1)$$

where Q is a simple closed convex set, $h(x)$ is a simple closed convex function and $f(x)$ is a convex function on Q with inexact first-order oracle, defined below. We assume that problem (1) is solvable with optimal solution x^* .

Definition 1. We say that a convex function $f(x)$ is equipped with a *first-order (δ, L) -oracle* on a convex set Q if for any point $x \in Q$, (δ, L) -oracle returns a pair $(f_\delta(x), g_\delta(x)) \in R \times E^*$ such that

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

In this definition, δ represents the error of the oracle [6]. The oracle is exact with $\delta = 0$. Also we can take $\delta = \delta_c + \delta_u$, where δ_c represents the error, which we can control and make as small as we would like to. On the opposite, δ_u represents the error, which we can not control [7]. Note that, by Definition 1,

$$0 \leq f(x) - f_\delta(x) \leq \delta, \quad \forall x \in Q. \quad (3)$$

To motivate Definition 1, we consider the following example. Let f be a convex function with Hölder-continuous subgradient. Namely, there exists $\nu \in [0, 1]$, and $M_\nu < +\infty$, such that

$$\|\nabla f(x) - \nabla f(y)\|_{E^*} \leq M_\nu \|x - y\|_E^\nu, \quad \forall x, y \in Q.$$

In [6], it was proved that, for such function for any $\delta_c > 0$, if

$$L \geq L(\delta_c) = \left[\frac{1-\nu}{1+\nu} \cdot \frac{1}{2\delta_c} \right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}, \quad (4)$$

then

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|_E^2 + \delta_c, \quad \forall x, y \in Q. \quad (5)$$

We assume also that the set Q is bounded with $\max_{x, y \in Q} \|x - y\|_E \leq D$. Finally, assume that the value and subgradient of f can be calculated only with some known, but uncontrolled error. Strictly speaking, there exist $\bar{\delta}_1, \bar{\delta}_2 > 0$ such that, for any point $x \in Q$, we can calculate approximations $\bar{f}(x)$ and $\bar{g}(x)$ with $|\bar{f}(x) - f(x)| \leq \bar{\delta}_1$ and $\|\bar{g}(x) - \nabla f(x)\|_{E^*} \leq \bar{\delta}_2$.

Let us show that, in this example, f can be equipped with inexact first-order oracle based on the pair $(\bar{f}(x), \bar{g}(x))$, where $f_\delta(x) = \bar{f}(x) - \bar{\delta}_1 - \bar{\delta}_2 D$ and $g_\delta(x) = \bar{g}(x)$.

Now we prove the first inequality from (2)

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle \\ &\geq \bar{f}(x) - \bar{\delta}_1 + \langle \bar{g}(x), y - x \rangle - \bar{\delta}_2 D = f_\delta(x) + \langle g_\delta(x), y - x \rangle \end{aligned}$$

Using inequality (5) we obtain the second inequality from (2), for any $y \in Q$,

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L(\delta_c)}{2} \|x - y\|_E^2 + \delta_c \\ &\leq \bar{f}(x) + \bar{\delta}_1 + \langle \bar{g}(x), y - x \rangle + \langle \nabla f(x) - \bar{g}(x), y - x \rangle + \frac{L(\delta_c)}{2} \|x - y\|_E^2 + \delta_c \\ &\leq f_\delta(x) + \langle g_\delta(x), y - x \rangle + \frac{L(\delta_c)}{2} \|x - y\|_E^2 + 2\bar{\delta}_1 + 2\bar{\delta}_2 D + \delta_c. \end{aligned}$$

Thus, $(f_\delta(x), g_\delta(x))$ is an inexact first-order oracle with $\delta_u = 2\bar{\delta}_1 + 2\bar{\delta}_2 D$, δ_c , and $L(\delta_c)$ given by (4).

To construct our algorithm for problem (1), we introduce, as it is usually done, proximal setup [2], which consists of choosing a norm $\|\cdot\|_E$, and a *prox-function* $d(x)$ which is continuous, convex on Q and

1. $d(x)$ is a continuously differentiable 1-strongly convex on Q with respect to $\|\cdot\|_E$, i.e., for any $x, y \in \text{rint } Q$,

$$d(y) - d(x) - \langle \nabla d(x), y - x \rangle \geq \frac{1}{2} \|y - x\|_E^2.$$

2. Without loss of generality, we assume that

$$\min_{x \in Q} d(x) = 0.$$

Then if $\bar{x} = \operatorname{argmin}_{x \in Q} d(x)$, we get

$$d(y) \geq \frac{1}{2} \|y - \bar{x}\|_E^2, \quad \forall y \in Q. \quad (6)$$

The corresponding *Bregman divergence* is defined as $V(x,y) = d(y) - d(x) - \langle \nabla d(x), y - x \rangle$ and satisfies

$$V(x,y) \geq \frac{1}{2} \|x - y\|_E^2, \quad \forall x,y \in Q. \quad (7)$$

We use prox-function in so called *composite prox-mapping*, which consists in solving auxiliary problem

$$\min_{x \in Q} \{ \langle g, x \rangle + d(x) + h(x) \}, \quad (8)$$

where $g \in E^*$ is given. We allow this problem to be solved inexactly in the following sense.

Definition 2. Assume that $\delta_p > 0$, $g \in E^*$ are given. We call a point $\tilde{x} = \tilde{x}(g, \delta_p) \in \text{rint } Q$ an *inexact composite prox-mapping* iff we can calculate \tilde{x} and there exists $p \in \partial h(\tilde{x})$ s.t. it holds that

$$\langle g + \nabla d(\tilde{x}) + p, u - \tilde{x} \rangle \geq -\delta_p, \quad \forall u \in Q. \quad (9)$$

We denote by

$$\tilde{x} = \underset{x \in Q}{\text{argmin}}^{\delta_p} \{ \langle g, x \rangle + d(x) + h(x) \}.$$

one of the possible inexact composite prox-mapping.

Note that if \tilde{x} is an exact solution of (8), inequality (9) holds with $\delta_p = 0$ due to first-order optimality condition.

We also use the following auxiliary fact

Lemma 2.1. (*Lemma 5.5.1 in [2]*) Let $F : Q \rightarrow \mathbb{R} \cup \{+\infty\}$ be a convex function such that $\Psi(x) = F(x) + d(x)$ is closed and convex on Q . Denote $\tilde{x} = \underset{x \in Q}{\text{argmin}}^{\delta_p} \Psi(x)$. Then

$$\Psi(y) \geq \Psi(\tilde{x}) + V(\tilde{x}, y) - \delta_p, \quad \forall y \in Q. \quad (10)$$

Hence, from (7)

$$\Psi(y) \geq \Psi(\tilde{x}) - \delta_p, \quad \forall y \in Q. \quad (11)$$

3 Universal Intermediate Gradient Method

In this section, we describe a general scheme of Universal Intermediate Gradient Method (UIGM) and prove general convergence rate. This scheme is based on two sequences α_k, B_k , $k \geq 0$. From now on, we assume that these sequences satisfy, for all $k \geq 0$,

$$0 < \alpha_{k+1} \leq B_{k+1} \leq A_k + \alpha_{k+1}, \quad (12)$$

where the sequence A_k is defined by recurrence $A_{k+1} = A_k + \alpha_{k+1}$. Particular choice of these two sequences and its consequence for the convergence rate are discussed in the next section.

For Algorithm 1 we combine Algorithm 2 from [5] with Algorithm 1 from [15] to get IGM with only one prox-mapping instead two as in [5]. After that we improve this method by techniques from [22] to get UIGM with exact prox-mapping. Last generalization use Lemma 5.5.1 from [2]. As a result we get algorithm that works in wide class of problems, adaptive and don't need to know exact Hölder and Lipschitz constants, uses only one prox-mapping and correctly work with errors of oracle and prox-mapping.

Algorithm 1 Universal Intermediate Gradient Method (UIGM)

Require: $\varepsilon > 0$ – desired accuracy, δ_u – uncontrolled oracle error, δ_p – prox-mapping error, L_s – initial guess for the Hölder constant, α_k – choose by some policy, for example (34).

1: Set $\delta_0 = \frac{\varepsilon}{4} + \delta_u$,

$$z_0 = x_0 = \operatorname{argmin}_{x \in Q}^{\delta_p} d(x), \quad (13)$$

2: Set $i_0 = 0$

3: Compute

$$y_0 = \operatorname{argmin}_{x \in Q}^{\delta_p} \left\{ d(x) + (2^{i_0} L_s)^{-1} [\langle g_{\delta_0}(x_0), x - x_0 \rangle + h(x)] \right\}. \quad (14)$$

4: If

$$f_{\delta_0}(y_0) \leq f_{\delta_0}(x_0) + \langle g_{\delta_0}(x_0), y_0 - x_0 \rangle + \frac{2^{i_0} L_s}{2} \|y_0 - x_0\|_E^2 + \delta_0, \quad (15)$$

go to Step 5. Otherwise, set $i_0 = i_0 + 1$ and go back to Step 3.

5: Define $L_0 = 2^{i_0} L_s$, $\alpha_0 = B_0 = A_0 = (L_0)^{-1}$.

6: **for** $k = 1, \dots$ **do**

7: Set $i_k = 0$.

8: Set $L_k = 2^{i_k} L_{k-1}$ and $\alpha_k = \alpha(L_k)$ by some policy, for example (34),

$$B_k = \alpha_k^2 L_k, \quad (16)$$

$$\delta_k = \frac{\alpha_k \varepsilon}{B_k 4} + \delta_u \quad (17)$$

$$x_k = \frac{\alpha_k}{B_k} z_{k-1} + \frac{B_k - \alpha_k}{B_k} y_{k-1}. \quad (18)$$

$$z_k = \operatorname{argmin}_{x \in Q}^{\delta_p} \left\{ d(x) + \sum_{j=0}^k \alpha_j [\langle g_{\delta_j}(x_j), x - x_j \rangle + h(x)] \right\}, \quad (19)$$

$$w_k = \frac{\alpha_k}{B_k} z_k + \frac{B_k - \alpha_k}{B_k} y_{k-1}. \quad (20)$$

9: If

$$f_{\delta_k}(w_k) \leq f_{\delta_k}(x_k) + \delta_k + \langle g_{\delta_k}(x_k), w_k - x_k \rangle + \frac{L_k}{2} \|w_k - x_k\|_E^2. \quad (21)$$

go to Step 10. Otherwise, set $i_k = i_k + 1$ and go back to Step 8.

10: Set

$$A_k = A_{k-1} + \alpha_k, \quad (22)$$

$$y_k = \frac{B_k}{A_k} w_k + \frac{A_k - B_k}{A_k} y_{k-1}. \quad (23)$$

The next theorem gives an upper bound for $A_k F(y_k)$. Its proof is an adaptation of the proof of Lemma 1 in [5] and Theorem 3 in [22].

Theorem 3.1. *Let f be a convex function with inexact first-order oracle, the dependence $L(\delta_c)$ being given by (4). Then all iterations of UIGM are well defined and, for all $k \geq 0$ we have*

$$A_k F(y_k) - E_k \leq \Psi_k^*, \quad (24)$$

where $E_k = 2 \left(\sum_{j=0}^k B_j \right) \delta_u + (2k+1)\delta_p + A_k \frac{\varepsilon}{2}$,

$$\Psi_k^* = \min_{x \in Q} \left\{ \Psi_k(x) = d(x) + \sum_{j=0}^k \alpha_j [f_{\delta_j}(x_j) + \langle g_{\delta_j}(x_j), x - x_j \rangle + h(x)] \right\}. \quad (25)$$

Proof. Let us prove first, that the "line-search" process of steps 6–9 is finite. By (4), (5), if $2^{i_k} L_{k-1} \geq L \left(\frac{\alpha_k \varepsilon}{B_k 4} \right)$, from (21) and (3), we get

$$f_{\delta_k}(w_k) - \delta_k \leq f(w_k) \leq f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), w_k - x_k \rangle + \frac{2^{i_k} L_{k-1}}{2} \|w_k - x_k\|_E^2 + \delta_k$$

and the stopping criterion in the inner cycle holds. Thus, we need to show that

$$2^{i_k} L_{k-1} \geq \left[\frac{\alpha_k}{B_k} \varepsilon \right]^{-\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \quad (26)$$

for i_k large enough. Indeed,

$$2^{i_k} L_{k-1} \left[\frac{\alpha_k}{B_k} \right]^{\frac{1-\nu}{1+\nu}} \stackrel{(16)}{=} \frac{B_k}{\alpha_k^2} \left[\frac{\alpha_k}{B_k} \right]^{\frac{1-\nu}{1+\nu}} = \left[\frac{B_k}{\alpha_k} \right]^{\frac{2\nu}{1+\nu}} \frac{1}{\alpha_k} \stackrel{(12)}{\geq} \frac{1}{\alpha_k}.$$

It remains to prove that $\alpha_k \rightarrow 0$ as $i_k \rightarrow \infty$.

$$\begin{aligned} \alpha_k^2 &= \frac{B_k}{2^{i_k} L_{k-1}} \stackrel{(12)}{\leq} \frac{A_k}{2^{i_k} L_{k-1}} \stackrel{(22)}{=} \frac{A_{k-1} + \alpha_k}{2^{i_k} L_{k-1}}, \\ \Rightarrow \quad \alpha_k^2 - \frac{\alpha_k}{2^{i_k} L_{k-1}} - \frac{A_{k-1}}{2^{i_k} L_{k-1}} &\leq 0. \end{aligned} \quad (27)$$

Thus, $\alpha_k \in [\alpha_k^-, \alpha_k^+]$, where α_k^- and α_k^+ are the solutions of

$$\alpha_k^2 - \frac{\alpha_k}{2^{i_k} L_{k-1}} - \frac{A_{k-1}}{2^{i_k} L_{k-1}} = 0.$$

The solutions are

$$\begin{aligned} \alpha_k^- &= \frac{1}{2^{i_k+1} L_{k-1}} - \left(\frac{1}{4^{i_k+1} L_{k-1}^2} + \frac{A_{k-1}}{2^{i_k} L_{k-1}} \right)^{1/2}, \\ \alpha_k^+ &= \frac{1}{2^{i_k+1} L_{k-1}} + \left(\frac{1}{4^{i_k+1} L_{k-1}^2} + \frac{A_{k-1}}{2^{i_k} L_{k-1}} \right)^{1/2}. \end{aligned}$$

Now from (27) we have that $\alpha_k^- \leq \alpha_k \leq \alpha_k^+$. From $\alpha_k^- \rightarrow 0$, $\alpha_k^+ \rightarrow 0$ as $i_k \rightarrow \infty$ we get $\alpha_k \rightarrow 0$.

Let us prove relation (24). For $k = 0$:

$$\begin{aligned}
\Psi_0^* &\stackrel{(25)}{=} \min_{x \in Q} \{d(x) + \alpha_0 f_{\delta_0}(x_0) + \alpha_0 \langle g_{\delta_0}, x - x_0 \rangle + \alpha_0 h(x)\} \\
&\stackrel{(11),(14)}{\geq} d(y_0) + \alpha_0 f_{\delta_0}(x_0) + \alpha_0 \langle g_{\delta_0}(x_0), y_0 - x_0 \rangle + \alpha_0 h(y_0) - \delta_p \\
&\stackrel{(6),(13)}{\geq} \alpha_0 \left(\frac{1}{2\alpha_0} \|y_0 - x_0\|_E^2 + f_{\delta_0}(x_0) + \langle g_{\delta_0}(x_0), y_0 - x_0 \rangle + h(y_0) \right) - \delta_p \\
&= \alpha_0 \left(\frac{2^{i_0} L_s}{2} \|y_0 - x_0\|_E^2 + f_{\delta_0}(x_0) + \langle g_{\delta_0}(x_0), y_0 - x_0 \rangle + h(y_0) \right) - \delta_p \\
&\stackrel{(15)}{\geq} \alpha_0 \left(f_{\delta_0}(y_0) - \frac{\varepsilon}{4} - \delta_u + h(y_0) \right) - \delta_p \\
&\stackrel{(3)}{\geq} \alpha_0 \left(f(y_0) - \frac{\varepsilon}{2} - 2\delta_u + h(y_0) \right) - \delta_p = A_0 F(y_0) - E_0.
\end{aligned}$$

Assume that (24) is valid for certain $k - 1 \geq 0$. We now prove that it holds for k .

$$\begin{aligned}
\Psi_k^* &\stackrel{(25)}{=} \min_{x \in Q} \Psi_k(x) \stackrel{(11),(19)}{\geq} \Psi_k(z_k) - \delta_p \\
&\stackrel{(25)}{=} \Psi_{k-1}(z_k) + \alpha_k [f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle + h(z_k)] - \delta_p \\
&\stackrel{(10)}{\geq} \Psi_{k-1}(z_{k-1}) + V(z_{k-1}, z_k) - 2\delta_p + \alpha_k [f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle + h(z_k)] \\
&\stackrel{(7)}{\geq} \Psi_{k-1}^* + \frac{1}{2} \|z_k - z_{k-1}\|_E - 2\delta_p + \alpha_k [f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle + h(z_k)] \\
&\stackrel{(24)}{\geq} A_{k-1} F(y_{k-1}) - E_{k-1} + \frac{1}{2} \|z_k - z_{k-1}\|_E - 2\delta_p \\
&\quad + \alpha_k [f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle + h(z_k)] \\
&= (A_k - B_k) F(y_{k-1}) - E_{k-1} + \frac{1}{2} \|z_k - z_{k-1}\|_E - 2\delta_p + (B_k - \alpha_k) f(y_{k-1}) \\
&\quad + (B_k - \alpha_k) h(y_{k-1}) + \alpha_k h(z_k) + \alpha_k [f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle] \\
&\stackrel{(20)}{\geq} (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) + \frac{1}{2} \|z_k - z_{k-1}\|_E - 2\delta_p \\
&\quad + (B_k - \alpha_k) f(y_{k-1}) + \alpha_k [f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle]
\end{aligned}$$

$$\begin{aligned}
&\geq (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) + \frac{1}{2} \|z_k - z_{k-1}\|_E - 2\delta_p \\
&+ (B_k - \alpha_k) (f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), y_{k-1} - x_k \rangle) + \alpha_k [f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle + h(z_k)] \\
&= (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) + \frac{1}{2} \|z_k - z_{k-1}\|_E + B_k f_{\delta_k}(x_k) \\
&+ \langle g_{\delta_k}(x_k), (B_k - \alpha_k)(y_{k-1} - x_k) + \alpha_k(z_k - x_k) \rangle - 2\delta_p.
\end{aligned}$$

From (18), we have

$$(B_k - \alpha_k)(y_{k-1} - x_k) + \alpha_k(z_k - x_k) = \alpha_k(z_k - z_{k-1}).$$

Therefore,

$$\begin{aligned}
\Psi_k^* &\geq (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p \\
&+ B_k f_{\delta_k}(x_k) + \alpha_k \langle g_{\delta_k}(x_k), (z_k - z_{k-1}) \rangle + \frac{1}{2} \|z_k - z_{k-1}\|_E \\
&= (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p \\
&+ B_k \left[f_{\delta_k}(x_k) + \frac{\alpha_k}{B_k} \langle g_{\delta_k}(x_k), z_k - z_{k-1} \rangle + \frac{1}{2B_k} \|z_k - z_{k-1}\|_E^2 \right].
\end{aligned}$$

As $B_k = 2^{i_k} L_{k-1} \alpha_k^2$, we have $\frac{1}{B_k} = 2^{i_k} L_{k-1} \frac{\alpha_k^2}{B_k^2}$ and, therefore,

$$\begin{aligned}
\Psi_k^* &\geq (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p \\
&+ B_k \left[f_{\delta_k}(x_k) + \frac{\alpha_k}{B_k} \langle g_{\delta_k}(x_k), z_k - z_{k-1} \rangle + \frac{2^{i_k} L_{k-1} \alpha_k^2}{2B_k^2} \|z_k - z_{k-1}\|_E^2 \right].
\end{aligned}$$

But

$$\frac{\alpha_k}{B_k} (z_k - z_{k-1}) \stackrel{(18),(20)}{=} w_k - x_k,$$

and we obtain

$$\begin{aligned}
\Psi_k^* &\geq (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p \\
&+ B_k \left[f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), w_k - x_k \rangle + \frac{2^{i_k} L_{k-1}}{2} \|w_k - x_k\|_E^2 \right] \\
&\stackrel{(21)}{\geq} (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p + B_k \left[f_{\delta_k}(w_k) - \frac{\alpha_k \varepsilon}{B_k 4} - \delta_u \right] \\
&\stackrel{(3)}{\geq} (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p + B_k \left[f(w_k) - \frac{\alpha_k \varepsilon}{B_k 2} - 2\delta_u \right] \\
&\stackrel{(23)}{\geq} A_k F(y_k) - E_{k-1} - B_k \left[\frac{\alpha_k \varepsilon}{B_k 2} + 2\delta_u \right] - 2\delta_p \\
&= A_k F(y_k) - E_k.
\end{aligned}$$

□

We are in position to establish the relation between the rate of growth of $\{A_k\}_{k=0}^{+\infty}$ with rate of convergence of UIGM. The proof of the next result is an adaptation of Theorem 2 in [5].

Corollary 3.2. *Let f be a convex function with inexact first-order oracle, the dependence $L(\delta_c)$ being given by (4). Then all iterations of UIGM are well defined and, for all $k \geq 0$, we have*

$$F(y_k) - F^* \leq \frac{d(x^*)}{A_k} + \frac{2\delta_u}{A_k} \sum_{j=0}^k B_j + \frac{(2k+1)\delta_p}{A_k} + \frac{\varepsilon}{2}. \quad (28)$$

Proof.

$$\begin{aligned} \Psi_k^* &= \min_{x \in Q} \left\{ d(x) + \sum_{j=0}^k \alpha_j [f_{\delta_j}(x_j) + \langle g_{\delta_j}(x_j), x - x_j \rangle + h(x)] \right\} \\ &\leq d(x^*) + \sum_{j=0}^k \alpha_j [f_{\delta_j}(x_j) + \langle g_{\delta_j}(x_j), x^* - x_j \rangle + h(x^*)] \\ &\stackrel{(2)}{\leq} d(x^*) + \sum_{j=0}^k \alpha_j [f(x^*) + h(x^*)] = d(x^*) + A_k F(x^*). \end{aligned}$$

By (24), we have $A_k F(y_k) - E_k \leq d(x^*) + A_k F(x^*)$ and so

$$F(y_k) - F^* \leq \frac{d(x^*)}{A_k} + \frac{E_k}{A_k} \leq \frac{d(x^*)}{A_k} + \frac{2\delta_u}{A_k} \sum_{j=0}^k B_j + \frac{(2k+1)\delta_p}{A_k} + \frac{\varepsilon}{2}.$$

□

Similarly as UFGM [22], UIGM can be equipped with an implementable stopping criterion. Assume that we know an upper bound D for the distance to the solution from the starting point $V(x_0, x^*) = d(x^*) \leq D$. Denote $l_k^{pd}(x) = \sum_{j=0}^k \alpha_j [f_{\delta_j}(x_j) + \langle g_{\delta_j}(x_j), x - x_j \rangle]$

and

$$\begin{aligned} \tilde{F}_k &= \min_{x \in Q} \left\{ \frac{1}{A_k} l_k^{pd}(x) + h(x) : d(x) \leq D \right\} \\ &= \min_{x \in Q} \max_{\beta \geq 0} \left\{ \frac{1}{A_k} l_k^{pd}(x) + h(x) + \beta(d(x) - D) \right\} \\ &= \max_{\beta \geq 0} \min_{x \in Q} \left\{ \frac{1}{A_k} l_k^{pd}(x) + h(x) + \beta(d(x) - D) \right\} \\ &\stackrel{\beta=1/A_k}{\geq} \frac{1}{A_k} \Psi_k^* - \frac{1}{A_k} D. \end{aligned} \quad (29)$$

Note that by the first inequality from (2) we get $\tilde{F}_k \leq F^*$. Then

$$F(y_k) - F^* \leq F(y_k) - \tilde{F}_k \stackrel{(24),(29)}{\leq} \frac{D}{A_k} + \frac{E_k}{A_k}$$

Thus, we can use stopping criterion

$$F(y_k) - \tilde{F}_k \leq \varepsilon + \frac{2\delta_u}{A_k} \sum_{j=0}^k B_j + \frac{(2k+1)\delta_p}{A_k}, \quad (30)$$

which ensures

$$F(y_k) - F^* \leq \varepsilon + \frac{2\delta_u}{A_k} \sum_{j=0}^k B_j + \frac{(2k+1)\delta_p}{A_k}, \quad (31)$$

as far as

$$A_k \geq \frac{2D}{\varepsilon}. \quad (32)$$

At the end we get an upper bound of the total number of oracle calls for UIGM with stopping criterion (30) to get an approximate solution of problem (1) satisfying (31).

Denote by $N(k)$ the total number of oracle calls after k iterations (without 0 iteration). We don't take 0 into account because it is some constant that depends on initial guess L_s . At each iteration we call oracle at points x_m and w_m and do it $(i_m + 1)$ times. Then total number of oracle calls per iteration equal to $2(i_m + 1)$. Note that $L_m = 2^{i_m} L_{m-1}$. Therefore, $i_m = \log_2 \frac{L_m}{L_{m-1}}$. Hence,

$$\begin{aligned} N(k) &= \sum_{m=1}^k 2(i_m + 1) = \sum_{m=1}^k 2(\log_2 \frac{L_m}{L_{m-1}} + 1) = \sum_{m=1}^k [2 + 2(\log_2 L_m - \log_2 L_{m-1})] \\ &= 2k + 2\log_2 L_k - 2\log_2 L_0. \end{aligned} \quad (33)$$

Note that

$$\begin{aligned} \frac{B_k}{2\alpha_k^2} &\stackrel{(16)}{=} 2^{i_k} L_{k-1} = L_k \stackrel{(4),(26)}{\leq} \left[\frac{1-\nu}{1+\nu} \frac{1}{\alpha_k \varepsilon} \right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} = \left[\frac{B_k(1-\nu)}{\varepsilon \alpha_k (1+\nu)} \right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}, \\ \Rightarrow \alpha_k^{\frac{-1-3\nu}{1+\nu}} &\leq 2B_k^{\frac{-2\nu}{1+\nu}} \left[\frac{M_\nu^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{1-\nu}{1+\nu}}} \left(\frac{1-\nu}{1+\nu} \right)^{\frac{1-\nu}{1+\nu}} \right]. \end{aligned}$$

Therefore,

$$L_k \leq \frac{B_k}{2} \left(2B_k^{\frac{-2\nu}{1+\nu}} \left[\frac{M_\nu^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{1-\nu}{1+\nu}}} \left(\frac{1-\nu}{1+\nu} \right)^{\frac{1-\nu}{1+\nu}} \right] \right)^{\frac{2(1+\nu)}{1+3\nu}} \leq 2^{\frac{1-\nu}{1+3\nu}} A_k^{\frac{1-\nu}{1+3\nu}} \left[\frac{M_\nu^{\frac{4}{1+3\nu}}}{\varepsilon^{\frac{2-2\nu}{1+3\nu}}} \left(\frac{1-\nu}{1+\nu} \right)^{\frac{2-2\nu}{1+3\nu}} \right].$$

Note that (31) holds if (32) holds. Thus, we can assume that, during the iterations,

$$A_k \leq \frac{2D}{\varepsilon}, \quad k \geq 0.$$

Hence,

$$L_{k+1} \leq 2^{\frac{1-\nu}{1+3\nu}} \left(\frac{2D}{\varepsilon} \right)^{\frac{1-\nu}{1+3\nu}} \left[\frac{M_\nu^{\frac{4}{1+3\nu}}}{\varepsilon^{\frac{2-2\nu}{1+3\nu}}} \left(\frac{1-\nu}{1+\nu} \right)^{\frac{2-2\nu}{1+3\nu}} \right].$$

Substituting this estimate in the expression (33), we obtain that on average UIGM has approximately two calls of oracle per iteration.

4 Power policy

In this section, we present particular choice of the two sequences of coefficients $\{\alpha_k\}_{k \geq 0}$ and $\{B_k\}_{k \geq 0}$. As it was done in [5], these sequences depend on a parameter $p \in [1, 2]$. In our case, the value $p = 1$ corresponds to Universal Dual Gradient Method, and the value $p = 2$ corresponds to Universal Fast Gradient Method. For the smooth case, namely $\nu = 1$, the method in [5] has convergence rate

$$F(y_k) - F^* \leq \Theta \left(\frac{d(x^*)}{k^p} \right) + \Theta(k^{p-1} \delta_u),$$

where $p \in [1, 2]$. Our goal to obtain convergence rate for the whole segment $\nu \in [0, 1]$ and get the above rate of convergence as a special case.

Given a value $p \in [1, 2]$, we choose sequences $\{\alpha_k\}_{k \geq 0}$ and $\{B_k\}_{k \geq 0}$ to be given by

$$\alpha_k = \frac{\left(\frac{k+2p}{2p} \right)^{p-1}}{2^{i_k}} L_{k-1}, \quad k \geq 0 \tag{34}$$

and, in accordance to (16),

$$B_k = \frac{\left(\frac{k+2p}{2p} \right)^{2p-2}}{2^{i_k}} L_{k-1}, \quad k \geq 0. \tag{35}$$

Now we should prove, that power policy can be used in UIGM.

Lemma 4.1. *Assume that f is a convex function with inexact first-order oracle. Then, the sequences $\{\alpha_k\}_{k \geq 0}$ and $\{B_k\}_{k \geq 0}$ given in (34) and (35), respectively, satisfy (12).*

Proof. From (34) we get that $\alpha_k > 0$ for $k \geq 0$. To prove that $\alpha_k \leq B_k$ for $k \geq 0$, we use (34), (35) and that $p \in [0, 1]$

$$\alpha_k = \frac{\left(\frac{k+2p}{2p} \right)^{p-1}}{2^{i_k} L_{k-1}} = \frac{\left(\frac{k}{2p} + 1 \right)^{p-1}}{2^{i_k} L_{k-1}} \leq \frac{\left(\frac{k}{2p} + 1 \right)^{2p-2}}{2^{i_k} L_{k-1}} = B_k.$$

Proof of $A_k \geq B_k$. For $k = 0$ it's correct by definition. Assume that $A_k \geq B_k$ is valid for certain $k - 1 \geq 0$. We now prove that it holds for k . For $m \in [0,1]$ and $x, y \geq 0$ function $f(x, y) = x^m + y^m - (x + y)^m$ has minimal value greater or equal to 0, hence,

$$\begin{aligned}
& x^m + y^m - (x + y)^m \geq 0, \\
\Rightarrow & x^m + y^m \geq (x + y)^m, \\
\Rightarrow & x^{p-1} + y^{p-1} \stackrel{m=p-1}{\geq} (x + y)^{p-1}, \quad p \in [1, 2], \\
\Rightarrow & ((k - 1 + 2p)^2)^{p-1} + (2(k + 2p))^{p-1} \geq ((k - 1 + 2p)^2 + 2(k + 2p))^{p-1}, \\
\Rightarrow & (k - 1 + 2p)^{2(p-1)} + (2p(k + 2p))^{p-1} \geq ((k - 1 + 2p)^2 + 2(k - 1 + 2p) + 2)^{p-1}, \\
\Rightarrow & (k - 1 + 2p)^{2(p-1)} + (2p(k + 2p))^{p-1} \geq (k + 2p)^{2(p-1)}, \\
\Rightarrow & \left(\frac{k - 1 + 2p}{2p}\right)^{2(p-1)} + \left(\frac{k + 2p}{2p}\right)^{p-1} \geq \left(\frac{k + 2p}{2p}\right)^{2(p-1)}, \\
& \Rightarrow \left(\frac{k - 1 + 2p}{2p}\right)^{2(p-1)} \geq \left(\frac{k + 2p}{2p}\right)^{2(p-1)} - \left(\frac{k + 2p}{2p}\right)^{p-1}, \\
& \Rightarrow \frac{\left(\frac{k-1+2p}{2p}\right)^{2(p-1)}}{L_{k-1}} \geq \frac{\left(\frac{k+2p}{2p}\right)^{p-1} \left(\left(\frac{k+2p}{2p}\right)^{p-1} - 1\right)}{L_{k-1}}, \\
& \Rightarrow \frac{\left(\frac{k-1+2p}{2p}\right)^{2(p-1)}}{L_{k-1}} \geq \frac{\left(\frac{k+2p}{2p}\right)^{p-1} \left(\left(\frac{k+2p}{2p}\right)^{p-1} - 1\right)}{2^{i_k} L_{k-1}}, \\
& \Rightarrow \frac{\left(\frac{k-1+2p}{2p}\right)^{2(p-1)}}{L_{k-1}} + \frac{\left(\frac{k+2p}{2p}\right)^{p-1}}{2^{i_k} L_{k-1}} \geq \frac{\left(\frac{k+2p}{2p}\right)^{2(p-1)}}{2^{i_k} L_{k-1}}, \\
& \Rightarrow B_{k-1} + \alpha_k \stackrel{(34),(35)}{\geq} B_k, \\
& \Rightarrow A_{k-1} + \alpha_k \geq B_k, \\
& \Rightarrow A_k \stackrel{(22)}{\geq} B_k.
\end{aligned}$$

□

Now we can obtain the rate of growth of $\{A_k\}_k = 0^{+\infty}$. Combining this rate with Corollary 3.2, we get the explicit rate of convergence of UIGM under the power policy (34).

Theorem 4.2. *Assume that f is a convex function with inexact first-order oracle, the dependence $L(\delta_c)$ being given by (4). Then, for the sequences (34) and (35), for all $k \geq 0$,*

$$F(y_k) - F^* \leq \inf_{\nu \in [0, 1]} \left(\frac{16M_\nu^{\frac{2}{1+\nu}} d(x^*)}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2p\nu-\nu+1}{1+\nu}}} + \frac{32M_\nu^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2\nu(p-1)}{1+\nu}}} \delta_p \right) + 4k^{p-1} \delta_u + \frac{\varepsilon}{2}. \quad (36)$$

Proof. The proof is divided into three steps. First, we prove a lower bound for α_m and A_m . Then, we prove upper bound for B_m . Finally, we use these bounds in Corollary 3.2 and obtain (36).

Lower bound for α_m and A_m . Since the inner cycle of UIGM for sure ends when $2^{i_m} L_{m-1} > L(\delta_m)$, we have $2^{i_m} L_{m-1} \leq 2L(\delta_m)$. Hence,

$$\begin{aligned}
2^{i_m} L_{m-1} &\stackrel{(4),(26)}{\leq} 2 \left[\frac{\alpha_m}{B_m} \varepsilon \right]^{-\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \leq 2 \left[\frac{\left(\frac{m+2p}{2p} \right)^{p-1}}{\varepsilon} \right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}, \\
\Rightarrow \alpha_m &= \frac{\left(\frac{m+2p}{2p} \right)^{p-1}}{2^{i_m} L_{m-1}} \geq \frac{\left(\frac{m+2p}{2p} \right)^{\frac{2p\nu-2\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2M_\nu^{\frac{2}{1+\nu}}}, \\
\Rightarrow A_k &= \sum_{m=0}^k \alpha_m \geq \sum_{m=0}^k \frac{\left(\frac{m+2p}{2p} \right)^{\frac{2p\nu-2\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2M_\nu^{\frac{2}{1+\nu}}} \geq \frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{2M_\nu^{\frac{2}{1+\nu}}} \sum_{m=0}^k \left(\frac{m+2p}{2p} \right)^{\frac{2p\nu-2\nu}{1+\nu}}.
\end{aligned}$$

Since

$$\begin{aligned}
\sum_{m=0}^k \left(\frac{m+2p}{2p} \right)^{\frac{2p\nu-2\nu}{1+\nu}} &\geq \int_0^k \left(\frac{x+2p}{2p} \right)^{\frac{2p\nu-2\nu}{1+\nu}} dx + \alpha_0 \\
&\geq \frac{2p(1+\nu)}{2p\nu-\nu+1} \left(\frac{k+2p}{2p} \right)^{\frac{2p\nu-\nu+1}{1+\nu}} \geq 2 \left(\frac{k+2p}{2p} \right)^{\frac{2p\nu-\nu+1}{1+\nu}},
\end{aligned}$$

we have

$$A_k = \sum_{m=0}^k \alpha_m \geq \frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{M_\nu^{\frac{2}{1+\nu}}} \left(\frac{k+2p}{2p} \right)^{\frac{2p\nu-\nu+1}{1+\nu}} \geq \frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{M_\nu^{\frac{2}{1+\nu}}} \left(\frac{k+2}{4} \right)^{\frac{2p\nu-\nu+1}{1+\nu}}. \quad (37)$$

Upper bound for B_m .

$$\begin{aligned}
B_m &\stackrel{(34),(35)}{=} \left(\frac{m+2p}{2p} \right)^{p-1} \alpha_m, \\
\sum_{m=0}^k B_m &= \sum_{m=0}^k \left(\frac{m+2p}{2p} \right)^{p-1} \alpha_m.
\end{aligned}$$

Therefore,

$$\sum_{m=0}^k B_m \leq \left(\frac{k+2p}{2p} \right)^{p-1} A_k. \quad (38)$$

Proof of (36). Now using (28),(37) and (38) we can get convergence rate.

$$\begin{aligned}
F(y_k) - F^* &\stackrel{(28)}{\leq} \frac{d(x^*)}{A_k} + \frac{2\delta_u}{A_k} \sum_{i=0}^k B_i + \frac{\varepsilon}{2} + \frac{(2k+1)\delta_p}{A_k} \\
&\stackrel{(37)}{\leq} \frac{4^{\frac{2p\nu-\nu+1}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} d(x^*)}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2p\nu-\nu+1}{1+\nu}}} + \frac{2\delta_u}{A_k} \sum_{i=0}^k B_i + \frac{4^{\frac{2p\nu-\nu+1}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} (2k+1)}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2p\nu-\nu+1}{1+\nu}}} \delta_p + \frac{\varepsilon}{2} \\
&\leq \frac{16M_\nu^{\frac{2}{1+\nu}} d(x^*)}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2p\nu-\nu+1}{1+\nu}}} + \frac{2\delta_u}{A_k} \sum_{i=0}^k B_i + \frac{32M_\nu^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2\nu(p-1)}{1+\nu}}} \delta_p + \frac{\varepsilon}{2} \\
&\stackrel{(38)}{\leq} \frac{16M_\nu^{\frac{2}{1+\nu}} d(x^*)}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2p\nu-\nu+1}{1+\nu}}} + 2\delta_u \left(\frac{k+2p}{2p} \right)^{p-1} + \frac{32M_\nu^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2\nu(p-1)}{1+\nu}}} \delta_p + \frac{\varepsilon}{2} \\
&\leq \frac{16M_\nu^{\frac{2}{1+\nu}} d(x^*)}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2p\nu-\nu+1}{1+\nu}}} + 4k^{p-1}\delta_u + \frac{32M_\nu^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2\nu(p-1)}{1+\nu}}} \delta_p + \frac{\varepsilon}{2}.
\end{aligned}$$

Since UIGM does not use ν as a parameter, we get

$$F(y_k) - F^* \leq \inf_{\nu \in [0,1]} \left(\frac{16M_\nu^{\frac{2}{1+\nu}} d(x^*)}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2p\nu-\nu+1}{1+\nu}}} + \frac{32M_\nu^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{1-\nu}{1+\nu}} (k+2)^{\frac{2\nu(p-1)}{1+\nu}}} \delta_p \right) + 4k^{p-1}\delta_u + \frac{\varepsilon}{2}.$$

□

Corollary 4.3. *At each iteration $m \geq 0$ of UIGM with sequences $\{\alpha_m\}$, $\{B_m\}$, $m \geq 0$ chosen in accordance with (34) and (35), we have, for any $p \in [1,2]$,*

$$\delta_m = O\left(\frac{\varepsilon}{m^{p-1}}\right) + \delta_u.$$

Proof. By (17), we have

$$\delta_m - \delta_u = \frac{\varepsilon\alpha_m}{4B_m} = \frac{\varepsilon}{4\left(\frac{m+2p}{2p}\right)^{p-1}} = O\left(\frac{\varepsilon}{(m)^{p-1}}\right).$$

□

From the rate of convergence (36) and the fact that UIGM does not include ν as a parameter, we get the following estimation for the number of iterations, which are necessary for getting first term of (36) smaller than $\varepsilon/6$ we need:

$$N = O\left[\inf_{\nu \in [0,1]} \left(\frac{M_\nu^2 d(x^*)^{1+\nu}}{\varepsilon^2} \right)^{\frac{1}{2p\nu-\nu+1}}\right].$$

The dependence of this bound in smoothness parameters is optimal (see [21]).

Let's compare our method and convergence rate with existing optimal methods. We assume that N the number of iterations, then $F(y_N) - F^* \leq B(N) + C(N)\delta_p + D(N)\delta_u + \frac{\varepsilon}{2}$. Here $B(N)$ is an accuracy of our method, $C(N)$ is a speed of collecting prox-mapping error and $D(N)$ is a speed of collecting oracle error. As a result we get next table.

(ν, p)	$B(N)$	$C(N)$	$D(N)$
$(0, 1)$	$O\left(\frac{M_0 d(x^*)^{1/2}}{N^{1/2}}\right)$	$O(M_0 d(x^*)^{1/2} N^{1/2})$	$O(1)$
$(1, 1)$	$O\left(\frac{M_1 d(x^*)}{N}\right)$	$O(M_1 d(x^*))$	$O(1)$
$(1, 2)$	$O\left(\frac{M_1 d(x^*)}{N^2}\right)$	$O\left(\frac{M_1 d(x^*)}{N}\right)$	$O(N)$

For non-smooth functions ($\nu = 0$), the convergence rate of UIGM for any $p \in [1, 2]$ agrees with rate of convergence of subgradient methods. This methods are robust for oracle error, but collect prox-mapping error. For smooth functions ($\nu = 1$) and $p = 1$ UIGM has the same convergence rate as a dual gradient method. This method is robust both for oracle error and prox-mapping error. And for $p = 2$ UIGM has the same rate as a fast gradient method. This method collects oracle error but kill prox-mapping error. This table shows three main regimes for UIGM and how it corresponds with classical methods.

5 Accelerating UIGM for strongly convex functions

In this section, we consider problem (1) with additional assumption of strong convexity of the objective F

$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_E^2, \quad \forall x, y \in Q,$$

where the constant $\mu > 0$ is assumed to be known. We also assume that the chosen prox-function has quadratic growth

$$d(x) \leq \frac{\Omega}{2} \|x\|_E^2, \tag{39}$$

where Ω is some dimensional-dependent constant, and that we are given a starting point x_0 and a number R_0 such that

$$\|x_0 - x^*\|_E^2 \leq R_0^2, \tag{40}$$

where x^* is an optimal point in (1).

Algorithm 2 Restart UIGM

Require: μ – strong convexity parameter, Ω – quadratic growth constant, ε – desired accuracy, x_0 – starting point.

- 1: Set $d_0(x) = d(x - x_0)$.
 - 2: **for** $m = 1, \dots$ **do**
 - 3: **while** $2\Omega > \mu A_k$ **do**
 - 4: Run UIGM with accuracy ε and prox-function $d_{m-1}(x)$.
 - 5: Set $x_m = y_k$.
 - 6: Set $d_m(x) = d(x - x_m)$.
-

Theorem 5.1. *Let F be strongly convex with constant μ and (39), (40) hold. Then, for any $m \geq 0$ restarts of UIGM with power policy (34), (35),*

$$F(x_m) - F(x^*) \leq \mu R_0^2 2^{-m-1} + 2 \left(\frac{\varepsilon}{2} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u + \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \mu^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\Omega^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} \delta_p \right), \quad (41)$$

$$\|x_m - x^*\|_E^2 \leq R_m^2 = R_0^2 2^{-m} + \frac{4}{\mu} \left(\frac{\varepsilon}{2} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u + \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \mu^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\Omega^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} \delta_p \right). \quad (42)$$

Proof. From (5), we have

$$\frac{\mu}{2} \|x_m - x^*\|_E^2 \leq \langle \nabla F(x^*), x_m - x^* \rangle + \frac{\mu}{2} \|x_m - x^*\|_E^2 \leq F(x_m) - F(x^*).$$

Then, by the first-order optimality condition,

$$\frac{\mu}{2} \|x_m - x^*\|_E^2 \leq F(x_m) - F(x^*).$$

From this fact and (41) we can easily prove (42).

To prove (41), we prove a stronger inequality by induction

$$F(x_m) - F(x^*) \leq \mu R_0^2 2^{-m-1} + 2(1 - 2^{-m}) \left(\frac{\varepsilon}{2} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u + \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \mu^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\Omega^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} \delta_p \right).$$

For $m = 1$, we have

$$\begin{aligned}
F(x_1) - F(x^*) &\stackrel{(28)}{\leq} \frac{d_0(x^*)}{A_k} + 2 \frac{\sum_{i=0}^k B_i}{A_k} \delta_u + \frac{(2k+1)\delta_p}{A_k} + \frac{\varepsilon}{2} \\
&\stackrel{(39)}{\leq} \frac{\Omega \|x_0 - x^*\|_E^2}{2A_k} + 2 \frac{\sum_{i=0}^k B_i}{A_k} \delta_u + \frac{(2k+1)\delta_p}{A_k} + \frac{\varepsilon}{2} \\
&\stackrel{(40)}{\leq} \frac{\Omega R_0^2}{2A_k} + 2 \frac{\sum_{i=0}^k B_i}{A_k} \delta_u + \frac{(2k+1)\delta_p}{A_k} + \frac{\varepsilon}{2} \\
&\stackrel{(38)}{\leq} \frac{\Omega R_0^2}{2A_k} + 2 \left(\frac{k+2p}{2p} \right)^{p-1} \delta_u + \frac{(2k+1)\delta_p}{A_k} + \frac{\varepsilon}{2}.
\end{aligned}$$

By the condition on the Step 3 of the algorithm, we have

$$A_{k-1} < \frac{2\Omega}{\mu} \leq A_k, \quad (43)$$

and

$$\begin{aligned}
\frac{2\Omega}{\mu} &\geq A_{k-1} \stackrel{(37)}{\geq} \frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{M_\nu^{\frac{2}{1+\nu}}} \left(\frac{k-1+2p}{2p} \right)^{\frac{2p\nu-\nu+1}{1+\nu}}, \\
\Rightarrow \frac{2^{\frac{1+\nu}{2p\nu-\nu+1}} \Omega^{\frac{1+\nu}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\mu^{\frac{1+\nu}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} &\geq \left(\frac{k-1+2p}{2p} \right), \\
\Rightarrow \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \Omega^{\frac{1+\nu}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\mu^{\frac{1+\nu}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} + 1 - 2p &\geq k, \\
\Rightarrow \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \Omega^{\frac{1+\nu}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\mu^{\frac{1+\nu}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} + 1 - 2p &\geq k.
\end{aligned}$$

Hence,

$$\left(\frac{k+2p}{2p} \right)^{p-1} \leq \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}}, \quad (44)$$

$$\Rightarrow 2k+1 \leq \frac{p 2^{\frac{4p\nu-\nu+3}{2p\nu-\nu+1}} \Omega^{\frac{1+\nu}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\mu^{\frac{1+\nu}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}}. \quad (45)$$

Finally, we have

$$\begin{aligned}
F(x_1) - F(x^*) &\stackrel{(44),(45)}{\leq} \frac{\Omega R_0^2}{2A_k} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u + \frac{p 2^{\frac{4p\nu-\nu+3}{2p\nu-\nu+1}} \Omega^{\frac{1+\nu}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\mu^{\frac{1+\nu}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}} A_k} \delta_p + \frac{\varepsilon}{2} \\
&\leq \mu R_0^2 2^{-2} + 2(1-2^{-1}) \left(\frac{\varepsilon}{2} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u + \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \mu^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\Omega^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} \delta_p \right).
\end{aligned}$$

So (41) is proved for $m = 1$. Now we assume that (41) holds for m and prove that it holds for $m + 1$.

From (28) we get

$$\begin{aligned}
F(x_{m+1}) - F(x^*) &\stackrel{(28)}{\leq} \frac{d_m(x^*)}{A_k} + 2 \frac{\sum_{i=0}^k B_i}{A_k} \delta_u + \frac{(2k+1)\delta_p}{A_k} + \frac{\varepsilon}{2} \\
&\stackrel{(39)}{\leq} \frac{\Omega \|x_m - x^*\|_E^2}{2A_k} + 2 \frac{\sum_{i=0}^k B_i}{A_k} \delta_u + \frac{(2k+1)\delta_p}{A_k} + \frac{\varepsilon}{2} \\
&\stackrel{(42)}{\leq} \frac{\Omega R_m^2}{2A_k} + 2 \frac{\sum_{i=0}^k B_i}{A_k} \delta_u + \frac{(2k+1)\delta_p}{A_k} + \frac{\varepsilon}{2} \\
&\stackrel{(38)}{\leq} \frac{\Omega R_m^2}{2A_k} + 2 \left(\frac{k+2p}{2p} \right)^{p-1} \delta_u + \frac{(2k+1)\delta_p}{A_k} + \frac{\varepsilon}{2} \\
&\stackrel{(44),(45)}{\leq} \frac{\Omega R_m^2}{2A_k} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u \\
&\quad + \frac{p 2^{\frac{4p\nu-\nu+3}{2p\nu-\nu+1}} \Omega^{\frac{1+\nu}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\mu^{\frac{1+\nu}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}} A_k} \delta_p + \frac{\varepsilon}{2} \\
&\stackrel{(43)}{\leq} \frac{\mu R_m^2}{4} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u \\
&\quad + \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \mu^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\Omega^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} \delta_p + \frac{\varepsilon}{2}
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(42)}{\leq} \mu R_0^2 2^{-m-2} + \frac{4\mu}{4\mu} (1 - 2^{-m}) \left(\frac{\varepsilon}{2} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u \right. \\
&+ \left. \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \mu^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\Omega^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} \delta_p \right) + \frac{\varepsilon}{2} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u \\
&+ \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \mu^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\Omega^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} \delta_p \\
&\leq \mu R_0^2 2^{-m-2} + (2 - 2^{-m}) \left(\frac{\varepsilon}{2} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u \right. \\
&+ \left. \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \mu^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\Omega^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} \delta_p \right) \\
&\leq \mu R_0^2 2^{-m-2} + 2(1 - 2^{-(m+1)}) \left(\frac{\varepsilon}{2} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u \right. \\
&+ \left. \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \mu^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\Omega^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} \delta_p \right).
\end{aligned}$$

So we have obtained that (41) holds for $m+1$ and by induction it holds for all $m \geq 1$ □

Corollary 5.2. *For getting $(\varepsilon + C_u)$ -solution of problem (1), where*

$$C_u = 2 \left(\frac{\varepsilon}{2} + 2 \left[\left(\frac{2^{\nu+1} \Omega^{\nu+1} M_\nu^2}{\mu^{\nu+1} \varepsilon^{1-\nu}} \right) \right]^{\frac{p-1}{2p\nu-\nu+1}} \delta_u + \frac{p 2^{\frac{2p\nu+2}{2p\nu-\nu+1}} \mu^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} M_\nu^{\frac{2}{2p\nu-\nu+1}}}{\Omega^{\frac{2\nu(p-1)}{2p\nu-\nu+1}} \varepsilon^{\frac{1-\nu}{2p\nu-\nu+1}}} \delta_p \right),$$

we need

$$\tilde{l} = \left\lceil \log \left(\frac{\mu R_0^2}{2\varepsilon} \right) \right\rceil \quad (46)$$

restarts and

$$\tilde{k} \leq \inf_{0 \leq \nu \leq 1} \left(\frac{\Omega^{1+\nu} 2^{4p\nu-\nu+3} M_\nu^2}{\mu^{1+\nu} \varepsilon^{1-\nu}} \right)^{\frac{1}{2p\nu-\nu+1}} + 1$$

iterations of UIGM per iteration. The total number of UIGM iterations is not more than

$$N = \left(\inf_{0 \leq \nu \leq 1} \left(\frac{\Omega^{1+\nu} 2^{4p\nu-\nu+3} M_\nu^2}{\mu^{1+\nu} \varepsilon^{1-\nu}} \right)^{\frac{1}{2p\nu-\nu+1}} + 1 \right) \cdot \left\lceil \log \left(\frac{\mu R_0^2}{2\varepsilon} \right) \right\rceil.$$

Proof.

$$F(x_{\tilde{l}}) - F(x^*) \stackrel{(41)}{\leq} \mu R_0^2 2^{-\tilde{l}-1} + C_u \stackrel{(46)}{\leq} \varepsilon + C_u.$$

We now estimate the total number of UIGM iterations, which is sufficient to obtain $(\varepsilon + C_u)$ -solution. First, we estimate the number \tilde{k} of UIGM iterations at each restart. By the stopping condition for the restart, we have

$$\begin{aligned}
A_{\tilde{k}} &\geq \frac{2\Omega}{\mu} \stackrel{(37)}{>} A_{\tilde{k}-1} \geq \frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{M_\nu^{\frac{2}{1+\nu}}} \left(\frac{\tilde{k}-1}{4} \right)^{\frac{2p\nu-\nu+1}{1+\nu}}, \\
&\Rightarrow \frac{2\Omega 2^{\frac{4p\nu-2\nu+2}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}{\mu \varepsilon^{\frac{1-\nu}{1+\nu}}} \geq \left(\frac{\tilde{k}-1}{4} \right)^{\frac{2p\nu-\nu+1}{1+\nu}}, \\
&\Rightarrow \left(\frac{\Omega^{1+\nu} 2^{4p\nu-\nu+3} M_\nu^2}{\mu^{1+\nu} \varepsilon^{1-\nu}} \right)^{\frac{1}{2p\nu-\nu+1}} \geq \tilde{k} - 1.
\end{aligned}$$

Since the algorithm does not use any particular choice of ν , we have

$$\tilde{k} \leq \inf_{0 \leq \nu \leq 1} \left(\frac{\Omega^{1+\nu} 2^{4p\nu-\nu+3} M_\nu^2}{\mu^{1+\nu} \varepsilon^{1-\nu}} \right)^{\frac{1}{2p\nu-\nu+1}} + 1.$$

Then the total number of UIGM is not more than $N = \tilde{k} \cdot \tilde{l}$, and we have

$$N = \left(\inf_{0 \leq \nu \leq 1} \left(\frac{\Omega^{1+\nu} 2^{4p\nu-\nu+3} M_\nu^2}{\mu^{1+\nu} \varepsilon^{1-\nu}} \right)^{\frac{1}{2p\nu-\nu+1}} + 1 \right) \cdot \left\lceil \log \left(\frac{\mu R_0^2}{2\varepsilon} \right) \right\rceil.$$

□

Now we compare our result with existing methods in the same manner as for convex functions. $u = C\delta_p + D\delta_u + \frac{\varepsilon}{2}$. Here C is a collecting prox-mapping error for given ε, μ and D is a collecting oracle error for desired ε, μ . As a result we get next table.

(ν, p)	N	C	D
$(0, 1)$	$O\left(\frac{\Omega M_0^2}{\mu \varepsilon} \cdot \log\left(\frac{\mu R_0^2}{2\varepsilon}\right)\right)$	$O\left(\frac{M_0^2}{\varepsilon}\right)$	$O(1)$
$(1, 1)$	$O\left(\frac{\Omega M_1}{\mu} \cdot \log\left(\frac{\mu R_0^2}{2\varepsilon}\right)\right)$	$O(M_1)$	$O(1)$
$(1, 2)$	$O\left(\left(\frac{\Omega M_1}{\mu}\right)^{\frac{1}{2}} \cdot \log\left(\frac{\mu R_0^2}{2\varepsilon}\right)\right)$	$O\left(\frac{\mu M_1}{\Omega}\right)^{\frac{1}{2}}$	$O\left(\frac{\Omega M_1}{\mu}\right)^{\frac{1}{2}}$

For non-smooth functions ($\nu = 0$), the convergence rate of Restart UIGM for any $p \in [1, 2]$ agrees with rate of convergence of subgradient methods. This methods are robust for oracle error, but collect prox-mapping error. For smooth functions ($\nu = 1$) and $p = 1$ Restart UIGM has the same convergence rate as a dual gradient method. This method is robust both for oracle error and prox-mapping error. And for $p = 2$ Restart UIGM has the same rate as a fast gradient method. This method collects oracle error but kill prox-mapping error. This table shows three main regimes for Restart UIGM and how it corresponds with classical methods.

6 Switching policy

In this section we describe another variant of coefficient policy. The key observation is that Fast gradient method(FGM) accumulates the error, but converges faster and Dual gradient method(DGM) doesn't accumulate the error, but works slower. That's why at the begging we make some steps of FGM until the error reaches some limit and then make only DGM steps. This policy was introduced in [5]. Now we should understand, what is the limit. If we want to get the total error equal to ε , then the the error from inexactness should be $\varepsilon/2$. Now we describe this idea in more details.

Let the switching policy is

$$\alpha_k = \begin{cases} \frac{k+4}{4} \cdot \frac{1}{L_k} & k = 0, \dots, s \quad - \text{FGM steps} \\ c_k \cdot \frac{1}{L_k} & k = s + 1, \dots, N \quad - \text{DGM steps} \end{cases}, \quad (47)$$

where s is the moment of switching and c_k is some constant, we will describe later how to choose them.

Firstly, we should prove, that the switching policy can be used in UIGM. Let check the correctness of inequalities (12) for the switching policy. For FGM steps it is easily follow from (4.1), because it is the power policy for $p = 2$. For DGM steps we need to prove that

$$0 < c_k \cdot \frac{1}{L_k} \leq c_k^2 \cdot \frac{1}{L_k} \leq A_{k-1} + c_k \cdot \frac{1}{L_k}$$

First two inequalities are satisfied if $c_k \geq 1$. So we get the first condition for c_k . The second condition comes from the last inequality because we need to get c_k such that $c_k^2 - c_k - A_{k-1}L_k \geq 0$. Hence

$$1 \leq c_k \leq \frac{1 + \sqrt{1 + A_{k-1}L_k}}{2} \quad (48)$$

So if these two conditions for c_k are satisfied we prove that the switching policy can be used in UIGM.

Secondly, we should prove convergence of the switching policy. For that we need to satisfy two inequalities from (28)

$$\frac{2\delta_u}{A_K} \sum_{j=0}^k B_j \leq \frac{\varepsilon}{6} \quad (49)$$

$$\frac{(2k+1)\delta_p}{A_k} \leq \frac{\varepsilon}{6}. \quad (50)$$

Note that from this two inequalities we get three main regimes:

- Only FGM steps. In this case, $\delta_u \ll \varepsilon$ and $\delta_p \ll \varepsilon$, and both inequalities (49) and (50) never fail.
- Only DGM steps. In this case, δ_u is rather big and (49) fails on the first step, so we try our best and do only slow DGM steps.

- Switching on the moment s . In this case, we do some FGM steps until the moment s , when (49) fails at the first time and next do only DGM steps.

First two regimes are easy for understanding but for the last one we write more details. Note that for FGM steps (50) always true because the left side decreases on each step. Note that from some moment $\sum_{j=0}^k B_j / \sum_{j=0}^k \alpha_j$ starts to increase and at the moment s it reaches the limit $\frac{\varepsilon}{12\delta_u}$. Now we need to check, that for DGM steps (49) will not fail.

$$\begin{aligned} \sum_{j=0}^k B_j &\leq \frac{\varepsilon}{12\delta_u} \sum_{j=0}^k \alpha_j \\ \sum_{j=0}^{k-1} B_j + B_k &\leq \frac{\varepsilon}{12\delta_u} \left(\sum_{j=0}^{k-1} \alpha_j + \alpha_k \right) \end{aligned}$$

We assume that on previous step (49) was correct, that's why we need

$$\begin{aligned} B_k &\leq \frac{\varepsilon}{12\delta_u} \alpha_k \\ \frac{c_k^2}{L_k} &\stackrel{(47)}{\leq} \frac{\varepsilon}{12\delta_u} \frac{c_k}{L_k} \end{aligned}$$

So we get the third condition on c_k

$$c_k \leq \frac{\varepsilon}{12\delta_u} \quad (51)$$

Hence when when we merge all conditions (48) and (51), we get

$$c_k = \min \left(\frac{\varepsilon}{12\delta_u}, \frac{1 + \sqrt{1 + A_{k-1} L_k}}{2} \right) \quad (52)$$

As a result, we've proved that the switching policy can be used with UIGM. We've proved that UIGM converges, when we do FGM steps until at the moment s (49) fails, then switch DGM with c_k defined by (52). Note, that now our method needs to know only $\varepsilon, \delta_u, \delta_p$ and doesn't need p as in the power policy, hence it converges as well or better than any p for power policy.

Theorem 6.1. *Assume that f is a convex function with inexact first-order oracle. Then, for the sequence (47), the moment s is first time when (49) fails and c_k defined by (52), for all $k \geq 0$,*

$$F(y_k) - F^* \leq \inf_{p \in [1,2]} O \left[\inf_{\nu \in [0,1]} \left(\frac{M_\nu^{\frac{2}{1+\nu}} d(x^*)}{\varepsilon^{\frac{1-\nu}{1+\nu}} k^{\frac{2p\nu-\nu+1}{1+\nu}}} + \frac{M_\nu^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{1-\nu}{1+\nu}} k^{\frac{2\nu(p-1)}{1+\nu}}} \delta_p \right) + k^{p-1} \delta_u \right] + \frac{\varepsilon}{2}.$$

The same argumentation is correct also for strongly convex functions. So now we get fully adaptive and universal coefficient policy and method.

7 Numerical illustration

For numerical illustration we choose a Poisson likelihood problems and as an application Positron Emission Tomography(PET). It plays an important role in medicine for detecting cancer and metabolic changes in human organ. PET can be treated as a Poisson likelihood model [3], [17]. The estimation of radioactivity density within an organ corresponds to the following convex non-smooth optimization problem:

$$\min_{x \in \Delta_n} \sum_{i=1}^m [[Ax]_i - w_i \log([Ax]_i)]$$

where Δ_n is a standard simplex. A is a data and refers to the likelihood matrix known from geometry of detector, and w is a data and refers to the vector of detected events, such that $w_i = [Ax]_i + b_i$, where b_i is Poisson noise for any $1 \leq i \leq m$. So we get a regression and our goal is to find x from data. For simplicity, we will not consider any penalty term for this application. Note that actually this problem has unbounded M_ν , because $\nabla \log y = 1/y$ is unbounded in $y = 0$. So here we assume that all points of our method are separated from zero and then M_ν is bounded by some constant.

We assume, that tomographic scanner can have some small random and systematic errors, so we get inexact data and hence inexact function and gradient. So we get inexact oracle. If method converges with inexact data it means that we have robust system and even with errors we will get rather precise tomography.

In this case, the entropy function $d(x) = \sum_{i=1}^n x_i \log(x_i)$ is a good choice for the simplex domain. Moreover, the prox-mapping can be computed by direct formula [5], which means that we have $\delta_p = 0$. If we choose another $d(x)$ it may be worse, because for the finding of the prox-mapping we need to solve additional optimization subproblem. For example we can approximately solve it by FGM with $\delta_p > 0$, because this subproblem is strongly convex and that's why FGM converges fast.

Code is written in Python 3. We conduct experiments using Ubuntu 14, machine: Intel Core i7-4510U CPU 2.00GHz 2.60GHz, 8Gb RAM. Matrix $A \in \mathbb{R}^{100 \times 200}$ and $w \in \mathbb{R}^{100}$ are generated uniformly randomly. For simplicity, we calculate inexact oracle as exact oracle plus the noise δ_u . Desired accuracy is $\varepsilon = 0.0001$

For small inexactness $\delta = 0.001\varepsilon$ UIGM give us next graphic.

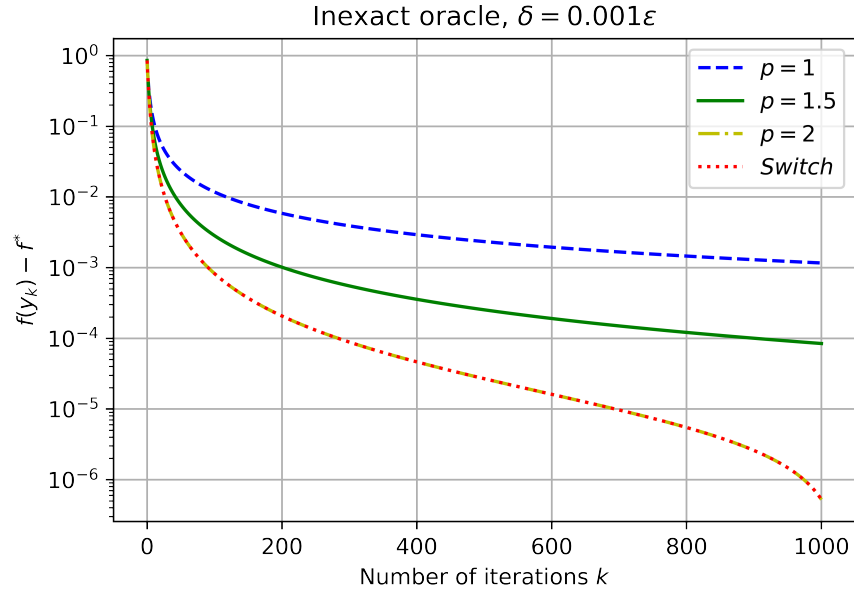


Figure 1: Comparison of different power policies and switch policy for small inexactness

From this graphic we can see that, power policy with $p = 2$ and switching policy are the fastest. For small inexactness all variants don't collect any noticeable error.

In next graphic, we can see, that for medium error $\delta = \epsilon$ switching policy starts to work as power policy with $p = 1$, because all our estimates of error collection come from theory but the real error in specific point can be less than theoretical estimate. Unfortunately we can't measure the real error.

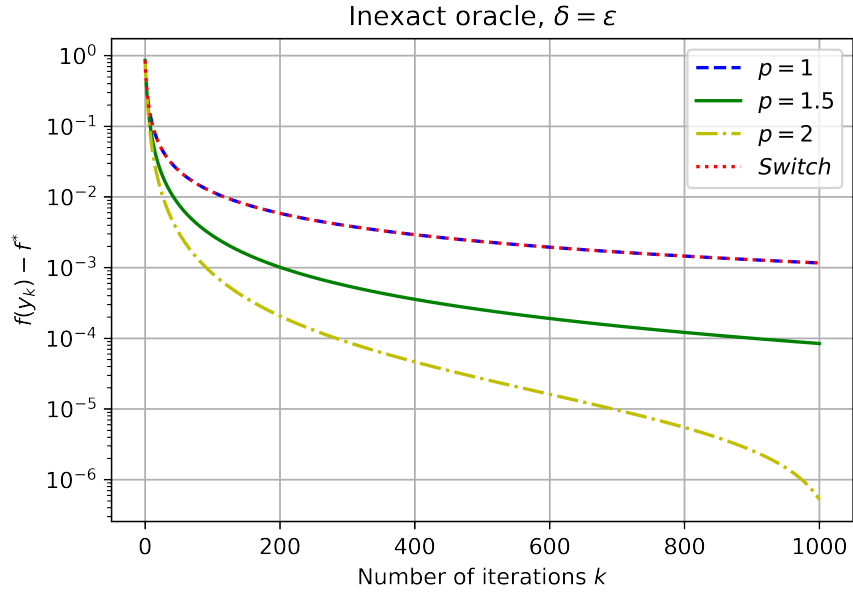


Figure 2: Comparison of different power policies and switch policy for medium inexactness

For big error $\delta = 1000\epsilon$ the power policy with $p = 2$ collects error and works worse than the power policy $p = 1.5$. So the method with intermediate rate is the best one, because it is rather fast and also robust.

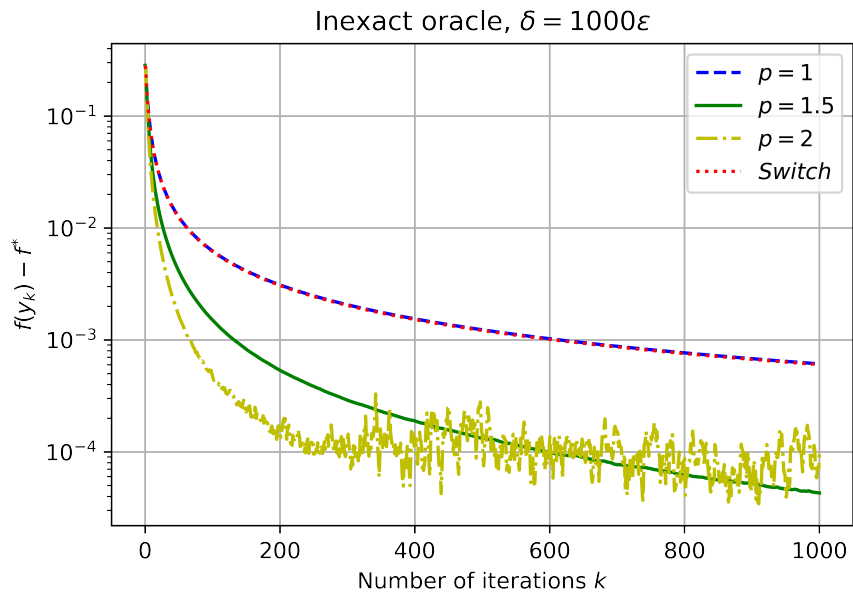


Figure 3: Comparison of different power policies and switch policy for big inexactness

As a result, we get that UIGM for some intermediate p can be better, than classical methods. Also we get that our method get a speed up for non-smooth problem in comparison with optimal DGM ($p = 1$). Unfortunately in practice switching policy may be worse, than power policy because of uncertainty of real error.

8 Conclusion

In this paper, we present new Universal Intermediate Gradient Method for convex optimization problem with inexact Hölder-continuous subgradient. Our method enjoys both the universality with respect to smoothness of the problem and interpolates between Universal Gradient Method and Universal Fast Gradient Method, thus, allowing to balance the rate of convergence of the method and rate of the error accumulation. Under additional assumption of strong convexity of the objective, we show how the restart technique can be used to obtain an algorithm with faster rate of convergence.

We note that Theorem 3.1 is primal-dual friendly. This means that, if UIGM is used to solve a problem, which is dual to a problem with linear constraints, it generates also a sequence of primal iterates and the rate for the primal-dual gap and linear constraints infeasibility is the same. This can be proved in the same way as in Theorem 2 of [8]. Also, based on the ideas from [9,10,25], UIGM for the strongly convex case can be modified to work without exact knowledge of strong convexity parameter μ . Finally, similarly to [7, 11, 15], UIGM can be modified to solve convex problems with stochastic inexact oracle.

Acknowledgements. This research was funded by Russian Science Foundation (project 17-11-01027).

References

- [1] Dilyara Baimurzina, Alexander Gasnikov, Evgenia Gasnikova, Pavel Dvurechensky, Egor Ershov, Anastasia Lagunovskaya, and Meruza Kubentaeva. Universal similar triangular method for searching equilibriums in traffic flow distribution models. *Comp.Math. & Math. Phys.*, 58, 2018.
- [2] Aaron Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization (Lecture Notes)*. Personal web-page of A. Nemirovski, 2015.
- [3] Aharon Ben-Tal, Tamar Margalit, and Arkadi Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12(1):79–108, 2001.
- [4] Alexandre d’Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.
- [5] Olivier Devolder, François Glineur, and Yurii Nesterov. Intermediate gradient methods for smooth convex problems with inexact oracle. 2013. CORE Discussion Paper 2013/17.

- [6] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.
- [7] Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1):121–145, 2016.
- [8] Pavel Dvurechensky, Alexander Gasnikov, Sergey Omelchenko, and Alexander Tiurin. Adaptive similar triangles method: a stable alternative to sinkhorn’s algorithm for regularized optimal transport. *arXiv preprint arXiv:1706.07622*, 2017.
- [9] Olivier Fercoq and Zheng Qu. Restarting accelerated gradient methods with a rough strong convexity estimate. *arXiv preprint arXiv:1609.07358*, 2016.
- [10] Olivier Fercoq and Zheng Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *arXiv preprint arXiv:1709.02300*, 2017.
- [11] Alexander Gasnikov, Pavel Dvurechensky, and Dmitry Kamzolov. Gradient and gradient-free methods for stochastic convex optimization with inexact oracle. *arXiv preprint arXiv:1502.06259*, 2015.
- [12] Alexander Gasnikov, Pavel Dvurechensky, Dmitry Kamzolov, Yurii Nesterov, Vladimir Spokoiny, Petr Stetsyuk, Alexandra Suvorikova, and Alexey Chernov. Universal method with inexact oracle and its applications for searching equilibriums in multistage transport problems. *arXiv preprint arXiv:1506.00292*, 2015.
- [13] Alexander Gasnikov, Sergey Kabanikhin, Ahmed Mohamed, and Maxim Shishlenin. Convex optimization in hilbert space with applications to inverse problems. *arXiv preprint arXiv:1703.00267*, 2017.
- [14] Alexander Gasnikov, Dmitry Kamzolov, and Mikhail Mendel. Universal composite prox-method for strictly convex optimization problems. *TRUDY MIPT*, 8(3):25–42, 2016.
- [15] Alexander Gasnikov and Yurii Nesterov. Universal fast gradient method for stochastic composite optimization problems. *Comp.Math. Math. Phys.*, 58(1):51–68, 2018.
- [16] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Generalized uniformly optimal methods for nonlinear programming. *ArXiv preprint arXiv:1508.07384*, 2015.
- [17] Niao He, Zaid Harchaoui, Yichen Wang, and Le Song. Fast and simple optimization for poisson likelihood models. *arXiv preprint arXiv:1608.01264*, 2016.
- [18] Anatoli Juditsky and Yuri Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

- [19] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.
- [20] A.S. Nemirovskii and Yu.E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21 – 30, 1985.
- [21] A.S. Nemirovsky and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. J. Wiley & Sons, New York, 1983.
- [22] Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.
- [23] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady AN USSR*, volume 269, pages 543–547, 1983.
- [24] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.
- [25] Vincent Roulet and Alexandre d’Aspremont. Sharpness, restart and acceleration. *arXiv preprint arXiv:1702.03828*, 2017.
- [26] Alp Yurtsever, Quoc Tran-Dinh, and Volkan Cevher. A universal primal-dual convex optimization framework. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS’15, pages 3150–3158, Cambridge, MA, USA, 2015. MIT Press.