Universal intermediate gradient method for convex problems with inexact oracle

Dmitry Kamzolov, Pavel Dvurechensky and Alexander V. Gasnikov

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 Nesterov (Math. Program. 152 (2015), pp. 381–404) 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 Devolder et al. (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 the 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.

ARTICLE HISTORY
Received 18 December 2017
Accepted 10 December 2019

KEYWORDS
Convex optimization; first-order methods; inexact oracle; intermediate gradient methods; complexity bounds

AMS SUBJECT CLASSIFICATIONS
90C25; 90C47; 90C60

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 or Fast Gradient Method, 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

CONTACT
Dmitry Kamzolov, kamzolov.dmitry@phystech.edu, dkamzolov@yandex.ru
© 2020 Informa UK Limited, trading as Taylor & Francis Group
However, these methods require to know the Hölder constant. In 2013, Nesterov proposed a Universal Fast Gradient Method [34] 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 [31]. In 2012, Lan proposed a Fast gradient method with one prox-mapping for stochastic optimization problems [29]. In 2016, Gasnikov and Nesterov proposed a Universal Triangle Method [20], 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 [25], where the authors introduce a method which is uniformly optimal for convex and non-convex problems with Hölder-continuous subgradient, and the work [39], in which a universal primal-dual method is proposed to solve linearly constrained convex problems.

Another line of research [7–10,17] 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 [9], gradient descent has slower rate of convergence, but does not accumulate the error of the oracle. On the opposite, Fast Gradient Method has a faster convergence rate, but accumulates the error linearly with the iteration counter. Later, in [8] 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 [2,21], inverse problems [22] and others.

We also consider the same problem under the additional assumption of strong convexity of the objective function and show how the restart technique [16,19,28,30–32,36] 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 Section 2, we state the problem. After that, in Section 3, we present Universal Intermediate Gradient Method and prove a convergence rate theorem with general choice of controlling sequence of coefficients. In Section 4, we analyze particular choice of controlling sequence of coefficients and prove a convergence rate theorem under this choice of coefficients. In Section 5, we present UIGM for strongly convex functions and prove convergence rate theorem under this additional assumption. In Section 6, we introduce another choice of coefficients that don’t need any additional information. In Section 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 $f : \text{dom} f \rightarrow R$ with $\text{dom} f \subseteq E$ we denote by $\nabla f(x) \in E^*$ one of its subgradients.

We consider the following convex composite optimization problem [33]:

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

where $Q$ is a simple closed convex function, $h(x)$ 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 2.1 ([9]):** We say that a convex function $f(x)$ is equipped with a first-order ($\delta, L$)–oracle on a convex set $Q$ if for any point $x \in Q$, ($\delta, L$)-oracle returns a pair $(f_\delta(x), g_\delta(x)) \in R \times E^*$ such that

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

In this definition, $\delta$ represents the error of the oracle [9]. The oracle is exact with $\delta = 0$. Also we can take $\delta = \delta_c + \delta_u$, where $\delta_c$ represents the error, which we can control and make as small as we would like to. On the opposite, $\delta_u$ represents the error, which we can not control [10]. Note that, by Definition 2.1,

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

To motivate Definition 2.1, we consider the following example. Let $f$ be a convex function with H"older-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 [9], it was proved that, for such function for any $\delta_c > 0$, if

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

then

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

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 $\delta_1, \delta_2 > 0$ such that, for any point $x \in Q$, we can
calculate approximations \( \tilde{f}(x) \) and \( \tilde{g}(x) \) with \( |\tilde{f}(x) - f(x)| \leq \delta_1 \) and \( \|\tilde{g}(x) - \nabla f(x)\|_{E^*} \leq \delta_2 \).

Let us show that, in this example, \( f \) can be equipped with inexact first-order oracle based on the pair \( (\tilde{f}(x), \tilde{g}(x)) \), where \( \tilde{f}_g(x) = \tilde{f}(x) - \delta_1 - \delta_2 D \) and \( g_\delta(x) = \tilde{g}(x) \).

Now we prove the first inequality from (2)

\[
 f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \\
 \geq \tilde{f}(x) - \delta_1 + \langle \tilde{g}(x), y - x \rangle - \delta_2 D = f_\delta(x) + \langle g_\delta(x), y - x \rangle
\]

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

\[
 f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L(\delta_c)}{2} \|x - y\|^2_E + \delta_c
\]

\[
 \leq \tilde{f}(x) + \delta_1 + \langle \tilde{g}(x), y - x \rangle + \langle \nabla f(x) - \tilde{g}(x), y - x \rangle + \frac{L(\delta_c)}{2} \|x - y\|^2_E + \delta_c
\]

\[
 \leq f_\delta(x) + \langle g_\delta(x), y - x \rangle + \frac{L(\delta_c)}{2} \|x - y\|^2_E + 2\delta_1 + 2\delta_2 D + \delta_c.
\]

Thus, \((f_\delta(x), g_\delta(x))\) is an inexact first-order oracle with \( \delta_u = 2\delta_1 + 2\delta_2 D, \delta_c, \) and \( L(\delta_c) \) given by (4).

To construct our algorithm for problem (1), we introduce, as it is usually done, proximal setup [4], 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 \( x, y \in \text{rint} \ Q \),

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

2. Without loss of generality, we assume that

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

Then if \( \bar{x} = \arg\min_{x \in Q} d(x) \), we get

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

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

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

We use prox-function in so-called composite prox-mapping, which consists of 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 2.2:** 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

$$\tilde{x} = \arg\min_{x \in Q} \delta_p \left\{ \langle g, x \rangle + d(x) + h(x) \right\}.$$

one of the possible inexact composite prox-mappings.

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 [4]):** Let $F : Q \to \mathbb{R} \cup \{+\infty\}$ be a convex function such that $\Psi(x) = F(x) + d(x)$ is closed and convex on $Q$. Denote $\tilde{x} = \arg\min_{x \in Q} \delta_p \Psi(x)$. 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 $\alpha_k, B_k, k \geq 0$. 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 [8] with Algorithm 1 from [20] to get IGM with only one prox-mapping instead two as in [8]. After that, we improve this method by techniques from [34] to get UIGM with exact prox-mapping. Last generalization use Lemma 5.5.1 from [4]. As a result, we get algorithm that works in wide class of problems, adaptive and does not need to know exact Hölder and Lipschitz constants, uses only one prox-mapping and correctly works with errors of oracle and prox-mapping.
Algorithm 1 Universal Intermediate Gradient Method (UIGM)

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

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

\[
z_0 = x_0 = \text{argmin}_{x \in Q} \delta_p d(x),
\]

(13)

2: Set \( i_0 = 0 \)

3: Compute

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

(14)

4: If

\[
f_{\delta_0}(y_0) \leq f_{\delta_0}(x_0) + \langle g_{\delta_0}(x_0), y_0 - x_0 \rangle + \frac{2^{i_0} L_s}{2} \| y_0 - x_0 \|^2 + \delta_0,
\]

goto 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 \( k = 1, \ldots \) 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),

\[
B_k = \alpha_k^2 L_k,
\]

(16)

\[
\delta_k = \frac{\alpha_k \varepsilon}{B_k} + \delta_u
\]

(17)

\[
x_k = \frac{\alpha_k}{B_k} z_{k-1} + \frac{B_k - \alpha_k}{B_k} y_{k-1}.
\]

(18)

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

(19)

\[
w_k = \frac{\alpha_k}{B_k} z_k + \frac{B_k - \alpha_k}{B_k} y_{k-1}.
\]

(20)

9: If

\[
f_{\delta_k}(w_k) \leq f_{\delta_k}(x_k) + \delta_k + \langle g_{\delta_k}(x_k), w_k - x_k \rangle + \frac{L_k}{2} \| w_k - x_k \|^2.
\]

goto Step 10. Otherwise, set \( i_k = i_k + 1 \) and go back to Step 8.

10: Set

\[
A_k = A_{k-1} + \alpha_k,
\]

(22)

\[
y_k = \frac{B_k}{A_k} w_k + \frac{A_k - B_k}{A_k} y_{k-1}.
\]

(23)

Ensure: \( y_k \).
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 [8] and Theorem 3 in [34].

**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

$$A_k F(y_k) - E_k \leq \Psi^*_k,$$

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

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

**Proof:** Let us prove first, that the ‘line-search’ process of steps 6–9 is finite. By (4), (5), if $2^{ik} L_{k-1} \geq L(\frac{\alpha_k \varepsilon}{4})$, from (18) and (3), we get

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

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

$$2^{ik} L_{k-1} \geq \left[ \frac{\alpha_k \varepsilon}{B_k} \right]^{\frac{1}{1+\nu}} \frac{2^\nu}{M_{\delta}^\nu}$$

for $i_k$ large enough. Indeed,

$$2^{ik} L_{k-1} \left[ \frac{\alpha_k}{B_k} \right]^{\frac{1}{1+\nu}} B_k \left[ \frac{\alpha_k}{B_k} \right]^{\frac{1}{1+\nu}} = \left[ \frac{B_k}{\alpha_k} \right]^{\frac{2\nu}{1+\nu}} \frac{1}{\alpha_k} \geq \frac{1}{\alpha_k}.$$  

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

$$\alpha_k^2 = \frac{B_k}{2^{ik} L_{k-1}} \leq \frac{A_k}{2^{ik} L_{k-1}} \Rightarrow \alpha_k^2 - \frac{\alpha_k}{2^{ik} L_{k-1}} - \frac{A_{k-1}}{2^{ik} L_{k-1}} \leq 0.$$  

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

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

The solutions are

$$\alpha_k^- = \frac{1}{2^{ik+1} L_{k-1}} \left( \frac{1}{4^{ik+1} L_{k-1}^2} + \frac{A_{k-1}}{2^{ik} L_{k-1}} \right)^{1/2},$$

$$\alpha_k^+ = \frac{1}{2^{ik+1} L_{k-1}} \left( \frac{1}{4^{ik+1} L_{k-1}^2} + \frac{A_{k-1}}{2^{ik} L_{k-1}} \right)^{1/2}.$$
Now from (27) we have that \( \alpha_k^- \leq \alpha_k \leq \alpha_k^+ \). From \( \alpha_k^- \to 0, \alpha_k^+ \to 0 \) as \( i_k \to \infty \) we get \( \alpha_k \to 0 \).

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

\[
\Psi_0^* \overset{(25)}{=} \min_{x \in Q} \left\{ d(x) + \alpha_0 f_{\delta_0}(x_0) + \alpha_0 \langle g_{\delta_0}, x - x_0 \rangle + \alpha_0 h(x) \right\} \\
\overset{(11),(14)}{\geq} d(y_0) + \alpha_0 f_{\delta_0}(x_0) + \alpha_0 \langle g_{\delta_0}(x_0), y_0 - x_0 \rangle + \alpha_0 h(y_0) - \delta_p \\
\overset{(6),(13)}{\geq} \alpha_0 \left( \frac{1}{2\alpha_0} \| y_0 - x_0 \|_E^2 + f_{\delta_0}(x_0) + \langle g_{\delta_0}(x_0), y_0 - x_0 \rangle + h(y_0) \right) - \delta_p \\
= \alpha_0 \left( \frac{1}{2} \| y_0 - x_0 \|_E^2 + f_{\delta_0}(x_0) + \langle g_{\delta_0}(x_0), y_0 - x_0 \rangle + h(y_0) \right) - \delta_p \\
\overset{(15)}{\geq} \alpha_0 \left( f(y_0) - \frac{\epsilon}{4} - \delta_u + h(y_0) \right) - \delta_p \\
\geq \alpha_0 \left( f(y_0) - \frac{\epsilon}{2} - 2\delta_u + h(y_0) \right) - \delta_p = A_0 F(y_0) - E_0.
\]

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

\[
\Psi_k^* \overset{(25)}{=} \min_{x \in Q} \Psi_k(x) \overset{(11),(19)}{\geq} \Psi_k(z_k) - \delta_p \\
\overset{(25)}{=} \Psi_{k-1}(z_k) + \alpha_k \left[ f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle + h(z_k) \right] - \delta_p \\
\overset{(10)}{\geq} \Psi_{k-1}(z_{k-1}) + V(z_{k-1}, z_k) - 2\delta_p + \alpha_k \left[ f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle + h(z_k) \right] \\
\overset{(7)}{\geq} \Psi_{k-1}^* + \frac{1}{2} \| z_k - z_{k-1} \|_E - 2\delta_p + \alpha_k \left[ f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle + h(z_k) \right] \\
\overset{(24)}{\geq} A_k f(y_{k-1}) - E_{k-1} + \frac{1}{2} \| z_k - z_{k-1} \|_E - 2\delta_p \\
+ \alpha_k \left[ f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle + h(z_k) \right] \\
= (A_k - B_k) f(y_{k-1}) - E_{k-1} + \frac{1}{2} \| z_k - z_{k-1} \|_E - 2\delta_p + (B_k - \alpha_k) f(y_{k-1}) \\
+ (B_k - \alpha_k) h(y_{k-1}) + \alpha_k h(z_k) + \alpha_k \left[ f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle \right] \\
\overset{(20)}{\geq} (A_k - B_k) f(y_{k-1}) - E_{k-1} + B_k h(w_k) + \frac{1}{2} \| z_k - z_{k-1} \|_E - 2\delta_p \\
+ (B_k - \alpha_k) f(y_{k-1}) + \alpha_k \left[ f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle \right] \\
\geq (A_k - B_k) f(y_{k-1}) - E_{k-1} + B_k h(w_k) + \frac{1}{2} \| z_k - z_{k-1} \|_E - 2\delta_p \\
+ (B_k - \alpha_k) \left[ f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), y_{k-1} - x_k \rangle \right] \\
+ \alpha_k \left[ f_{\delta_k}(x_k) + \langle g_{\delta_k}(x_k), z_k - x_k \rangle + h(z_k) \right] \\
= (A_k - B_k) f(y_{k-1}) - E_{k-1} + B_k h(w_k) + \frac{1}{2} \| z_k - z_{k-1} \|_E + B_k f_{\delta_k}(x_k) \\
+ \langle g_{\delta_k}(x_k), (B_k - \alpha_k) (y_{k-1} - x_k) + \alpha_k (z_k - x_k) \rangle - 2\delta_p.
\]
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,
\[
\Psi^*_k \geq (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p \\
+ B_k f_{\delta_k}(x_k) + \alpha_k (g_{\delta_k}(x_k), (z_k - z_{k-1})) + \frac{1}{2} ||z_k - z_{k-1}||_E \\
= (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p \\
+ B_k \left[ f_{\delta_k}(x_k) + \alpha_k (g_{\delta_k}(x_k), z_k - z_{k-1}) + \frac{1}{2B_k} ||z_k - z_{k-1}||_E^2 \right].
\]

As \(B_k = 2^{\ell_k} L_{k-1} \alpha_k^2\), we have \(\frac{1}{B_k} = 2^{\ell_k} L_{k-1} \alpha_k^2\) and, therefore,
\[
\Psi^*_k \geq (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p \\
+ B_k \left[ f_{\delta_k}(x_k) + \frac{\alpha_k}{B_k} (g_{\delta_k}(x_k), z_k - z_{k-1}) + \frac{2^{\ell_k} L_{k-1} \alpha_k^2}{2B_k^2} ||z_k - z_{k-1}||_E^2 \right].
\]

But
\[
\frac{\alpha_k}{B_k} (z_k - z_{k-1}) \overset{(18),(20)}{=} w_k - x_k,
\]
and we obtain
\[
\Psi^*_k \geq (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p \\
+ B_k \left[ f_{\delta_k}(x_k) + (g_{\delta_k}(x_k), w_k - x_k) + \frac{2^{\ell_k} L_{k-1}}{2} ||w_k - x_k||_E^2 \right] \\
\overset{(21)}{\geq} (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p + B_k \left[ f_{\delta_k}(w_k) - \frac{\alpha_k \varepsilon}{B_k} - \delta_u \right] \\
\overset{(3)}{\geq} (A_k - B_k) F(y_{k-1}) - E_{k-1} + B_k h(w_k) - 2\delta_p + B_k \left[ f(w_k) - \frac{\alpha_k \varepsilon}{B_k} - 2\delta_u \right] \\
\overset{(23)}{\geq} A_k F(y_k) - E_k - B_k \left[ \frac{\alpha_k \varepsilon}{B_k} + 2\delta_u \right] - 2\delta_p \\
= A_k F(y_k) - E_k. \quad \blacksquare
\]

We are in a 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 [8].

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

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

\[
\leq d(x^*) + \sum_{j=0}^{k} \alpha_j \left[ f_{\delta_j}(x_j) + \langle g_{\delta_j}(x_j), x^* - x_j \rangle + h(x^*) \right]
\]

\[
 \overset{(2)}{\leq} d(x^*) + \sum_{j=0}^{k} \alpha_j \left[ f(x^*) + h(x^*) \right] = d(x^*) + A_k F(x^*).
\]

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 [34], 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 \( \bar{F}_k^d(x) = \sum_{j=0}^{k} \alpha_j \left[ f_{\delta_j}(x_j) + \langle g_{\delta_j}(x_j), x - x_j \rangle \right] \) and

\[
\bar{F}_k = \min_{x \in \mathcal{Q}} \left\{ \frac{1}{A_k} \bar{F}_k^d(x) + h(x) : d(x) \leq D \right\}
\]

\[
= \min_{x \in \mathcal{Q}} \max_{\beta \geq 0} \left\{ \frac{1}{A_k} \bar{F}_k^d(x) + h(x) + \beta(d(x) - D) \right\}
\]

\[
= \max_{\beta = 1/A_k} \min_{x \in \mathcal{Q}} \left\{ \frac{1}{A_k} \bar{F}_k^d(x) + h(x) + \beta(d(x) - D) \right\}
\]

\[
\geq \frac{1}{A_k} \Psi_k^* - \frac{1}{A_k} D.
\]

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

\[
F(y_k) - F^* \leq F(y_k) - \bar{F}_k \overset{(24),(29)}{\leq} \frac{D}{A_k} + \frac{E_k}{A_k}
\]

Thus, we can use stopping criterion

\[
F(y_k) - \bar{F}_k \leq \varepsilon + \frac{2 \delta_u}{A_k} \sum_{j=0}^{k} B_j + \frac{(2k + 1) \delta_p}{A_k},
\]

which ensures

\[
F(y_k) - F^* \leq \varepsilon + \frac{2 \delta_u}{A_k} \sum_{j=0}^{k} B_j + \frac{(2k + 1) \delta_p}{A_k},
\]
as long as
\[ A_k \geq \frac{2D}{\varepsilon}. \]  

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 do not 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 \( (im + 1) \) times. Then total number of oracle calls per iteration is equal to \( 2(im + 1) \). Note that \( L_m = 2^{im}L_{m-1} \).

Therefore, \( i_m = \log_2 \frac{L_m}{L_{m-1}} \). Hence,

\[
N(k) = \sum_{m=1}^{k} 2(i_m + 1) = \sum_{m=1}^{k} 2 \left( \log_2 \frac{L_m}{L_{m-1}} + 1 \right) = \sum_{m=1}^{k} \left[ 2 + 2(\log_2 L_m - \log_2 L_{m-1}) \right] = 2k + 2 \log_2 L_k - 2 \log_2 L_0. \]  

(33)

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,

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

Substituting this estimate in the expression (33), we obtain that on average UIGM needs 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 [8], 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 [8] 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

\[
\alpha_k = \left( \frac{k+2p}{2p} \right)^{p-1} 2^{k-1} L_{k-1}, \quad k \geq 0
\]  

and, in accordance to (16),

\[
B_k = \left( \frac{k+2p}{2p} \right)^{2p-2} 2^{k-1} L_{k-1}, \quad k \geq 0.
\]  

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 \( p \in [0, 1] \)

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

Proof of \( A_k \geq B_k \). For \( k = 0 \) it holds 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,

\[
x^m + y^m - (x + y)^m \geq 0,\]

\[
x^m + y^m \geq (x + y)^m,
\]

\[
x^{p-1} + y^{p-1} \geq (x + y)^{p-1}, \quad p \in [1, 2],
\]

\[
((k - 1 + 2p)^2)^{p-1} + (2(k + 2p))^{p-1} \geq ((k - 1 + 2p)^2 + 2(k + 2p))^{p-1},
\]

\[
(k - 1 + 2p)^{2(p-1)} + (2p(k + 2p))^{p-1} \geq ((k - 1 + 2p)^2 + 2(k - 1 + 2p) + 2)^{p-1},
\]

\[
(k - 1 + 2p)^{2(p-1)} + (2p(k + 2p))^{p-1} \geq (k + 2p)^{2(p-1)},
\]

\[
(k - 1 + 2p)^{2(p-1)} + (2p(k + 2p))^{p-1} \geq ((k - 1 + 2p)^2 + 2(k - 1 + 2p) + 2)^{p-1},
\]

\[
(k - 1 + 2p)^{2(p-1)} + (2p(k + 2p))^{p-1} \geq (k + 2p)^{2(p-1)},
\]
Then, we prove upper bound for Corollary 3.2, we get the explicit rate of convergence of UIGM under the power policy (34).

Proof: The proof is divided into three steps. First, we prove an upper bound for \( \alpha_k \) and \( A_k \).

Now we can obtain the rate of growth of \( \{A_k\}^\infty_{k=0} \). 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 \( k \geq 0 \),

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

**Proof:** The proof is divided into three steps. First, we prove a lower bound for \( \alpha_m \) 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 \( \alpha_m \) and \( A_m \). Since the inner cycle of UIGM for sure ends when \( 2^m L_{m-1} > L(\delta_m) \), we have \( 2^m L_{m-1} \leq 2L(\delta_m) \). Hence,

\[
2^m L_{m-1} \overset{(4),(26)}{\leq} 2 \left[ \frac{\alpha_m}{B_m} \right]^{\frac{1-v}{1+v}} M_v^{\frac{2}{1+v}} \leq 2 \left[ \frac{\left( \frac{m+2p}{2p} \right)^{p-1}}{\varepsilon^{\frac{1}{1+v}}} \right]^{\frac{1-v}{1+v}} M_v^{\frac{2}{1+v}},
\]

\[
\Rightarrow \alpha_m = \frac{\left( \frac{m+2p}{2p} \right)^{p-1}}{2^m L_{m-1}} \geq \frac{\left( \frac{m+2p}{2p} \right)^{2p-v-2v} \varepsilon^{\frac{1-v}{1+v}}}{2M_v^{\frac{2}{1+v}}}.
\]
\[ A_k = \sum_{m=0}^{k} \alpha_m \geq \sum_{m=0}^{k} \left( \frac{m+2p}{2p} \right)^{\frac{2p-2v}{1+v}} 2M_v^\frac{1}{1+v} \geq \frac{1}{2M_v^\frac{1}{1+v}} \sum_{m=0}^{k} \left( \frac{m+2p}{2p} \right)^{\frac{2p-2v}{1+v}}. \]

Since
\[
\sum_{m=0}^{k} \left( \frac{m+2p}{2p} \right)^{\frac{2p-2v}{1+v}} \geq \int_0^k \left( \frac{x+2p}{2p} \right)^{\frac{2p-2v}{1+v}} \, dx + \alpha_0
\]
\[
\geq \frac{2p(1+v)}{2pv - v + 1} \left( \frac{k + 2p}{2p} \right)^{\frac{2p-1}{1+v}} \geq 2 \left( \frac{k + 2p}{2p} \right)^{\frac{2p-1}{1+v}}.
\]

we have
\[
A_k = \sum_{m=0}^{k} \alpha_m \geq \frac{\varepsilon^{\frac{1-v}{2}}}{2M_v^\frac{1}{1+v}} \left( \frac{k + 2p}{2p} \right)^{\frac{2p-1}{1+v}} \geq \frac{\varepsilon^{\frac{1-v}{2}}}{2M_v^\frac{1}{1+v}} \left( \frac{k + 2}{4} \right)^{\frac{2p-1}{1+v}}. (37)
\]

Upper bound for \( B_m \).
\[
B_m^{(32),(35)} \equiv \left( \frac{m + 2p}{2p} \right)^{p-1} \alpha_m,
\]
\[
\sum_{m=0}^{k} B_m = \sum_{m=0}^{k} \left( \frac{m + 2p}{2p} \right)^{p-1} \alpha_m.
\]

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

Proof of (36). Now using (28), (37) and (38) we can get convergence rate.
\[
F(y_k) - F^* \leq \frac{d(x^*)}{A_k} + 2\delta_u \sum_{i=0}^{k} B_i + \frac{\varepsilon}{2} + \frac{(2k + 1)\delta_p}{A_k}
\]
\[
\leq \frac{4}{\varepsilon^{\frac{1-v}{1+v}} (k + 2)^{\frac{2p-1}{1+v}}} \frac{2}{A_k} \sum_{i=0}^{k} B_i + \frac{4}{\varepsilon^{\frac{1-v}{1+v}} (k + 2)^{\frac{2p-1}{1+v}}} (2k + 1) \delta_p + \frac{\varepsilon}{2}
\]
\[
\leq \frac{16M_v^\frac{2}{1+v} d(x^*)}{\varepsilon^{\frac{1-v}{1+v}} (k + 2)^{\frac{2p-1}{1+v}}} + 2\delta_u \sum_{i=0}^{k} B_i + \frac{32M_v^\frac{2}{1+v}}{\varepsilon^{\frac{1-v}{1+v}} (k + 2)^{\frac{2p-1}{1+v}}} \delta_p + \frac{\varepsilon}{2}
\]
\[
\leq \frac{16M_v^\frac{2}{1+v} d(x^*)}{\varepsilon^{\frac{1-v}{1+v}} (k + 2)^{\frac{2p-1}{1+v}}} + 2\delta_u \left( \frac{k + 2p}{2p} \right)^{p-1} + \frac{32M_v^\frac{2}{1+v}}{\varepsilon^{\frac{1-v}{1+v}} (k + 2)^{\frac{2p-1}{1+v}}} \delta_p + \frac{\varepsilon}{2}
\]
\[
\leq \frac{16M_v^\frac{2}{1+v} d(x^*)}{\varepsilon^{\frac{1-v}{1+v}} (k + 2)^{\frac{2p-1}{1+v}}} + 4k^{p-1} \delta_u + \frac{32M_v^\frac{2}{1+v}}{\varepsilon^{\frac{1-v}{1+v}} (k + 2)^{\frac{2p-1}{1+v}}} \delta_p + \frac{\varepsilon}{2}.
\]
Since UIGM does not use $\nu$ as a parameter, we get

$$ F(y_k) - F^* \leq \inf_{\nu \in [0,1]} \left( \frac{16M_0^2}{\varepsilon^{1+\nu}} d(x^*)^{1+\nu} \right) + \frac{32M_0^2}{\varepsilon^{1+\nu}} \delta_p^2 + 4k^{p-1} \delta_u + \frac{\varepsilon}{2}. $$

\[\blacksquare\]

**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 $p \in [1, 2]$,

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

**Proof:** By (17), we have

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

From the rate of convergence (36) and the fact that UIGM does not require the knowledge of $\nu$ to define the iterations, we get the following estimate for the number of iterations, which are necessary to guarantee the first term of (36) to be smaller than $\varepsilon/6$

$$ N = O\left[ \inf_{\nu \in [0,1]} \left( \frac{M_0^2 d(x^*)^{1+\nu}}{\varepsilon^2} \right)^{\frac{1}{2p-1}} \right]. $$

The dependence of this bound on the smoothness parameters is optimal (see [31]).

Let us compare the proposed method and its convergence rate with the existing optimal methods for different classes of problems. If $N$ is the number of iterations, then $F(y_N) - F^* \leq B(N) + C(N)\delta_p + D(N)\delta_u + \frac{\varepsilon}{2}$. Here $B(N)$ characterises the convergence rate, $C(N)$ characterises the accumulation of the prox-mapping error and $D(N)$ characterises the accumulation of the oracle error. The result can be summarized in the following table.

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

For non-smooth functions ($\nu = 0$), the convergence rate of UIGM for any $p \in [1, 2]$ coincides with the rate of convergence of subgradient methods. These methods are robust to oracle error, but accumulate the 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. Finally, for $p = 2$ UIGM has the same rate as a fast gradient method. This method accumulates the oracle error, but the effective prox-mapping error decreases with iterations. 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 $F$

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

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

$$d(x) \leq \frac{\Omega}{2} \|x\|^2_E,$$  \hspace{1cm} (39)

where $\Omega$ is some dimension-dependent constant, and that we are given a starting point $x_0$ and a number $R_0$ such that

$$\|x_0 - x_*\|^2_E \leq R_0^2,$$ \hspace{1cm} (40)

where $x_*$ is an optimal point in (1).

Algorithm 2 Restart UIGM

**Require:** $\mu$ – strong convexity parameter, $\Omega$ – quadratic growth constant, $\epsilon$ – desired accuracy, $x_0$ – starting point.

1: Set $d_0(x) = d(x - x_0)$.

2: for $m = 1, \ldots$ do

3: \hspace{1cm} while $2\Omega > \mu A_k$ do

4: \hspace{4cm} Run UIGM with accuracy $\epsilon$ and prox-function $d_{m-1}(x)$.

5: \hspace{1cm} Set $x_m = y_k$.

6: \hspace{1cm} Set $d_m(x) = d(x - x_m)$.

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

$$F(x_m) - F(x_*) \leq \mu R_0^2 2^{-m-1}$$

$$+ 2 \left( \frac{\epsilon}{2} + 2 \left\lceil \left( \frac{2^{v+1} \Omega^{v+1} M_v^2}{\mu^{v+1} \epsilon^{1-v}} \right)^{\frac{p-1}{2pv-v-1}} \right\rceil \delta_u + \frac{p^2}{2p^2 - v - 1} \frac{\mu^2}{\Omega^{2v-p-v+1}} \frac{M_v^2}{\epsilon^{2p-2v+1}} \delta_p \right),$$

$$\|x_m - x_*\|^2_E \leq R_m^2 = R_0^2 2^{-m}$$

$$+ \frac{4}{\mu} \left( \frac{\epsilon}{2} + 2 \left\lceil \left( \frac{2^{v+1} \Omega^{v+1} M_v^2}{\mu^{v+1} \epsilon^{1-v}} \right)^{\frac{p-1}{2pv-v-1}} \right\rceil \delta_u + \frac{p^2}{2p^2 - v - 1} \frac{\mu^2}{\Omega^{2v-p-v+1}} \frac{M_v^2}{\epsilon^{2p-2v+1}} \delta_p \right).$$

(41)

(42)
Proof: From (39), by the first-order optimality condition, we have

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

Whence,

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

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

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

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

For \( m = 1 \), we have

\[
F(x_1) - F(x^*) \leq \frac{d_0(x^*)}{A_k} + 2 \sum_{i=0}^{k} B_i \delta_u + \frac{(2k + 1) \delta_p}{A_k} + \frac{\varepsilon}{2}
\]

\[
\leq \frac{\Omega \| x_0 - x^* \|^2_E}{2A_k} + 2 \sum_{i=0}^{k} B_i \delta_u + \frac{(2k + 1) \delta_p}{A_k} + \frac{\varepsilon}{2}
\]

\[
\leq \frac{\Omega R_0^2}{2A_k} + 2 \sum_{i=0}^{k} B_i \delta_u + \frac{(2k + 1) \delta_p}{A_k} + \frac{\varepsilon}{2}
\]

\[
\leq \frac{\Omega R_0^2}{2A_k} + 2 \left( \frac{k + 2p}{2p} \right)^{p-1} \delta_u + \frac{(2k + 1) \delta_p}{A_k} + \frac{\varepsilon}{2}.
\]

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

\[
A_{k-1} < \frac{2\Omega}{\mu} \leq A_k,
\]

and

\[
\frac{2\Omega}{\mu} \geq A_{k-1} \geq \frac{\varepsilon^{1-v}}{2^{v+1} \mu^{v+1}} \left( \frac{k - 1 + 2p}{2p} \right)^{\frac{2p(v-1)+1}{1+v}},
\]

\[
\Rightarrow \frac{2^{p(v-1)+1} \Omega^{p(v-1)+1} M_v^{2p(v-1)+1}}{\mu^{p(v-1)+1} \varepsilon^{1-v}} \geq \left( \frac{k - 1 + 2p}{2p} \right)^{\frac{2p(v-1)+1}{1+v}},
\]

\[
\Rightarrow p^2 \frac{2^{p(v-1)+1} \Omega^{p(v-1)+1} M_v^{2p(v-1)+1}}{\mu^{p(v-1)+1} \varepsilon^{2p(v-1)+1}} + 1 - 2p \geq k,
\]

\[
\Rightarrow p^2 \frac{2^{p(v-1)+1} \Omega^{p(v-1)+1} M_v^{2p(v-1)+1}}{\mu^{p(v-1)+1} \varepsilon^{2p(v-1)+1}} + 1 - 2p \geq k.
\]
Hence,

\[
\left( \frac{k + 2p}{2p} \right)^{p-1} \leq \left[ \left( \frac{2^{v+1} \Omega^{1+v} M_v^2}{\mu^{v+1} e^{1-v}} \right) \right]^{\frac{p-1}{2p-v}} ,
\]

(44)

\[
\Rightarrow 2k + 1 \leq \frac{p2^{2p-v+3} \Omega^{1+v} M_v^{2p-v+1}}{\mu^{p-v+1} e^{p-v+1}} A_k .
\]

(45)

Finally, we have

\[
F(x_1) - F(x^*) \leq \frac{\Omega R_0^2}{2A_k} + 2 \left[ \left( \frac{2^{v+1} \Omega^{1+v} M_v^2}{\mu^{v+1} e^{1-v}} \right) \right]^{\frac{p-1}{2p-v}} \delta_u
\]

\[
+ \frac{p2^{2p-v+3} \Omega^{1+v} M_v^{2p-v+1}}{\mu^{p-v+1} e^{p-v+1}} A_k \delta_p + \frac{\varepsilon}{2}
\]

\[
\leq \mu R_0^2 2^{-2} + 2 \left( 1 - 2^{-1} \right) \left( \frac{\varepsilon}{2} + 2 \left[ \left( \frac{2^{v+1} \Omega^{1+v} M_v^2}{\mu^{v+1} e^{1-v}} \right) \right]^{\frac{p-1}{2p-v}} \delta_u
\]

\[
+ \frac{p2^{2p-v+3} \Omega^{1+v} M_v^{2p-v+1}}{\mu^{p-v+1} e^{p-v+1}} A_k \delta_p \right) .
\]

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

\[
F(x_{m+1}) - F(x^*) \leq \frac{d_m(x^*)}{A_k} + 2 \sum_{i=0}^{k} B_i \delta_u + \frac{(2k + 1) \delta_p}{A_k} + \frac{\varepsilon}{2}
\]

(39)

\[
\leq \frac{\Omega \|x_m - x^*\|_E^2}{2A_k} + 2 \sum_{i=0}^{k} B_i \delta_u + \frac{(2k + 1) \delta_p}{A_k} + \frac{\varepsilon}{2}
\]

(42)

\[
\leq \frac{\Omega R_m^2}{2A_k} + 2 \sum_{i=0}^{k} B_i \delta_u + \frac{(2k + 1) \delta_p}{A_k} + \frac{\varepsilon}{2}
\]

(38)

\[
\leq \frac{\Omega R_m^2}{2A_k} + 2 \left( \frac{k + 2p}{2p} \right)^{p-1} \delta_u + \frac{(2k + 1) \delta_p}{A_k} + \frac{\varepsilon}{2}
\]

(44),(45)

\[
\leq \frac{\Omega R_m^2}{2A_k} + 2 \left[ \left( \frac{2^{v+1} \Omega^{1+v} M_v^2}{\mu^{v+1} e^{1-v}} \right) \right]^{\frac{p-1}{2p-v}} \delta_u
\]

\[
+ \frac{p2^{2p-v+3} \Omega^{1+v} M_v^{2p-v+1}}{\mu^{p-v+1} e^{p-v+1}} A_k \delta_p + \frac{\varepsilon}{2}
\]

(43)

\[
\leq \frac{\mu R_m^2}{4} + 2 \left[ \left( \frac{2^{v+1} \Omega^{1+v} M_v^2}{\mu^{v+1} e^{1-v}} \right) \right]^{\frac{p-1}{2p-v}} \delta_u
\]
\begin{align*}
\leq \mu R_0^2 2^{-m-2} + \frac{2v(p-1)}{4\mu} \left( 1 - 2^{-m} \right) \left( \frac{\varepsilon}{2} + 2 \left[ \left( \frac{2^{v+1} \Omega^{v+1} M_v^2}{\mu^{v+1} \varepsilon^{1-v}} \right) \right] \right)^{\frac{p-1}{2^{p(v-1) v}}} \delta_u \\
+ \frac{2^{p(v-1) v} \mu^{p(v-1) v} M_v^{2^{p(v-1) v+1} v}}{\Omega^{2^{p(v-1) v+1} v}} \delta_p
\end{align*}

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

**Corollary 5.2:** To obtain \((\varepsilon + C_u)\)-solution of problem (1), where

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

we need

\[ \tilde{i} = \left\lceil \log \left( \frac{\mu R_0^2}{2\varepsilon} \right) \right\rceil \]

restarts and

\[ \tilde{k} \leq \inf_{0 \leq v \leq 1} \left( \frac{\Omega^{1+v} 2^{4p(v-1)v+3} M_v^2}{\mu^{1+v} \varepsilon^{1-v}} \right) \frac{1}{2^{p(v-1) v+1}} + 1 \]
iterations of UIGM per iteration. The total number of UIGM iterations is no more than

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

**Proof:**

\[
F(x_l) - F(x^*) \leq \mu R_0^2 \frac{\epsilon + Cu}{2^{\nu}} \leq \epsilon + Cu.
\]

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

\[
A_{\tilde{k}} \geq \frac{2\Omega}{\mu} \geq A_{\tilde{k}-1} \geq \frac{\epsilon^{1+v}}{M^{\nu\nu}} \left( \frac{\tilde{k} - 1}{4} \right)^{\frac{2^{p\nu-v+1}}{1+v}},
\]

\[
\Rightarrow \frac{2\Omega \frac{2^{p\nu-v+1}}{1+v} M^{\nu\nu}}{\mu \epsilon^{1+v}} \geq \left( \tilde{k} - 1 \right)^{\frac{2^{p\nu-v+1}}{1+v}},
\]

\[
\Rightarrow \left( \frac{\Omega^{1+v} 2^{4p\nu - v + 3} \mu^2}{\mu^{1+v} \epsilon^{1-v}} \right)^{\frac{1}{2^{p\nu-v+1}}} \geq \tilde{k} - 1.
\]

Since the algorithm does not use any particular choice of \(\nu\), we have

\[
\tilde{k} \leq \inf_{0 \leq \nu \leq 1} \left( \frac{\Omega^{1+v} 2^{4p\nu - v + 3} \mu^2}{\mu^{1+v} \epsilon^{1-v}} \right)^{\frac{1}{2^{p\nu-v+1}}} + 1.
\]

Then the total number of UIGM is no more than \(N = \tilde{k} \cdot \tilde{l}\), and we have

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

Now we compare our result with existing methods in the same manner as for convex functions. The objective residual is characterised by the quantity \(C_u = C\delta_p + D\delta_u + \frac{\epsilon}{2}\), where \(C\) corresponds to prox-mapping error for given \(\epsilon, \mu\) and \(D\) corresponds to oracle error for given \(\epsilon, \mu\). The results are summarised in the following table.

For non-smooth functions \((\nu = 0)\), the convergence rate of Restart UIGM for any \(p \in [1,2]\) coincides with the rate of convergence of subgradient methods. These methods are robust to oracle error, but accumulate 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. Finally, for \(p = 2\) Restart UIGM has the same rate as a fast gradient method. This method collects oracle error but the effective prox-mapping error decreases with iterations. This table shows three main regimes for Restart UIGM and how it corresponds to classical methods.
6. Switching policy

In this section, we describe another variant of policy for coefficient sequence \( \{\alpha_k, B_k\}_{k \geq 0} \). The key observation is that Fast Gradient Method (FGM) accumulates the error, but converges faster and Dual Gradient Method (DGM) does not accumulate the oracle error, but converges slower. That is why the idea of the switching policy is to start with a number of steps of FGM until the error reaches some limit and then switch to DGM steps. A similar policy was introduced in [8]. Now we need to understand, what is the limit of the error corresponding to the switching moment. If we want to get the total error equal to \( \varepsilon \), then the error from the inexactness should be \( \varepsilon / 2 \). Now we describe this idea in more details.

Let the switching policy be

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

where \( s \) is the moment of switching and \( c_k \) is some constant, which we describe below.

Firstly, we should prove, that the switching policy is feasible for the UIGM. Let us check the correctness of inequalities (12) for the switching policy. For FGM steps the correctness easily follows from (4.1), since it is the power policy with \( p = 2 \). 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,

\[
1 \leq c_k \leq \frac{1 + \sqrt{1 + 4A_{k-1}L_k}}{2}
\] (48)

So if these two conditions for \( c_k \) are satisfied we have proved that the switching policy is feasible for UIGM.

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

\[
\frac{2d_u}{A_k} \sum_{j=0}^{k} B_j \leq \frac{\varepsilon}{6}
\] (49)
\[
\frac{(2k+1)\delta_p}{A_k} \leq \frac{\varepsilon}{6}.
\]

Note that for these 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) always hold.
- Only DGM steps. In this case, \( \delta_u \) is rather large and (49) does not hold already on the first iteration, so we perform only slow DGM steps.
- Switching at the moment \( s \). In this case, we do some FGM steps until the moment \( s \), when (49) does not hold any more for the first time and next make only DGM steps.

First two regimes are easy for understanding and for the last one we requires a more careful analysis. Note that for FGM steps (50) always true because the left hand side decreases on each step. Further, 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) holds.

\[
\sum_{j=0}^{k} B_j \leq \frac{\varepsilon}{12\delta_u} \sum_{j=0}^{k} \alpha_j
\]

\[
\sum_{j=0}^{k-1} B_j + B_k \leq \frac{\varepsilon}{12\delta_u} \left( \sum_{j=0}^{k-1} \alpha_j + \alpha_k \right).
\]

We assume that on the previous step (49) holds, that is why we need

\[
B_k \leq \frac{\varepsilon}{12\delta_u} \alpha_k
\]

\[
\frac{c_k^2}{L_k} \leq \frac{\varepsilon}{12\delta_u} \frac{c_k}{L_k}.
\]

So we get the third condition on \( c_k \)

\[
c_k \leq \frac{\varepsilon}{12\delta_u}.
\]

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

\[
c_k = \min \left( \frac{\varepsilon}{12\delta_u}, 1 + \sqrt{1 + \frac{A_{k-1}L_k}{2}} \right).
\]

As a result, we prove that the switching policy is feasible for UIGM and that UIGM converges, when we do FGM steps until at the moment \( s \) (49) does not hold any more, then switch to DGM with \( c_k \) defined by (52). Note that now our method needs to know only \( \varepsilon, \delta_u, \delta_p \) and does not need the parameter \( p \) as an input as opposed to the power policy. Hence it converges no slower than UIGM with power policy for any \( p \).
Theorem 6.1: Assume that $f$ is a convex function with inexact first-order oracle. Then, for the sequence (47), the moment $s$ being the first time when (49) does not hold and $c_k$ defined by (52), for all $k \geq 0$,

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

The same argument holds 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). PET plays an important role in medicine for detecting cancer and metabolic changes in human body. PET can be treated as a Poisson likelihood model [3,27]. The estimation of radioactivity density within an organ corresponds to the following convex non-smooth optimization problem:

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

where $\Delta_n$ is the standard simplex. $A$ is a data and refers to the likelihood matrix known from geometry of the detector, and $w$ is the 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 estimate $x$ from the data. For simplicity, we will not consider any regularization for this application. Note that in fact, the objective in this problem has unbounded $M_\nu$ since $\nabla \log y = 1/y$, which is not bounded as $y \to 0$. So we assume that the iterates of our method are separated from zero and then $M_\nu$ is bounded by some constant. Nevertheless, this constant may be large and it is desirable to adapt to some local value of this constant which can be much smaller than the global estimate.

We assume, that tomographic scanner can have some small random and systematic errors, which leads to inexact data and hence inexact function values and gradients. If a method converges with inexact data, we have robust system and even with errors we can obtain rather precise tomography.

Since the minimization problem is on the standard simplex, the entropy function $d(x) = \sum_{i=1}^{n} x_i \log(x_i)$ is a good. Moreover, the prox-mapping can be calculated by an explicit formula [8], 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 an additional optimization subproblem. For example we can approximately solve it by FGM with $\delta_p > 0$, because this subproblem is strongly convex and FGM had linear convergence in this case.

We conduct experiments using Ubuntu 14, Python 3, and computer with Intel Core i7-4510U CPU 2.00 GHz 2.60 GHz, 8 Gb 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 $\delta_u$. Desired accuracy is $\epsilon = 0.0001$.

For small inexactness $\delta = 0.001\epsilon$ UIGM give us next graphic (Figure 1).
Figure 1. Comparison of different power policies and switching policy for small inexactness.

Figure 2. Comparison of different power policies and switching policy for medium inexactness.

From this plot, we can see that the power policy with $p = 2$ and switching policy are the fastest. For small inexactness all the tested choices of the parameters do not lead to any noticeable error accumulation.

In the next plot, we can see that for a medium error $\delta = \varepsilon$ the switching policy starts to work as power policy with $p = 1$. The reason could be that all our estimates of error accumulation come from theory but the real error at a specific point can be less than the theoretical estimate. Unfortunately the real error sometimes is not available for measurement (Figure 2).

For a large error $\delta = 1000\varepsilon$ the power policy with $p = 2$ leads to error accumulation and the method works slower than with the power policy $p = 1.5$. So the method with intermediate rate is the best one since it is rather fast and also robust (Figure 3).

As a result, we get that UIGM for some intermediate $p$ can be better than classical methods. Also we get that our method has a speed up for non-smooth problem in comparison
with optimal DGM ($p = 1$). Unfortunately, in practice the 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 [13]. See also [1,6,11,12,14,26]. Also, based on the ideas from [15,16,36], UIGM for the strongly convex case can be modified to work without exact knowledge of strong convexity parameter $\mu$. Finally, similarly to [10,17,18,20,23], UIGM can be modified to solve convex problems with stochastic inexact oracle. In the future it would be interesting to generalize this method to work with inexact model of the objective function [24,37,38], which is a generalization of inexact oracle. Another direction is to generalize it for non-convex optimization using the ideas of adaptive methods with line-search for non-convex optimization with inexact oracle [5,35].

Acknowledgments

The authors are very grateful to the anonymous reviewers for their suggestions and comments which allowed to improve the readability, clarity, and correctness of the paper.
Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This research was funded by Russian Science Foundation (project 17-11-01027). The work of A.V. Gasnikov on Section 4 was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project ‘5–100’.

Notes on contributors

**Dmitry Kamzolov**

got B.Sc. and M.Sc. degrees from the Faculty of Control and Applied Mathematics of Moscow Institute of Physics and Technology (MIPT) in 2016 and 2018, respectively. Also, he obtained an M.Sc. in Operations Research, Combinatorics and Optimization from the University of Grenoble Alpes. Since 2018 he has been pursuing a Ph.D. in MIPT.

**Pavel Dvurechensky**

got B.Sc and M.Sc degrees from the Faculty of Control and Applied Mathematics of Moscow Institute of Physics and Technology in 2008 and 2010, respectively. He obtained a Ph.D. from the same university in 2014. Since 2015 he has been a research associate at Weierstrass Institute for Applied Analysis and Stochastics in Berlin.

**Alexander V. Gasnikov**

received the B.Sc, M.Sc, Ph.D., Dr. habil. degrees from the Department of Control and Applied Mathematics of Moscow Institute of Physics and Technology (Russia) in 2004, 2006, 2007, and 2016, respectively. Since 2011, he has been an Associate Professor with the Department of Mathematical Foundations of Control, MIPT. Also, since 2015, he has been the Lead Researcher with the Institute for Information Transmission Problems of Russian Academy of Science, Moscow, Russia, and an Associate Professor with the Faculty of Computer Science High school of economics.

**ORCID**

Dmitry Kamzolov [http://orcid.org/0000-0001-8488-9692](http://orcid.org/0000-0001-8488-9692)
Pavel Dvurechensky [http://orcid.org/0000-0003-1201-2343](http://orcid.org/0000-0003-1201-2343)
Alexander V. Gasnikov [http://orcid.org/0000-0002-7386-039X](http://orcid.org/0000-0002-7386-039X)

**References**

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

[2] D. R. Baimurzina, A. V. Gasnikov, E. V. Gasnikova, P. E. Dvurechensky, E. I. Ershov, M. B. Kuben-taeva, and A. A. Lagunovskaya, *Universal method of searching for equilibria and stochastic equilibria in transportation networks*, Comput. Math. Math. Phys. 59 (2019), pp. 19–33. arXiv:1701.02473.

[3] A. Ben-Tal, T. Margalit, and A. Nemirovski, *The ordered subsets mirror descent optimization method with applications to tomography*, SIAM J. Optim. 12 (2001), pp. 79–108.

[4] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization (Lecture notes)*, Personal web-page of A. Nemirovski, 2015. Available at [http://www2.isye.gatech.edu/nemirovs/Lect_ModConvOpt.pdf](http://www2.isye.gatech.edu/nemirovs/Lect_ModConvOpt.pdf).

[5] L. Bogolubsky, P. Dvurechensky, A. Gasnikov, G. Gusev, Y. Nesterov, A. M. Raigorodskii, A. Tikhonov, and M. Zhukovskii, *Learning supervised pagerank with gradient-based and gradient-free optimization methods*, in Advances in Neural Information Processing Systems 29, D. D. Lee,
[6] A. Chernov, P. Dvurechensky, and A. Gasnikov, Fast primal-dual gradient method for strongly convex minimization problems with linear constraints, Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings, Springer International Publishing, 2016, pp. 391–403.

[7] A. d’Aspremont, Smooth optimization with approximate gradient, SIAM J. Optim. 19 (2008), pp. 1171–1183.

[8] O. Devolder, F. Glineur, and Y. Nesterov, Intermediate gradient methods for smooth convex problems with inexact oracle (2013), CORE Discussion Paper 2013/17.

[9] O. Devolder, F. Glineur, and Y. Nesterov, First-order methods of smooth convex optimization with inexact oracle, Math. Program. 146 (2014), pp. 37–75.

[10] P. Dvurechensky and A. Gasnikov, Stochastic intermediate gradient method for convex problems with stochastic inexact oracle, J. Optim. Theory Appl. 171 (2016), pp. 121–145.

[11] P. Dvurechensky, A. Gasnikov, and A. Kroshnin, Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn’s algorithm, Stockholm, Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, Vol. 80, arXiv:1802.04367, 2018, pp. 1367–1376.

[12] P. Dvurechensky, A. Gasnikov, E. Gasnikova, S. Matsievsky, A. Rodomanov, and I. Usik, Primal-dual method for searching equilibrium in hierarchical congestion population games, Supplementary Proceedings of the 9th International Conference on Discrete Optimization and Operations Research and Scientific School (DOOR 2016) Vladivostok, Russia, September 19–23, 2016, arXiv:1606.08988, 2016, pp. 584–595.

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

[14] P. Dvurechensky, D. Dvinskikh, A. Gasnikov, C.A. Uribe, and A. Nedić, Decentralize and randomize: faster algorithm for wasserstein barycenters, in Advances in Neural Information Processing Systems 31, NeurIPS 2018, arXiv:1806.03915, Curran Associates, Inc., 2018, pp. 10783–10793.

[15] O. Fercoq and Z. Qu, Restarting accelerated gradient methods with a rough strong convexity estimate, preprint arXiv:1609.07358 (2016).

[16] O. Fercoq and Z. Qu, Adaptive restart of accelerated gradient methods under local quadratic growth condition, IMA J. Numer. Anal. 39 (2019), pp. 2069–2095.

[17] A. Gasnikov, P. Dvurechensky, and D. Kamzolov, Gradient and gradient-free methods for stochastic convex optimization with inexact oracle, Dynamics systems and control processes. Proceedings of the International Conference Dedicated to the 90th Anniversary of the Birth of Academic N. N. Krasovsky. (2015), pp. 111–117. Available at https://arxiv.org/abs/1502.06259.

[18] A. Gasnikov, P. Dvurechensky, and Y. Nesterov, Stochastic gradient methods with inexact oracle, Proc. Moscow Inst. Phys. Technol. 8 (2016), pp. 41–91. in Russian, first appeared in arXiv:1411.4218.

[19] A. Gasnikov, D. Kamzolov, and M. Mendel, Universal composite prox-method for strictly convex optimization problems, TRUDY MIPT 8 (2016), pp. 25–42. Available at https://arxiv.org/abs/1603.07701.

[20] A. Gasnikov and Y. Nesterov, Universal fast gradient method for stochastic composite optimization problems, Comput. Math. Math. Phys. 58 (2018), pp. 51–68. Available at https://arxiv.org/abs/1604.05275.

[21] A. Gasnikov, P. Dvurechensky, D. Kamzolov, Y. Nesterov, V. Spokoiny, P. Stetsyuk, A. Suvorikova, and A. Chernov, Universal method with inexact oracle and its applications for searching equilibriums in multistage transport problems, TRUDY MIPT 7 (2015), pp. 143–155. Available at https://arxiv.org/abs/1506.00292.

[22] A. Gasnikov, S. Kabanikhin, A. Mohamed, and M. Shishlenin, Convex optimization in Hilbert space with applications to inverse problems, preprint arXiv:1703.00267 (2017).
[23] A.V. Gasnikov and P.E. Dvurechensky, *Stochastic intermediate gradient method for convex optimization problems*, Doklady Math. 93 (2016), pp. 148–151.

[24] A.V. Gasnikov and A.I. Tyurin, *Fast gradient descent for convex minimization problems with an oracle producing a (δ, l)-model of function at the requested point*, Comput. Math. Math. Phys. 59 (2019), pp. 1085–1097. doi:10.1134/S0965542519070078.

[25] S. Ghadimi, G. Lan, and H. Zhang, *Generalized uniformly optimal methods for nonlinear programming*, J. Sci. Comput. 79 (2019), pp. 1854–1881.

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

[27] N. He, Z. Harchaoui, Y. Wang, and L. Song, *Fast and simple optimization for poisson likelihood models*, preprint arXiv:1608.01264 (2016).

[28] A. Juditsky and Y. Nesterov, *Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization*, Stochastic Syst. 4 (2014), pp. 44–80. doi:10.1287/10-SSY010.

[29] G. Lan, *An optimal method for stochastic composite optimization*, Math. Program. 133 (2012), pp. 365–397.

[30] A. Nemirovskii and Y. Nesterov, *Optimal methods of smooth convex minimization*, USSR Comput. Math. Math. Phys. 25 (1985), pp. 21–30.

[31] A. Nemirovsky and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, J. Wiley & Sons, New York, 1983.

[32] Y. Nesterov, *A method for unconstrained convex minimization problem with the rate of convergence O(1/k²)*, in Doklady AN USSR, Vol. 269, 1983, pp. 543–547.

[33] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program. 140 (2013), pp. 125–161.

[34] Y. Nesterov, *Universal gradient methods for convex optimization problems*, Math. Program. 152 (2015), pp. 381–404.

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

[36] V. Roulet and A. d’Aspremont, *Sharpness, restart and acceleration*, in *Advances in Neural Information Processing Systems*, 2017, pp. 1119–1129.

[37] Y. Nesterov, *Inexact model: a framework for optimization and variational inequalities*, arXiv:1902.00990 (2019).

[38] F.S. Stonyakin, D. Dvinskikh, P. Dvurechensky, A. Kroshnin, V. Piskunova, *Inexact model: a framework for optimization and variational inequalities*, arXiv:1902.09901 (2019).

[39] A. Yurtsever, Q. Tran-Dinh, and V. Cevher, *A universal primal-dual convex optimization framework*, Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS’15, Montreal, Canada, MIT Press, Cambridge, MA, USA, 2015, pp. 3150–3158.