Math. Program., Ser. A (2015) 152:381–404
DOI 10.1007/s10107-014-0790-0

FULL LENGTH PAPER

# Universal gradient methods for convex optimization problems

**Yu Nesterov**

Received: 18 April 2013 / Accepted: 16 May 2014 / Published online: 28 May 2014


**Abstract** In this paper, we present new methods for black-box convex minimization. They do not need to know in advance the actual level of smoothness of the objective function. Their only essential input parameter is the required accuracy of the solution. At the same time, for each particular problem class they automatically ensure the best possible rate of convergence. We confirm our theoretical results by encouraging numerical experiments, which demonstrate that the fast rate of convergence, typical for the smooth optimization problems, sometimes can be achieved even on nonsmooth problem instances.

**Keywords** Convex optimization · Black-box methods · Complexity bounds · Optimal methods · Weakly smooth functions

**Mathematics Subject Classification** 90C25 · 90C47 · 68Q25

## 1 Introduction

**Motivation** In Convex Optimization, the majority of numerical schemes are developed for particular problem classes. In the Black-Box framework, two main classes of

The research results presented in this paper have been supported by a grant "Action de recherche concertè ARC 04/09-315" from the "Direction de la recherche scientifique - Communautè française de Belgique". The author also acknowledges the support from Laboratory of Structural Methods of Data Analysis in Predictive Modelling, through RF government grant 11.G34.31.0073, and RFBR research projects 13-01-12007 ofi_m, 14-01-00722-a.

Yu. Nesterov (✉)
Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium
e-mail: yurii.nesterov@uclouvain.be

convex problems, the smooth problems, and nonsmooth ones are treated by completely different techniques.

This separation looks very natural. Indeed, differentiable functions allow constructing *monotone* minimization sequences, for which the convergence results can be easily obtained. Smooth functions can be locally approximated by first- and second-order models, which are very helpful in developing efficient minimization schemes.

The class of nonsmooth convex functions looks much more difficult. For them, there is no hope to get a good local approximation model. It is very difficult to construct relaxation sequences. Moreover, even if a descent direction is found, there is no guarantee that we can advance along it by a sufficiently long step. Therefore, the majority of methods for nonsmooth convex optimization rely only on separation properties. Cutting planes provide us with information about the half-spaces containing the optimal solution. Using this very restricted knowledge, it is still possible to develop some optimization methods. But their computational abilities are incomparably weaker than the abilities of smooth minimization schemes.

Above observations are confirmed by theoretical results. It is well known that for the class of smooth problems $C^{1,1}(\mathbb{R}^n)$, composed by functions with Lipschitz-continuous gradients, the optimal iteration complexity bound for finding $\epsilon$-solution of corresponding optimization problem by a first-order method is of the order $O(\frac{1}{\epsilon^{1/2}})$. For nonsmooth problems from the class $C^{1,0}(\mathbb{R}^n)$, where we can rely only on Lipschitz continuity of function values, such a bound is established on the level of $O(\frac{1}{\epsilon^2})$ (see, e.g. [9]).

Such a big difference in the complexity bounds stimulated an interest to the intermediate classes of convex problems. One of the possibilities consists in considering functions from the class $C^{1,\nu}(\mathbb{R}^n)$, $\nu \in [0, 1]$, which have Hölder continuous gradients:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L_\nu \|x - y\|^\nu, \quad x, y \in \mathbb{R}^n. \tag{1.1}$$

General Complexity Theory [8] established for this class the following lower iteration complexity bound:

$$O\left(\left(\frac{L_\nu R^{1+\nu}}{\epsilon}\right)^{\frac{2}{1+3\nu}}\right), \tag{1.2}$$

where $R$ is the distance from a starting point to the solution. The first optimal methods for such problems were developed in [7].[1] The main advantage of these schemes is an automatic adjustment to the proper level of smoothness parameter $\nu$. However, these methods need to know other characteristics of the problem (estimate of Lipschitz constant $L_\nu$, estimate of the distance to optimum), which are not readily available. Moreover, it was necessary to decide in advance on the total number of steps of the method. This requirement is not very practical. Indeed, in order to make such a decision, we need to know the rate of convergence of the method. However, this is possible only if we know the Hölder parameter. This hidden contradiction probably

[1] English translation of this paper was included in Sect. 2.3 in [4].

explains why these theoretically attractive procedures were never seriously tested in computational practice.

In the last decade, we can see a restoration of interest in the gradient methods. New problems in image processing, data mining, and statistics require computationally cheap minimization procedures, which can quickly deliver an approximate solutions with a moderate accuracy. This demand was satisfied by new families of problem-oriented methods (e.g. [2,10,11]), which increase the rate of convergence of the gradient schemes much above the limits of Black-Box Complexity Theory [8]. This can be done, of-course, only by an appropriate use of problem structure, violating one of the main assumptions of the Black-Box concept.

However, it appears that the Black-Box methods did not reach yet the limits of their performance. The old idea of automatic adjustment to Hölder parameter was revived in [5], where a new version of Level Method [6] was adapted to smooth problems, ensuring the best possible complexity bounds for all values of the smoothness parameter. The only drawback of this approach is related to a high iteration cost of the Level Method.

Minimization of functions with Hölder-continuous gradient was discussed in [3] in the framework of *inexact oracle*. It was shown that the answer $(f(x), \nabla f(x))$ of an *exact* oracle for a convex function satisfying Hölder condition (1.1) can be treated as "inexact" information for some function from $C^{1,1}(\mathbb{R}^n)$:

$$0 \le f(y) - f(x) - \langle \nabla f(x), y - x\rangle \le \tilde{\delta} + \tfrac{1}{2}\tilde{L}\|y-x\|^2, \quad x, y \in \mathbb{R}^n, \quad (1.3)$$

where $\tilde{L}$ and $\tilde{\delta}$ are some "inexactness" parameters. It was shown that these parameters can be chosen as appropriate functions of $\nu$. Therefore, functions from $C^{1,\nu}(\mathbb{R}^n)$ can be minimized by an "inexact" version of Fast Gradient Methods for $C^{1,1}(\mathbb{R}^n)$. The resulting complexity bounds appear to be optimal (1.2). However, in order to apply this approach, we need to employ a lot of additional information (the values of parameters $\nu$, $L_\nu$, and $R$ involved in the bound (1.2), and the total number of steps of the method).

In this paper, we construct new *universal methods* for minimizing composite functions with black-box component satisfying Hölder condition for its gradient. They do not need à priori knowledge of the parameter $\nu$, and they have a cheap cost of one iteration.

In order to apply universal methods to *composite minimization* problem (see [11])

$$\min_{x\in Q} [\, f(x) + \Psi(x)\,] \quad (1.4)$$

with simple convex part $\Psi$ (see Sect. 2 for details), we suggest to use for $f$ a *Damped Relaxation Condition* (DR)

$$f(x_+) \le \delta + \min_{y\in Q}\left[ f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x}\rangle + \tfrac{1}{2}\hat{L}\|y-\bar{x}\|^2\right], \quad (1.5)$$

verified at some test points $\bar{x}$ and $x_+$ of the methods. The tolerance parameter $\delta > 0$ in (1.5) depends only on the required *accuracy* $\epsilon > 0$ of approximate solution. Similar conditions were used in [7] and [3] with $\delta$ being a function of smoothness parameter

$\nu$ and total number of iterations. We show that all necessary information on $\nu$ and $L_\nu$ can be accumulated in the constant $\hat{L}$, which can be easily adapted by an appropriate "line-search" strategy.

For different methods, the dependence of $\delta$ in $\epsilon$ must be different. For the simplest Primal and Dual Gradient Methods, it is enough to take $\delta = \frac{\epsilon}{2}$. For the Fast Gradient Method [10], we use condition (1.5) with much smaller value of $\delta$, allowing to maintain a damped version of the estimating sequence condition

$$A_k(f(x_k) - \tfrac{\epsilon}{2}) \leq \min_{x \in Q} \phi_k(x), \quad k \geq 0, \tag{1.6}$$

where $\{x_k\}_{k \geq 0}$ is the minimizing sequence generated by the method, and $\{\phi_k(\cdot)\}_{k \geq 0}$ is the sequence of estimating functions (see Sect. 4 for details).

For our methods, the space of variables of the composite minimization problem (1.4) can be endowed with arbitrary norm. Hence, we apply the machinery of Bregman distances. The proposed schemes adjust automatically to the actual level of smoothness of the smooth part of the objective function. The only essential input parameter for these schemes is the required accuracy $\epsilon > 0$.

**Contents** The paper is organized as follows. In Sect. 2 we introduce the problem formulation and discuss the main properties of Bregman mapping as applied to functions with Hölder continuous gradients. After that, we prove a convergence result for Universal Primal Gradient Method and derive its complexity bound. We show that this method needs on average at most two calls of oracle per iteration. Moreover, this method can be equipped with a reliable stopping criterion, provide that we have a reasonably good bound for the initial Bregman distance to the optimal set.

In Sect. 3, we prove similar results for Universal Dual Gradient Method. This method needs on average four calls of oracle per iteration. Both these methods are based on DR-condition (1.5).

In Sect. 4, in order to derive Universal Fast Gradient Method, we introduce condition (1.6). We show that this scheme is uniformly optimal for minimizing composite function, which has Hölder-continuous gradients of its smooth part. This scheme has a reliable stopping criterion. It needs on average four calls of oracle per iteration.

In Sect. 5, we present preliminary computational results. We consider three families of random test problems. All of them are nonsmooth problems with Lipschitz-continuous objective function. It is shown that quite often the Universal Fast Gradient Method is able to accelerate and demonstrates the rate of convergence typical for smooth minimization schemes. The choice of appropriate norms is always very important.

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

$$\|s\| \stackrel{\text{def}}{=} \max_{x \in E}\{\langle s, x \rangle : \ \|x\| \leq 1\}.$$

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

## 2 Universal primal gradient method

Consider the following minimization problem:

$$\min_{x \in Q} \left[ \tilde{f}(x) \stackrel{\mathrm{def}}{=} f(x) + \Psi(x) \right], \tag{2.1}$$

where $Q$ is a simple closed convex set, $\Psi$ is a simple closed convex function. Function $f$ is assumed to be subdifferentiable on $Q$. In order to characterize variability of its (sub)gradients, we introduce the following values:

$$M_\nu \equiv M_\nu(f) = \sup_{\substack{x,y \in Q,\\ x \neq y}} \frac{\|\nabla f(x) - \nabla f(y)\|_*}{\|x - y\|^\nu}, \quad \nu \geq 0. \tag{2.2}$$

It is clear that the value $M_\nu$ depends on the smoothness parameter $\nu$ in a regular way. For example, if the set $Q$ is bounded, then for any $\nu' \in [\nu, 1]$ we have

$$M_\nu \leq M_{\nu'} \,(\mathrm{diam}\, Q)^{\nu' - \nu}.$$

On the other hand, definition (2.2) can be written in the following form:

$$\ln M_\nu = \sup_{\substack{x,y \in Q,\\ x \neq y}} [\ln \|\nabla f(x) - \nabla f(y)\|_* - \nu \ln \|x - y\|].$$

Thus, $M_\nu$ is a log-convex function of $\nu$. For certain $\nu \in [0, 1]$, the constant $M_\nu$ can be infinite. However, if $M_0$ and $M_1$ are finite, then

$$M_\nu \leq M_0^{1-\nu} M_1^\nu, \quad 0 \leq \nu \leq 1. \tag{2.3}$$

In any case, if $M_\nu < \infty$, then

$$\|\nabla f(x) - \nabla f(y)\|_* \leq M_\nu \|x - y\|^\nu, \quad x, y \in Q. \tag{2.4}$$

This inequality ensures that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M_\nu}{1+\nu} \|x - y\|^{1+\nu}, \quad x, y \in Q. \tag{2.5}$$

Our main assumption is as follows.

**Assumption 1** $\hat{M}(f) \stackrel{\mathrm{def}}{=} \inf\limits_{0 \leq \nu \leq 1} M_\nu(f) < +\infty.$

The majority of functions with Hölder continuous gradient can be obtained by duality relations. Recall the following definition.

**Definition 1** Let $f \in C^1(Q)$. It is called *uniformly convex* on $Q$ of degree $p \geq 2$ if

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \sigma_p \| y - x \|^p, \quad x, y \in Q,$$

where $\sigma_p = \sigma_p(f)$ is the parameter of uniform convexity.

Adding such $f$ to a convex function does not change the parameter. If $p = 2$, then this is just a definition of strong convexity.

For $f(x) \in C^1$ define its *conditional Fenchel dual*

$$f_*(s) = \sup_{x \in Q} [\langle s, x \rangle - f(x)].$$

Note that $\nabla f_*(s) = x_f(s) = \arg\max_{x \in Q} [\langle s, x \rangle - f(x)]$.

**Lemma 1** *If $f$ is $p$-uniformly convex on $Q$, then $f_* \in C^{1,\nu}$ with parameters*

$$\nu = \frac{1}{p-1}, \quad M_\nu(f_*) = \left( \frac{1}{\sigma_p} \right)^{\frac{1}{p-1}}.$$

*Proof* For points $s_1$ and $s_2$, denote $x_i = x_f(s_i)$, $i = 1, 2$. Then

$$\langle s_i - \nabla f(x_i), x_{3-i} - x_i \rangle \leq 0, \quad i = 1, 2.$$

Adding these inequalities, we get

$$\langle s_1 - s_2, x_1 - x_2 \rangle \geq \langle \nabla f(x_1) - \nabla f(x_2), x_1 - x_2 \rangle \geq \sigma_p \| x_1 - x_2 \|^p.$$

□

Note that for bounded $Q$ we have $M_0(f^*) \leq \operatorname{diam} Q$. In this case, under conditions of Lemma 1, we have $f_* \in C^{1,0} \bigcap C^{1,\nu}$.

Let us give an example of a practical application (see [1]).

*Example 1* Consider a gas transmission network, consisting of $m$ pipe lines connecting $n$ nodes. Denote by $x^{(i)}$ the gas flow for $i$th pipe, $i = 1, \ldots, m$. Then the conservation laws at the nodes can be written in the form of linear equality constraints $Ax = d$, where $d \in \mathbb{R}^m$ is the vector of demand flows at the nodes. The actual flows in the pipes are defined by the pressures established at the sources and sinks. In accordance to the minimal energy principle, in order to compute them, we need to solve the following minimization problem:

$$\min_x \left\{ f(x) \stackrel{\text{def}}{=} \sum_{i=1}^m \alpha_i |x^{(i)}|^3 : \; Ax = d \right\},$$

where the parameter $\alpha_i > 0$, $i = 1, \dots, m$, depends on the diameter of the arc, its length, etc. It is easy to check that the objective function of the dual problem $\min_{y \in \mathbb{R}^m} \phi(y)$ with

$$\phi(y) \stackrel{\rm def}{=} \max_x [\langle y, d - Ax \rangle - f(x)]$$

belongs to $C^{1,\nu}(\mathbb{R}^n)$ with $\nu = \frac{1}{2}$. In this situation, the most natural choice of norm for the primal space is $\|x\|_{3,\alpha} \stackrel{\rm def}{=} \left[\sum_{i=1}^m \alpha_i |x^{(i)}|^3\right]^{1/3}$, $x \in \mathbb{R}^m$.

For solving the problem (2.1), we introduce a *prox-function* $d(x)$. This is a continuously differentiable strongly convex function with convexity parameter equal to one:

$$d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2}\|x - y\|^2, \quad x, y \in \mathrm{rint}\, Q. \tag{2.6}$$

We assume that $d(x)$ attains its minimum on $Q$ at some point $x_0$, and $d(x_0) = 0$. Thus,

$$d(x) \stackrel{(2.6)}{\geq} \frac{1}{2}\|x - x_0\|^2, \quad x \in Q. \tag{2.7}$$

This prox-function defines the *Bregman distance* $\xi(x, y) \stackrel{\rm def}{=} d(y) - d(x) - \langle \nabla d(x), y - x \rangle$. Clearly, $\xi(x, x) \equiv 0$, and

$$\xi(x, y) \stackrel{(2.6)}{\geq} \frac{1}{2}\|x - y\|^2, \quad x, y \in Q. \tag{2.8}$$

Now for any $x \in Q$ we can define the *Bregman mapping*

$$\mathcal{B}_M(x) = \arg\min_{y \in Q} \left\{ \psi_M(x, y) \stackrel{\rm def}{=} f(x) + \langle \nabla f(x), y - x \rangle + M\xi(x, y) + \Psi(y) \right\}. \tag{2.9}$$

We assume that this point is easily computable either in a closed form, or by some cheap computational procedure. The first-order optimality condition for optimization problem in (2.9) is as follows:

$$\langle \nabla f(x) + M(\nabla d(\mathcal{B}_M(x)) - \nabla d(x)) + \nabla \Psi(\mathcal{B}_M(x)), y - \mathcal{B}_M(x) \rangle \geq 0, \quad y \in Q. \tag{2.10}$$

Denote $\psi^*_M(x) = \psi_M(x, \mathcal{B}_M(x))$.

**Lemma 2** *Let function $f$ satisfy condition (2.4). Then for any $\delta > 0$ and*

$$M \geq \left[\frac{1 - \nu}{1 + \nu} \cdot \frac{1}{\delta}\right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \tag{2.11}$$

*we have*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} M \|y - x\|^2 + \frac{\delta}{2}, \quad x, y \in Q. \tag{2.12}$$

*Therefore,*

$$\tilde{f}(\mathcal{B}_M(x)) \leq \psi_M^*(x) + \frac{\delta}{2}. \tag{2.13}$$

*Proof* Recall that all nonnegative $\tau$ and $s$ satisfy the Young's inequality

$$\frac{1}{p}\tau^p + \frac{1}{q}s^q \geq \tau s,$$

where $p, q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. Therefore, taking $p = \frac{2}{1+\nu}$, $q = \frac{2}{1-\nu}$, and $\tau = t^{1+\nu}$, we get

$$t^{1+\nu} \leq \frac{1+\nu}{2s} t^2 + \frac{1-\nu}{2} s^{\frac{1+\nu}{1-\nu}}, \quad s > 0,\ t \geq 0. \tag{2.14}$$

Let us choose $s = \left[\frac{1+\nu}{1-\nu} \cdot \frac{\delta}{M_\nu}\right]^{\frac{1-\nu}{1+\nu}}$. Then $\frac{1-\nu}{1+\nu} M_\nu s^{\frac{1+\nu}{1-\nu}} = \delta$. Therefore,

$$\begin{aligned} \frac{M_\nu}{1+\nu} t^{1+\nu} &\overset{(2.14)}{\leq} \frac{1}{2s} M_\nu t^2 + \frac{\delta}{2} \\ &= \frac{1}{2}\left[\frac{1-\nu}{1+\nu} \cdot \frac{1}{\delta}\right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} t^2 + \frac{\delta}{2} \overset{(2.11)}{\leq} \frac{1}{2} M t^2 + \frac{\delta}{2}. \end{aligned} \tag{2.15}$$

This inequality, together with (2.5), justifies (2.12). Further, denoting $x_+ = \mathcal{B}_M(x)$, we obtain:

$$\begin{aligned} f(x_+) &\overset{(2.5)}{\leq} f(x) + \langle \nabla f(x), x_+ - x \rangle + \frac{M_\nu}{1+\nu} \|x_+ - x\|^{1+\nu} \\ &\overset{(2.15)}{\leq} f(x) + \langle \nabla f(x), x_+ - x \rangle + \frac{M}{2} \|x_+ - x\|^2 + \frac{\delta}{2} \\ &\overset{(2.8)}{\leq} f(x) + \langle \nabla f(x), x_+ - x \rangle + M\xi(x, x_+) + \frac{\delta}{2}. \end{aligned}$$

Therefore, $\tilde{f}(x_+) = f(x_+) + \Psi(x_+) \leq \psi_M^*(x) + \frac{\delta}{2}$. □

Note that the right-hand side of inequality (2.11) is continuous in $\nu$. As $\nu \to 1$, it becomes

$$M \geq M_1. \tag{2.16}$$

*Remark 1* Inequality (2.12) was already used by several authors (see, for example, Sect. 2.3(c) in [3]).

Let us look now at the simplest Universal Primal Gradient Method equipped with a backtracking line search procedure with restore. Denote by $x^*$ the optimal solution to (2.1).

**Universal Primal Gradient Method (PGM)**

**Initialization.** Choose $L_0 > 0$ and accuracy $\epsilon > 0$.

**For** $k \geq 0$ **do:**

1. Find the smallest $i_k \geq 0$ such that for $x_k^+ \stackrel{\mathrm{def}}{=} \mathcal{B}_{2^{i_k} L_k}(x_k)$ we have

$$f(x_k^+) \leq f(x_k) + \langle \nabla f(x_k), x_k^+ - x_k \rangle + 2^{i_k - 1} L_k \| x_k^+ - x_k \|^2 + \tfrac{1}{2}\epsilon. \tag{2.17}$$

2. Set $x_{k+1} = \mathcal{B}_{2^{i_k} L_k}(x_k)$, and $L_{k+1} = 2^{i_k - 1} L_k$.

Denote $\gamma(M, \epsilon) \stackrel{\mathrm{def}}{=} \left[ \frac{1-\nu}{1+\nu} \cdot \frac{1}{\epsilon} \right]^{\frac{1-\nu}{1+\nu}} M^{\frac{2}{1+\nu}}$, $S_k = \sum_{i=1}^{k+1} \frac{1}{L_i}$, and $\tilde{f}_k^* = \frac{1}{S_k} \sum_{i=0}^{k} \frac{1}{L_{i+1}} \tilde{f}(x_{i+1})$.

**Theorem 1** *Let $f$ satisfy condition (2.4). Assume that $L_0 \leq \gamma(M_\nu, \epsilon)$. Then for all $k \geq 0$ we have $L_{k+1} \leq \gamma(M_\nu, \epsilon)$. Moreover, for all $y \in Q$*

$$\tilde{f}_k^* \leq \frac{1}{S_k} \sum_{i=0}^{k} \frac{1}{L_{i+1}} [f(x_i) + \langle \nabla f(x_i), y - x_i \rangle] + \Psi(y) + \frac{\epsilon}{2} + \frac{2}{S_k} \xi(x_0, y). \tag{2.18}$$

*Therefore,* $\tilde{f}_k^* - \tilde{f}(x^*) \leq \frac{\epsilon}{2} + \frac{2\gamma(M_\nu, \epsilon)}{k+1} \xi(x_0, x^*)$.

*Proof* In view of Lemma 2, the line-search procedure of Step 1 in method (2.17) is well defined, and

$$2L_{k+1} = 2^{i_k} L_k \leq 2\gamma(M_\nu, \epsilon). \tag{2.19}$$

Let us fix an arbitrary point $y \in Q$. Denote $r_k(y) \stackrel{\mathrm{def}}{=} \xi(x_k, y)$. Then

$$\begin{aligned} r_{k+1}(y) &= d(y) - d(x_{k+1}) - \langle \nabla d(x_{k+1}), y - x_{k+1} \rangle \\ &\stackrel{(2.10)}{\leq} d(y) - d(x_{k+1}) - \langle \nabla d(x_k), y - x_{k+1} \rangle + \frac{1}{2L_{k+1}} \langle \nabla f(x_k) \\ &\quad + \nabla \Psi(x_{k+1}), y - x_{k+1} \rangle. \end{aligned}$$

Note that

$$
\begin{aligned}
& d(y) - d(x_{k+1}) - \langle \nabla d(x_k), y - x_{k+1} \rangle \\
& \overset{(2.6)}{\leq} d(y) - d(x_k) - \langle \nabla d(x_k), x_{k+1} - x_k \rangle - \frac{1}{2}\|x_{k+1} - x_k\|^2 - \langle \nabla d(x_k), y - x_{k+1} \rangle \\
& = r_k(y) - \frac{1}{2}\|x_{k+1} - x_k\|^2.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
r_{k+1}(y) - r_k(y) \; & \leq \; \frac{1}{2L_{k+1}} \langle \nabla f(x_k) + \nabla \Psi(x_{k+1}), y - x_{k+1} \rangle - \frac{1}{2}\|x_{k+1} - x_k\|^2 \\
& = \frac{1}{2L_{k+1}} \langle \nabla \Psi(x_{k+1}), y - x_{k+1} \rangle \\
& \quad - \frac{1}{2L_{k+1}} \left( \langle \nabla f(x_k), x_{k+1} - x_k \rangle + L_{k+1} \|x_{k+1} - x_k\|^2 \right) \\
& \quad + \frac{1}{2L_{k+1}} \langle \nabla f(x_k), y - x_k \rangle \\
& \leq \frac{1}{2L_{k+1}} \left( \Psi(y) - \Psi(x_{k+1}) + f(x_k) - f(x_{k+1}) + \frac{1}{2}\epsilon + \langle \nabla f(x_k), y - x_k \rangle \right).
\end{aligned}
$$

Thus, we obtain the following inequality:

$$
\begin{aligned}
& \frac{1}{2L_{k+1}} \tilde{f}(x_{k+1}) + r_{k+1}(y) \\
& \quad \leq \frac{1}{2L_{k+1}} \left( f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \Psi(y) + \frac{\epsilon}{2} \right) + r_k(y).
\end{aligned}
$$

Summing up these inequalities, we obtain

$$
\tilde{f}_k^* \leq \frac{1}{S_k} \sum_{i=0}^{k} \frac{1}{L_{i+1}} \left[ f(x_i) + \langle \nabla f(x_i), y - x_i \rangle \right] + \Psi(y) + \frac{\epsilon}{2} + \frac{2}{S_k} r_0(y).
$$

It remains to use inequality (2.19). □

It is important that method (2.17) does not include $\nu$ as a parameter. Therefore, in view of Theorem 1, in order to get an $\epsilon$-solution of problem (2.1) we need

$$
4\xi(x_0, x^*) \inf_{0 \leq \nu \leq 1} \left( \frac{M_\nu}{\epsilon} \right)^{\frac{2}{1+\nu}} \tag{2.20}
$$

iterations at most. In this estimate, among all classes of functions with Hölder continuous gradient, we choose the class which better fits our particular objective function. Note that the expression (2.20) is log-quasiconvex in $\nu$. Hence, if $M_0$ and $M_1$ are finite, there are good chances that the optimal $\nu$ belongs to the interior of the interval (0, 1).

Inequality (2.18) gives us a reliable stopping criterion for method (2.17). Indeed, assume we have a bound for the size of optimal solution:

$$\xi(x_0, x^*) \leq D. \tag{2.21}$$

Denote $\ell_k^p(y) \stackrel{\text{def}}{=} \frac{1}{S_k} \sum_{i=0}^k \frac{1}{L_{i+1}} [f(x_i) + \langle \nabla f(x_i), y - x_i \rangle]$, and define

$$\hat{f}_k = \min_{y \in Q} \left\{ \ell_k^p(y) + \Psi(y) : \ \xi(x_0, y) \leq D \right\}.$$

Then

$$\tilde{f}_k^* - \tilde{f}(x^*) \leq \tilde{f}_k^* - \hat{f}_k \leq \frac{2\gamma(M_\nu, \epsilon)}{k+1} D. \tag{2.22}$$

Note that $\hat{f}_k$ can be computed. Thus, inequality (2.22) provides us with an implementable stopping criterion $\tilde{f}_k^* - \hat{f}_k \leq \epsilon$.

Finally, let us estimate $N(k)$, the total number of computations of the function value in method (2.17) after $k$ iterations. Note that

$$L_{k+1} = \frac{1}{2} 2^{i_k} L_k.$$

Therefore, $i_k - 1 = \log_2 \frac{L_{k+1}}{L_k}$. Hence, for any $\nu \in [0, 1]$, we have

$$\begin{aligned} N(k) &= \sum_{j=0}^{k} (i_j + 1) = 2(k+1) + \log_2 L_{k+1} - \log_2 L_0 \\ &\stackrel{(2.19)}{\leq} 2(k+1) + \frac{1-\nu}{1+\nu} \log_2 \left( \frac{1-\nu}{1+\nu} \cdot \frac{1}{\epsilon} \right) + \frac{2}{1+\nu} \log_2 M_\nu - \log_2 L_0. \end{aligned}$$

Finally, we come to the following upper bound:

$$N(k) \leq 2(k+1) - \log_2 L_0 + \inf_{0 \leq \nu \leq 1} \left[ \frac{1-\nu}{1+\nu} \log_2 \left( \frac{1-\nu}{1+\nu} \cdot \frac{1}{\epsilon} \right) + \frac{2}{1+\nu} \log_2 M_\nu \right]. \tag{2.23}$$

Thus on average, up to negligible logarithmic terms, method (2.17) requires two computations of function values per iteration.

The complexity estimates in (2.20) are optimal only for $\nu = 0$. In Sect. 4 we show that much better (and optimal) bounds can be achieved by a fast gradient scheme.

## 3 Universal dual gradient method

Dual gradient method is based on updating a simple model for objective function of problem (2.1). Its justification is based on the following simple result.

**Lemma 3** *Let $\phi : Q \to R \bigcup \{+\infty\}$ be a convex function such that for some $M \geq 0$ the difference $\phi(x) - Md(x)$ is closed and convex on $Q$. Denote $\bar{x} = \arg\min_{x \in Q} \phi(x)$. Then*

$$\phi(y) \geq \phi(\bar{x}) + M\xi(\bar{x}, y), \quad y \in Q. \tag{3.1}$$

*Proof* Denote $F(y) = \phi(y) - Md(y)$. Let us choose an arbitrary $y \in Q$ and $\alpha \in (0, 1]$. Then

$$\begin{aligned} F(\bar{x}) + Md(\bar{x}) &= \phi(\bar{x}) \leq \phi(\bar{x} + \alpha(y - \bar{x})) \\ &\leq (1-\alpha)F(\bar{x}) + \alpha F(y) + Md(\bar{x} + \alpha(\bar{x} - y)) \\ &\leq (1-\alpha)F(\bar{x}) + \alpha F(y) + M[d(\bar{x}) + \alpha\langle \nabla d(\bar{x} + \alpha(y - \bar{x})), y - \bar{x}\rangle]. \end{aligned}$$

Thus, $F(y) \geq F(\bar{x}) - M\langle \nabla d(\bar{x} + \alpha(y - \bar{x})), y - \bar{x}\rangle$. Since we can choose $\alpha$ arbitrary small, from continuous differentiability of function $d(\cdot)$ we get $F(y) \geq F(\bar{x}) - M\langle \nabla d(\bar{x}), y - \bar{x}\rangle$. And this is exactly inequality (3.1). □

**Universal Dual Gradient Method (DGM)**

**Initialization.** Choose $L_0 > 0$. Define $\phi_0(x) = \xi(x_0, x)$.

**For** $k \geq 0$ **do:**

1. Find the smallest $i_k \geq 0$ such that for the point

$$x_{k,i_k} = \arg\min_{x \in Q} \left\{ \phi_k(x) + \frac{1}{2^{i_k} L_k}[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \Psi(x)] \right\}$$

we have $\tilde{f}\left(\mathcal{B}_{2^{i_k} L_k}(x_{k,i_k})\right) \leq \psi^*_{2^{i_k} L_k}(x_{k,i_k}) + \frac{\epsilon}{2}$.

2. Set $x_{k+1} = x_{k,i_k}$, $y_k = \mathcal{B}_{2^{i_k} L_k}(x_{k,i_k})$, $L_{k+1} = 2^{i_k - 1} L_k$, and

$$\phi_{k+1}(x) = \phi_k(x) + \frac{1}{2L_{k+1}}[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \Psi(x)].$$

(3.2)

Assume that $L_0 \leq \gamma(M_\nu, \epsilon)$, and $M_\nu < \infty$. Note that the termination criterion of Step 1 in method (3.2) is satisfied if $2^{i_k} L_k \geq \gamma(M_\nu, \epsilon)$. Therefore, same as in the proof of Theorem 1, we have

$$L_k \leq \gamma(M_\nu, \epsilon), \quad k \geq 1. \tag{3.3}$$

Denote $S_k = \sum_{i=0}^{k} \frac{1}{L_{i+1}}$, and $\phi^*_k = \min_{y \in Q} \phi_k(y)$.

**Lemma 4** *For any $k \geq 0$ we have*

$$\sum_{i=0}^{k} \frac{1}{2L_{i+1}} \tilde{f}(y_i) \leq \phi_{k+1}^* + S_k \cdot \frac{\epsilon}{4}. \tag{3.4}$$

*Proof* Let us prove (3.4) by induction. For $k = 0$ we have

$$\begin{aligned}\frac{1}{2L_1}\tilde{f}(y_0) - S_0 \cdot \frac{\epsilon}{4} = \frac{1}{2L_1}\left[\tilde{f}(y_0) - \frac{\epsilon}{2}\right] &\leq \frac{1}{2^{i_0}L_0}\psi_{2^{i_0}L_0}^*(y_0) \\ &= \frac{1}{2^{i_0}L_0}[f(x_0) + \langle \nabla f(x_0), y_0 - x_0\rangle + \Psi(y_0)] + \xi(x_0, y_0) \\ &= \phi_1(y_0) = \min_{x\in Q} \phi_1(x) = \phi_1^*.\end{aligned}$$

Assume that (3.4) is true for some $k \geq 0$. In view of Lemma 3, for any $k \geq 0$ we have

$$\phi_k(x) \geq \phi_k(x_k) + \xi(x_k, x), \quad x \in Q.$$

Therefore,

$$\begin{aligned}&\min_{x\in Q} \phi_{k+2}(x) \\ &\quad = \min_{x\in Q}\left\{\phi_{k+1}(x) + \tfrac{1}{2L_{k+2}}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \Psi(x)]\right\} \\ &\quad \geq \min_{x\in Q}\left\{\phi_{k+1}(x_{k+1}) + \xi(x_{k+1}, x) + \tfrac{1}{2L_{k+2}}[f(x_{k+1})\right. \\ &\qquad \left. + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \Psi(x)]\right\} \\ &\quad \geq \phi_{k+1}(x_{k+1}) + \tfrac{1}{2L_{k+2}}\left[\tilde{f}(y_{k+1}) - \tfrac{\epsilon}{2}\right] \overset{(3.4)}{\geq} -S_{k+1}\cdot\tfrac{\epsilon}{4} + \sum_{i=0}^{k+1}\tfrac{1}{2L_{i+1}}\tilde{f}(y_i).\end{aligned}$$

Thus, we have proved that (3.4) is valid for all $k \geq 0$. □

Now we can prove the main convergence result for Universal Dual Gradient Method. Denote $\tilde{f}_k^* = \frac{1}{S_k}\sum_{i=0}^{k}\frac{1}{L_{i+1}}\tilde{f}(y_i)$.

**Theorem 2** *Let $f$ satisfies condition (2.4) with $M_\nu < \infty$, and $L_0 \leq \gamma(M_\nu, \epsilon)$. Then all $L_k$ generated by method (3.2) satisfy condition (3.3). Moreover, for all $k \geq 0$ we have*

$$\tilde{f}_k^* - \tilde{f}(x^*) \leq \frac{\epsilon}{2} + \frac{2\gamma(M_\nu, \epsilon)}{k+1}\xi(x_0, x^*). \tag{3.5}$$

*Proof* Indeed, in view of inequality (3.4), we have

$$\frac{1}{2} S_k \tilde{f}_k^* = \sum_{i=0}^{k} \frac{1}{2L_{i+1}} \tilde{f}(y_i) \le \min_{x \in Q} \phi_{k+1}(x) + S_k \cdot \frac{\epsilon}{4} \ \le \ \frac{1}{2} S_k \tilde{f}(x^*)$$
$$+ \xi(x_0, x^*) + S_k \cdot \frac{\epsilon}{4}.$$

It remains to use inequality (3.3). □

Note that the worst-case complexity bound for the number of iterations of method (3.2) coincides with the bound (2.20). However, the number of function evaluations at each iteration of (3.2) is twice more than (2.23).

Same as method (2.17), Universal Dual Gradient Method can be equipped with a stopping criterion. Denote $\ell_k^d(y) = \sum_{i=0}^{k} \frac{1}{L_{i+1}} [f(x_i) + \langle \nabla f(x_i), y - x_i \rangle]$. Assume that $\xi(x_0, x^*) \le D$ and the constant $D$ is known. Denote

$$\hat{f}_k = \min_{y \in Q} \left\{ \frac{1}{S_k} \ell_k^d(y) + \Psi(y) : \ \xi(x_0, y) \le D \right\}.$$

Note that

$$\begin{aligned} \hat{f}_k &= \min_{x \in Q} \max_{\beta \ge 0} \left\{ \frac{1}{S_k} \ell_k^d(y) + \Psi(y) + \beta\,(\xi(x_0, y) - D) \right\} \\ &= \max_{\beta \ge 0} \min_{x \in Q} \left\{ \frac{1}{S_k} \ell_k^d(y) + \Psi(y) + \beta\,(\xi(x_0, y) - D) \right\} \\ &\overset{\beta = 2/S_k}{\ge} \frac{2}{S_k} \phi_{k+1}^* - \frac{2}{S_k} D. \end{aligned}$$

Since $\tilde{f}_k^* \overset{(3.4)}{\le} \frac{2}{S_k} \phi_{k+1}^* + \frac{\epsilon}{2}$, we conclude that the stoping criterion $\tilde{f}_k^* - \hat{f}_k \le \epsilon$ ensures $\tilde{f}_k^* - \tilde{f}(x^*) \le \epsilon$ as far as $S_k \ge \frac{4}{\epsilon} D$.

## 4 Universal fast gradient method

Finally, let us apply to problem (2.1) the following method.

**Universal Fast Gradient Method (FGM)**

**Initialization.** Choose $L_0 > 0$. Define $\phi_0(x) = \xi(x_0, x)$, $y_0 = x_0$, $A_0 = 0$.

**For** $k \geq 0$ **do:**

1. Find $v_k = \arg\min\limits_{x \in Q} \phi_k(x)$.

2. Find the smallest $i_k \geq 0$ such that coefficient $a_{k+1,i_k} > 0$, computed from

equation $\boxed{a^2_{k+1,i_k} = \frac{1}{2^{i_k} L_k}(A_k + a_{k+1,i_k})}$ and used in the definitions

$A_{k+1,i_k} = A_k + a_{k+1,i_k}$, $\tau_{k,i_k} = \frac{a_{k+1,i_k}}{A_{k+1,i_k}}$, $x_{k+1,i_k} = \tau_{k,i_k} v_k + (1 - \tau_{k,i_k}) y_k$,

$\hat{x}_{k+1,i_k} = \arg\min\limits_{y \in Q} \left\{ \xi(v_k, y) + a_{k+1,i_k} [\langle \nabla f(x_{k+1,i_k}), y \rangle + \Psi(y)] \right\}$,

$y_{k+1,i_k} = \tau_{k,i_k} \hat{x}_{k+1,i_k} + (1 - \tau_{k,i_k}) y_k$, ensures the following relation:

$$f(y_{k+1,i_k}) \leq f(x_{k+1,i_k}) + \langle \nabla f(x_{k+1,i_k}), y_{k+1,i_k} - x_{k+1,i_k} \rangle$$
$$+ 2^{i_k - 1} L_k \| y_{k+1,i_k} - x_{k+1,i_k} \|^2 + \frac{\epsilon}{2} \tau_{k,i_k}.$$

3. Set $x_{k+1} = x_{k+1,i_k}$, $y_{k+1} = y_{k+1,i_k}$, $a_{k+1} = a_{k+1,i_k}$, $\tau_k = \tau_{k,i_k}$.

Define $A_{k+1} = A_k + a_{k+1}$, $L_{k+1} = 2^{i_k - 1} L_k$, and

$\phi_{k+1}(x) = \phi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle + \Psi(x)]$.

(4.1)

**Theorem 3** *Let $f$ satisfies condition (2.4) with certain $M_\nu < +\infty$. Then all iterations of method (4.1) are well defined. Moreover, for all $k \geq 0$ we have*

$$A_k \left( \tilde{f}(y_k) - \frac{\epsilon}{2} \right) \leq \phi_k^* \stackrel{\mathrm{def}}{=} \min_{x \in Q} \phi_k(x), \tag{4.2}$$

*where $A_k \geq \left[ \frac{1}{2^{2+4\nu} M_\nu^2} \epsilon^{1-\nu} k^{1+3\nu} \right]^{\frac{1}{1+\nu}}$. Consequently, for all $k \geq 1$ we have*

$$\tilde{f}(y_k) - \tilde{f}(x^*) \leq \left[ \frac{2^{2+4\nu} M_\nu^2}{\epsilon^{1-\nu} k^{1+3\nu}} \right]^{\frac{1}{1+\nu}} \xi(x_0, x^*) + \frac{\epsilon}{2}. \tag{4.3}$$

*Proof* Let us prove first, that the "line-search" process of Item 2 in (4.1) is finite. In view of inequality (2.12), we need to show that

$$2^{i_k} L_k \geq \left[\frac{1}{\epsilon \tau_{k,i_k}}\right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}$$

for $i_k$ large enough. Indeed, in view of the characteristic equation for $a_{k+1,i_k}$, we have

$$2^{i_k} L_k \tau_{k,i_k}^{\frac{1-\nu}{1+\nu}} = \frac{A_{k+1,i_k}}{a_{k+1,i_k}^2} \cdot \left(\frac{a_{k+1,i_k}}{A_{k+1,i_k}}\right)^{\frac{1-\nu}{1+\nu}} = \left(\frac{1}{\tau_{k,i_k}}\right)^{\frac{2\nu}{1+\nu}} \cdot \frac{1}{a_{k+1,i_k}} \geq \frac{1}{a_{k+1,i_k}}.$$

It remains to note that $a_{k+1,i_k} \to 0$ as $i_k \to \infty$.

Let us prove relation (4.2). For $k = 0$ it is evident. Assume that it is valid for certain $k \geq 0$. Then for any $y \in Q$ we have

$$\begin{aligned} \phi_k(y) &\overset{(3.1)}{\geq} \phi_k^* + \xi(v_k, y) \overset{(4.2)}{\geq} A_k \left(\tilde{f}(y_k) - \frac{\epsilon}{2}\right) + \xi(v_k, y) \\ &\geq A_k \left(f(x_{k+1}) + \langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle + \Psi(y_k) - \frac{\epsilon}{2}\right) + \xi(v_k, y). \end{aligned}$$

Therefore,

$$\begin{aligned} \phi_{k+1}(y) &\geq \xi(v_k, y) + A_k \left(f(x_{k+1}) + \langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle + \Psi(y_k) - \frac{\epsilon}{2}\right) \\ &\quad + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), y - x_{k+1}\rangle + \Psi(y)] \\ &= \xi(v_k, y) + A_k \left(f(x_{k+1}) + \Psi(y_k) - \frac{\epsilon}{2}\right) \\ &\quad + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), y - v_k\rangle + \Psi(y)]. \end{aligned}$$

In view of definition of point $\hat{x}_{k+1}$, we have

$$\begin{aligned} \phi_{k+1}^* &\geq \xi(v_k, \hat{x}_{k+1}) + A_k \left(f(x_{k+1}) + \Psi(y_k) - \frac{\epsilon}{2}\right) \\ &\quad + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), \hat{x}_{k+1} - v_k\rangle + \Psi(\hat{x}_{k+1})] \\ &\overset{(2.8)}{\geq} \frac{1}{2}\|\hat{x}_{k+1} - v_k\|^2 + A_{k+1} f(x_{k+1}) + A_{k+1}\Psi(y_{k+1}) - \frac{\epsilon}{2} A_k \\ &\quad + a_{k+1}\langle \nabla f(x_{k+1}), \hat{x}_{k+1} - v_k\rangle. \end{aligned}$$

Since $\hat{x}_{k+1} - v_k = \frac{1}{\tau_k}(y_{k+1} - x_{k+1})$, we obtain

$$\begin{aligned} \phi_{k+1}^* &\geq \frac{1}{2\tau_k^2}\|y_{k+1} - x_{k+1}\|^2 + A_{k+1} f(x_{k+1}) + A_{k+1}\Psi(y_{k+1}) - \frac{\epsilon}{2} A_k \\ &\quad + A_{k+1}\langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle \\ &= A_{k+1}(f(x_{k+1}) + \langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle \\ &\quad + 2^{i_k - 1} L_k \|y_{k+1} - x_{k+1}\|^2 + \Psi(y_{k+1})) - \frac{\epsilon}{2} A_k \end{aligned}$$

$$\geq A_{k+1}(f(y_{k+1}) - \frac{\epsilon}{2}\tau_k + \Psi(y_{k+1})) - \frac{\epsilon}{2}A_k \;=\; A_{k+1}\left(\tilde{f}(y_{k+1}) - \frac{\epsilon}{2}\right).$$

Thus, inequality (4.2) is proved for all $k \geq 0$. Since $\phi_k(y) \leq A_k\tilde{f}(y) + \xi(x_0, y)$ for all $y \in Q$, we obtain

$$\tilde{f}(y_k) - \tilde{f}(x^*) \overset{(4.2)}{\leq} \frac{\xi(x_0, x^*)}{A_k} + \frac{\epsilon}{2}, \quad k \geq 1.$$

It remains to estimate the growth of coefficients $A_k$.

In view of Lemma 2, the number of internal steps $i_k$ in Item 2 of (4.1) satisfies inequality

$$2^{i_k} L_k \leq 2\left[\frac{1}{\epsilon\tau_k}\right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}.$$

Therefore, $\frac{a_{k+1}^2}{A_{k+1}} = \frac{1}{2^{i_k}L_k} \;\geq\; \frac{1}{2M_\nu^{\frac{2}{1+\nu}}}[\epsilon\tau_k]^{\frac{1-\nu}{1+\nu}}$, which is $a_{k+1}^2 \geq \frac{[\epsilon a_{k+1}]^{\frac{1-\nu}{1+\nu}}}{2M_\nu^{\frac{2}{1+\nu}}} A_{k+1}^{\frac{2\nu}{1+\nu}}$. Thus, we come to the following estimate:

$$a_{k+1} \geq \frac{\epsilon^{\frac{1-\nu}{1+3\nu}} A_{k+1}^{\frac{2\nu}{1+3\nu}}}{2^{\frac{1+\nu}{1+3\nu}} M_\nu^{\frac{2}{1+3\nu}}}. \tag{4.4}$$

Denote $B_k = A_k^\gamma$, where $\gamma = \frac{1+\nu}{1+3\nu} \geq \frac{1}{2}$. Since $A_{k+1} \geq A_k$, we have

$$B_{k+1} - B_k \geq \frac{A_{k+1} - A_k}{A_{k+1}^{1-\gamma} + A_k^{1-\gamma}} \;\geq\; \frac{a_{k+1}}{2A_{k+1}^{1-\gamma}} \;\overset{(4.4)}{\geq}\; \frac{\epsilon^{\frac{1-\nu}{1+3\nu}}}{2^{\frac{2+4\nu}{1+3\nu}} M_\nu^{\frac{2}{1+3\nu}}}.$$

Thus, we have proved that $A_k \geq \left[\frac{k\cdot\epsilon^{\frac{1-\nu}{1+3\nu}}}{2^{\frac{2+4\nu}{1+3\nu}} M_\nu^{\frac{2}{1+3\nu}}}\right]^{\frac{1+3\nu}{1+\nu}} \;=\; \frac{k^{\frac{1+3\nu}{1+\nu}}\epsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2+4\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}$. □

From the rate of convergence (4.3), we get the following upper bound for the number of iterations, which are necessary for getting $\epsilon$-solution of problem (2.1):

$$k \leq \inf_{0\leq\nu\leq 1}\left[\left(\frac{2^{\frac{3+5\nu}{2}} M_\nu}{\epsilon}\right)^{\frac{2}{1+3\nu}} \xi(x_0, x^*)^{\frac{1+\nu}{1+3\nu}}\right]. \tag{4.5}$$

As compared with (2.20), the dependence of this bound in smoothness parameters is now optimal (see [8]).

Same as the gradient methods (2.17) and (3.2), Fast Gradient Method (4.1) can be equipped with an implementable stopping criterion. Assume that $\xi(x_0, x^*) \leq D$. Denote $\ell_k^{pd}(y) = \sum_{i=1}^k a_i[f(x_i) + \langle\nabla f(x_i), x - x_i\rangle]$, and $\hat{f}_k = \min_{y\in Q}\{\frac{1}{A_k}\ell_k^{pd}(y) +$

$\Psi(y):\ \xi(x_0,y)\le D\}$. Note that $\tilde f(y_k) \overset{(4.2)}{\le} \frac{\epsilon}{2}+\frac{1}{A_k}\phi_k^*$. Using the reasoning presented in the end of Sect. 3, we obtain

$$\hat f_k = \max_{\beta\ge 0}\min_{y\in Q}\left\{\frac{1}{A_k}\ell_k^{pd}(y)+\Psi(y)+\beta(\xi(x_0,y)-D)\right\} \overset{\beta=1/A_k}{\ge} \frac{1}{A_k}\phi_k^* - \frac{1}{A_k}D.$$

Thus, we can use stopping criterion

$$\tilde f(y_k) - \hat f_k \le \frac{\epsilon}{2}, \tag{4.6}$$

which ensures $\tilde f(y_k) - \tilde f(x^*)\le\epsilon$ as far as

$$A_k \ge \frac{2}{\epsilon}D. \tag{4.7}$$

It remains to estimate from above the total number of calls of oracle of method (4.1), which is sufficient to get an $\epsilon$-solution of problem (2.1). Let us assume that this method is equipped with the stopping criterion (4.6). Then we can be sure that

$$A_k \le \frac{2}{\epsilon}D, \quad k\ge 0. \tag{4.8}$$

Denote by $N(k)$ the total number of calls of oracle after $k$ iterations. At each iteration of this method we call the oracle $2(i_k+1)$ times (at point $x_{k+1,i_k}$ and at the prediction point $y_{k+1,i_k}$). Therefore, using the same reasoning as in the end of Sect. 2, we conclude that

$$N(k) = 4(k+1) + 2\log_2 L_{k+1} - 2\log_2 L_0. \tag{4.9}$$

Note that

$$L_{k+1} = \frac{1}{2}2^{i-k}L_k \;=\; \frac{A_{k+1}}{a_{k+1}^2} \overset{(2.11)}{\le} \left[\frac{1}{\epsilon\tau_k}\right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \;=\; \left[\frac{A_{k+1}}{\epsilon a_{k+1}}\right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \tag{4.10}$$

Therefore $\left[\frac{1}{a_{k+1}}\right]^{\frac{1+3\nu}{1+\nu}} \le A_{k+1}^{\frac{-2\nu}{1+\nu}}\left[\frac{1}{\epsilon}\right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}$, and we conclude that

$$\begin{aligned} L_{k+1} &\le A_{k+1}\left[A_{k+1}^{\frac{-2\nu}{1+\nu}}\left[\frac{1}{\epsilon}\right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}\right]^{\frac{2(1+\nu)}{1+3\nu}} = A_{k+1}^{\frac{1-\nu}{1+3\nu}}\left[\frac{1}{\epsilon}\right]^{\frac{2(1-\nu)}{1+3\nu}} M_\nu^{\frac{4}{1+3\nu}} \\ &\overset{(4.8)}{\le} (2D)^{\frac{1-\nu}{1+3\nu}}\left[\frac{1}{\epsilon}\right]^{\frac{3(1-\nu)}{1+3\nu}} M_\nu^{\frac{4}{1+3\nu}}. \end{aligned}$$

Substituting this estimate in the expression (4.9), we obtain that on average method (4.1) has at most four calls of oracle per iteration.

## 5 Numerical experiments

In our numerical experiments we tried to check the actual level of adaptivity of the above methods to the local topological structure of the objective function. For that, we have chosen three families of nonsmooth minimization problems.

**1. Matrix games** In this problem, given by an $n \times m$ payoff matrix $A$, we need to find a saddle point of the following problem:

$$\begin{aligned}\min_{x\in\Delta_n} \max_{y\in\Delta_m} \langle x, Ay\rangle &= \min_{x\in\Delta_n} \left\{ \psi_p(x) \stackrel{\mathrm{def}}{=} \max_{1\le j\le m} \langle x, Ae_j\rangle \right\} \\ &= \max_{y\in\Delta_m} \left\{ \psi_d(y) \stackrel{\mathrm{def}}{=} \min_{1\le i\le n} \langle e_i, Ay\rangle \right\}, \end{aligned} \quad (5.1)$$

where $e_{(\cdot)}$ denote the basis vectors in the corresponding spaces, and $\Delta_{(\cdot)}$ denotes the standard simplex. This problem can be posed as a minimization problem

$$\min_{x\in\Delta_n, y\in\Delta_m} \left\{ \psi_{pd}(x,y) = \psi_p(x) - \psi_d(y) \right\}. \quad (5.2)$$

The optimal value of this problem is zero. For our experiments, we generated matrix $A$ randomly, with uniform distribution of its entries in the interval $[-1, 1]$.

For feasible set of this problem, $\mathcal{F} = \{z = (x, y) : \ x \in \Delta_n, \ y \in \Delta_m\}$, a natural prox-function is the *entropy*:

$$\eta(z) = \sum_{i=1}^{n} z^{(i)} \ln z^{(i)}.$$

This function is strongly convex with respect to $\ell_1$-norm with the convexity parameter one. Note that $\ell_1$-norm is very good for measuring simplexes since in this "biggest" norm their sizes are still small. At the same time, we can measure the subgradients of the objective function in (5.2) in the "smallest" $\ell_\infty$-norm. In view of our strategy for generating the matrix $A$, we get Lipschitz-continuous function $\psi_{pd}$ with the constant equal to one.

We will refer to the methods based on the entropy function as methods with the *Entropy Setup*. If a method is using the standard Euclidean norm, we say that it is based on the Euclidean setup.

In the table below, we give computational results for two universal methods, the Fast Gradient Method (4.1), and the Primal Gradient Method (2.17), both with Entropy Setup. In our problem instance, $n = 896$ and $m = 128$.

| $Eps$ | $\text{FGM}_{Entropy}$ | | | $\text{PGM}_{Entropy}$ | | |
|---|---|---|---|---|---|---|
| | Iter | Gap | Lip | Iter | Gap | Lip |
| $2^{-5}$ | 516 | $6.0 \cdot 10^{-2}$ | $1.3 \cdot 10^{2}$ | 722 | $8.2 \cdot 10^{-2}$ | 8.0 |
| $2^{-6}$ | 1,127 | $2.9 \cdot 10^{-2}$ | $2.6 \cdot 10^{2}$ | 2,065 | $5.2 \cdot 10^{-2}$ | $1.6 \cdot 10^{1}$ |
| $2^{-7}$ | 1,937 | $1.6 \cdot 10^{-2}$ | $2.0 \cdot 10^{2}$ | 5,675 | $3.4 \cdot 10^{-2}$ | $3.2 \cdot 10^{1}$ |
| $2^{-8}$ | 4,684 | $7.9 \cdot 10^{-3}$ | $2.0 \cdot 10^{3}$ | 15,731 | $2.3 \cdot 10^{-2}$ | $6.4 \cdot 10^{1}$ |
| $2^{-9}$ | 8,129 | $3.8 \cdot 10^{-3}$ | $8.2 \cdot 10^{3}$ | 44,829 | $1.5 \cdot 10^{-2}$ | $1.3 \cdot 10^{2}$ |
| $2^{-10}$ | 17,556 | $2.1 \cdot 10^{-3}$ | $4.1 \cdot 10^{3}$ | 122,959 | $1.0 \cdot 10^{-2}$ | $2.6 \cdot 10^{2}$ |

(5.3)

In the first column we indicate the required accuracy. For each method, there are three subcolumns. First one indicates the number of iterations. Second one shows the upper estimate for the achieved accuracy.[2] The third column shows the current level of "Lipschitz constant", generated by the method. Note that per one iteration of FGM we need on average to call the oracle four times. PGM needs on average only two calls.

It is clear, that both methods behave much better than the worst-case complexity bounds. For FGM, decrease of required accuracy in two times results in a doubling of the number of iterations. This dependence is typical for the complexity bounds of the type $O(\frac{1}{\epsilon})$, and not to the theoretical bound $O(\frac{1}{\epsilon^2})$. For PGM, the average increase is higher (approximately, in three times). But it is still much better than the theoretical bound.

From Table (5.3) we can see that FGM generates much better model of the objective function. Its accuracy is usually only twice bigger than the actual residual. The estimates of PGM are much weaker. One of the possible reasons for this difference consists in much more aggressive behavior of FGM in introducing new gradients in the model.

Finally, the third subcolumn shows that the actual level of the Lipschitz constants are much lower than the theoretical prediction.

Let us look now how efficient FGM is in solving the same problem by the Euclidean setup. These results are presented in the left part of Table (5.4).

[2] Since in this problem the optimal value is known, we use it in the stopping criterion.

| *Eps* | FGM$_{Euclid}$ | | | WDA$_{Entropy}$ | | |
|---|---|---|---|---|---|---|
| | Iter | Gap | Lip | Iter | Gap | Lip |
| $2^{-5}$ | 886 | $1.7 \cdot 10^{-2}$ | $1.0 \cdot 10^{6}$ | 1,569 | $4.4 \cdot 10^{-2}$ | 1.0 |
| $2^{-6}$ | 3,249 | $9.0 \cdot 10^{-2}$ | $8.4 \cdot 10^{6}$ | 6,086 | $2.2 \cdot 10^{-2}$ | 1.0 |
| $2^{-7}$ | 11,803 | $4.8 \cdot 10^{-2}$ | $6.7 \cdot 10^{7}$ | 20,655 | $1.1 \cdot 10^{-2}$ | 1.0 |
| $2^{-8}$ | 45,417 | $2.5 \cdot 10^{-2}$ | $5.4 \cdot 10^{8}$ | 78,832 | $5.5 \cdot 10^{-3}$ | 1.0 |
| $2^{-9}$ | 178,866 | $1.3 \cdot 10^{-2}$ | $4.3 \cdot 10^{9}$ | 283,352 | $2.7 \cdot 10^{-3}$ | 1.0 |
| $2^{-10}$ | out of time | | | out of time | | |

(5.4)

They confirm that the right choice of prox-function is crucial for the efficient solution of optimization problems. Behavior of FGM with Euclidean setup just corresponds to the worst-case theoretical bound $O(\frac{1}{\epsilon^2})$ for Lipschitz-continuous functions (increase of the number of iterations in four times after dividing accuracy by two).

In the right part of this table we present the results of the standard black-box subgradient scheme as applied to the same problem. This is Weighted Dual Averaging (WDA) [12] with Entropy Setup. For choosing its parameters correctly, we need to know only an estimate for the diameter of the feasible set. Each iteration of this method needs one call of oracle. For our problem, WDA works in an exact correspondence to its worst-case complexity bound $O(\frac{1}{\epsilon^2})$. The second column of this part demonstrates that the lower bound generated by this scheme is almost exact.

**2. Continuous Steiner problem** In this problem we are given by centers $a_i \in \mathbb{R}^n$, $i = 1, \ldots, m$. It is necessary to find the optimal location of the service center $x$, which minimizes the total distance to all other centers. Thus, our problem is as follows:

$$\min_{x \in Q} f(x) \stackrel{\mathrm{def}}{=} \sum_{i=1}^{m} \|x - a_i\|. \tag{5.5}$$

where $Q \subseteq \mathbb{R}^n$ is a closed convex set. All norms in this problem are Euclidean.

Clearly, the level of smoothness of problem (5.5) is much higher than that of (5.2). So, we can expect that it is easier for the universal schemes. Let us look at the results of the experiments for random problem with $n = 256$, $m = 512$, and $Q = \mathbb{R}^n_+$. We choose $m > n$ in order to increase the density of nonsmooth points. The centers were generated randomly in the box $0 \leq x^{(i)} \leq \frac{1}{n^{1/2}}$, $i = 1, \ldots, n$ (which has Euclidean diameter one). All methods have origin as a starting point. The initial value of the objective is $f_0 = 295.226$. The optimal solution found by the schemes is $f^* = 147.336$. The table below has the same structure as (5.3).

| $Eps$ | $\mathrm{FGM}_{Euclid}$ | | | $\mathrm{PGM}_{Euclid}$ | | |
|---|---|---|---|---|---|---|
| | Iter | Gap | Lip | Iter | Gap | Lip |
| $2^{-5}$ | 205 | $3.1 \cdot 10^{-2}$ | $2.6 \cdot 10^{2}$ | 9,925 | $3.1 \cdot 10^{-2}$ | $2.6 \cdot 10^{2}$ |
| $2^{-6}$ | 307 | $1.5 \cdot 10^{-2}$ | $5.1 \cdot 10^{2}$ | 19,895 | $1.5 \cdot 10^{-2}$ | $5.1 \cdot 10^{2}$ |
| $2^{-7}$ | 277 | $6.8 \cdot 10^{-3}$ | $2.6 \cdot 10^{2}$ | 39,803 | $7.8 \cdot 10^{-3}$ | $2.6 \cdot 10^{2}$ |
| $2^{-8}$ | 611 | $3.9 \cdot 10^{-3}$ | $5.1 \cdot 10^{2}$ | 77,138 | $3.9 \cdot 10^{-3}$ | $5.1 \cdot 10^{2}$ |
| $2^{-9}$ | 827 | $1.9 \cdot 10^{-3}$ | $5.1 \cdot 10^{2}$ | 155,038 | $2.0 \cdot 10^{-3}$ | $2.6 \cdot 10^{2}$ |
| $2^{-10}$ | 1,226 | $9.8 \cdot 10^{-4}$ | $2.6 \cdot 10^{2}$ | out of time | | |
| $2^{-11}$ | 1,655 | $4.8 \cdot 10^{-4}$ | $2.6 \cdot 10^{2}$ | | | |
| $2^{-12}$ | 2,385 | $2.4 \cdot 10^{-4}$ | $5.1 \cdot 10^{2}$ | | | |
| $2^{-13}$ | 3,388 | $1.2 \cdot 10^{-4}$ | $5.1 \cdot 10^{2}$ | | | |

(5.6)

Note that the rate of convergence of FGM (4.1) is unexpectedly high. Increase of the accuracy in four times results in doubling the number of iterations. From the complexity point of view, this corresponds to the level $O(\frac{1}{\epsilon^{1/2}})$, which is typical for Fast Gradient Methods of *smooth* minimization. The predicted accuracy by FGM is still very good, and the level of Lipschitz constants is unexpectedly small. The results of PGM (2.17) are not so impressive. It doubles the number of iterations after dividing accuracy by two, which corresponds to $O(\frac{1}{\epsilon})$ level of complexity. It seems that a weak point of this method is the quality of termination criterion.

**3. Universal methods and smoothing technique** Let us compare the practical performance of method (4.1) as applied to the primal version of problem (5.1)

$$\min_{x \in \Delta_n} \psi_p(x), \tag{5.7}$$

with its performance as applied to the smoothed version of this function

$$\tilde{\psi}_p(x) = \mu \ln \left( \sum_{j=1}^{m} e^{\langle x, A e_j \rangle / \mu} \right).$$

The value of smoothing parameter $\mu > 0$ for this function is chosen in accordance with the theoretical recommendation (4.8) in [10]. For our experiments, we choose $n = m = 512$ and apply FGM with Entropy Setup.

| *Eps* | FGM for $\tilde{\psi}_p(x)$ | | | FGM for $\psi_p(x)$ | | |
|---|---|---|---|---|---|---|
| | Iter | Gap | Lip | Iter | Gap | Lip |
| $2^{-5}$ | 47 | $3.0 \cdot 10^{-2}$ | $4.0 \cdot 10^{0}$ | 555 | $3.1 \cdot 10^{-2}$ | $1.0 \cdot 10^{3}$ |
| $2^{-6}$ | 103 | $1.5 \cdot 10^{-2}$ | $8.0 \cdot 10^{0}$ | 1,956 | $1.5 \cdot 10^{-2}$ | $1.6 \cdot 10^{4}$ |
| $2^{-7}$ | 226 | $7.6 \cdot 10^{-3}$ | $1.6 \cdot 10^{1}$ | 8,048 | $7.8 \cdot 10^{-3}$ | $2.6 \cdot 10^{5}$ |
| $2^{-8}$ | 464 | $3.9 \cdot 10^{-3}$ | $3.2 \cdot 10^{1}$ | 34,355 | $3.9 \cdot 10^{-3}$ | $1.0 \cdot 10^{6}$ |
| $2^{-9}$ | 953 | $1.9 \cdot 10^{-3}$ | $1.3 \cdot 10^{2}$ | 135,419 | $2.0 \cdot 10^{-3}$ | $8.4 \cdot 10^{6}$ |
| $2^{-10}$ | 1,881 | $9.7 \cdot 10^{-4}$ | $1.3 \cdot 10^{2}$ | out of time | | |
| $2^{-11}$ | 3,653 | $4.9 \cdot 10^{-4}$ | $2.6 \cdot 10^{2}$ | | | |
| $2^{-12}$ | 7,077 | $2.4 \cdot 10^{-4}$ | $2.0 \cdot 10^{3}$ | | | |
| $2^{-13}$ | 13,771 | $1.2 \cdot 10^{-4}$ | $1.0 \cdot 10^{3}$ | | | |

(5.8)

These results confirm that smoothing is still a very powerful technique. Its computational efficiency is often higher even than that of the advanced Black Box Methods.

**Acknowledgments** The author is very thankful to three anonymous referees for careful reading and many suggestions, which significantly improved the initial variant of the paper.

## References

1. Babonneau, F., Nesterov, Yu., Vial, J.-P.: Design and operating of gas transmission networks. In: Operations Research, pp. 1–14, Feb. 2012. doi:10.1287/opre.1110.1001
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
3. Devolder, O., Glineur, F., Nesterov, Yu.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. (2013). doi:10.1007/s10107-013-0677-5
4. Elster, K.-H. (ed.): Modern Mathematical Methods in Optimization. Academie Verlag, Berlin (1993)
5. Lan, G.: Bundle-level methods uniformly optimal for smooth and nonsmooth convex optimization. Math. Program. (2013). doi:10.1007/s10107-013-0737-x
6. Lemarechal, C., Nemirovskii, A., Nesterov, Yu.: New variants of bundle methods. Math. Program. **69**, 111–147 (1995)
7. Nemirovskii, A., Nesterov, Yu.: Optimal methods for smooth convex optimization. Zh. Vychisl. Mat. i Mat. Fiz. **25**(3), 356–369 (1985). in Russian
8. Nemirovsky, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)
9. Nesterov, Yu.: Introductory Lectures on Convex Optimization. Kluwer, Boston (2004)
10. Nesterov, Yu.: Smooth minimization of non-smooth functions. Math. Program. A **103**(1), 127–152 (2005)

11. Nesterov, Yu.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
12. Nesterov, Yu.: Primal-dual subgradient methods for convex problems. Math. Program. **120**(1), 261–283 (2009)