

Primal-dual accelerated gradient methods with small-dimensional relaxation oracle

Yurii Nesterov^a and Alexander Gasnikov^{b,c} and Sergey Guminov^{b,c} and Pavel Dvurechensky^{d,c}

^a Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Louvain-la-Neuve, Belgium; National Research University Higher School of Economics, Moscow, Russia; ^b Moscow Institute of Physics and Technology, Dolgoprudny, Russia; ^c Institute for Information Transmission Problems RAS, Moscow, Russia; ^d Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

ARTICLE HISTORY

Compiled May 14, 2019

Abstract

In this paper, a new variant of accelerated gradient descent is proposed. The proposed method does not require any information about the objective function, uses exact line search for the practical accelerations of convergence, converges according to the well-known lower bounds for both convex and non-convex objective functions, possesses primal-dual properties and can be applied in the non-euclidian set-up. As far as we know this is the first such method possessing all of the above properties at the same time. We also present a universal version of the method which is applicable to non-smooth problems. We demonstrate how in practice one can efficiently use the combination of line-search and primal-duality by considering a convex optimization problem with a simple structure (for example, linearly constrained).

KEYWORDS

accelerated gradient descent, line-search, primal-dual methods, convex optimization, nonconvex optimization

AMS CLASSIFICATION

90C25, 68Q25

1. Introduction

The first accelerated gradient method for smooth convex optimization problems dates back to 1980s [23]. This method has optimal [19] convergence rate $f(x^k) - f(x_*) = O(1/k^2)$, where k is the iteration counter, f is the objective function and x_* is an optimal point. It is less known that there were earlier versions of optimal methods. The key difference is that, on each step, those versions used small-dimensional relaxation oracle (sDR-oracle), i.e. auxiliary minimization over some small-dimensional subspace. Thus, these early methods are optimal under additional assumption of availability of the sDR-oracle. This assumption is rarely satisfied in practice, since solving auxiliary minimization problem on each step can be very costly or can lead to divergence of the method due to accumulating error in practice. This seems to be the main reason why mostly the fixed-step accelerated methods [23] have been developing in

the last decades [25, 26, 28]. This choice of the direction of research by the community is supported by the fact that availability of sDR-oracle does not improve the worst-case theoretical guarantee of the optimal gradient-type procedure for smooth convex optimization problems [22].

There has been a resurging interest in first-order methods using sDR-oracle [17, 7, 14]. One of the reasons is that, in practice, small-dimensional relaxation allows local adaptation to the curvature of the objective function, which can dramatically improve the practical convergence rate, the classical example being conjugate gradient methods. At the same time, there are problem classes for which the computational overhead of the sDR-oracle is minimal [17]. This includes popular machine learning problems known as generalized linear models, including SVM, linear regression, logistic regression, etc., and optimization problems with linear equality constraints. In the first case, the problem itself is of the form

$$f(x) = F(A^T x) \rightarrow \min_{x \in \mathbb{R}^n}, \quad (1)$$

where $A \in \mathbb{R}^{n \times m}$. In the latter case, the dual problem has the same form as (1). Even if the matrix A is large and dense, and $F(y)$ can be calculated in $O(n)$ arithmetic operations, the whole small-dimensional relaxation has almost the same complexity as a single calculation of the gradient of $f(Ax)$. Hence, in this case, the operation complexities of the fixed-step methods and the methods with small-dimensional relaxation are almost identical.

In this paper, we propose a new accelerated gradient method utilizing a sDR-oracle. Instead of using pre-defined sequences of parameters of the method, such as step sizes, we choose these parameters by minimizing the objective in some specially constructed direction. This makes our method related to conjugate gradient methods. The difference is that our method has $O(1/k^2)$ convergence rate for general convex objectives. Moreover, our method is adaptive to the smoothness of the objective and does not require the Lipschitz constant of the gradient to be known, unlike classic accelerated gradient descent method [24]. The method also converges to a stationary point for non-convex objectives, which means that it is capable of adapting to the local convexity of the objective. We also estimate the rate of convergence of the method in terms of the norm of the gradient for convex, γ -weakly-quasi-convex and non-convex objectives.

Further on, we analyze potentially non-smooth functions with Hölder-continuous subgradient and propose a generalization of our method for this case. We prove that our method is universal in the sense of [29], i.e. does not require any a priori knowledge of the smoothness of the objective and automatically according to the lower bounds for the class of convex objectives with Hölder-continuous subgradient.

Special attention is devoted to the analysis of the primal-dual properties of the proposed methods for the class of strongly convex linearly constrained problems. In this case the dual problem has the form (1) and is solved by our method with the ultimate goal to reconstruct the solution to the primal problem.

Finally, we describe how these methods can be accelerated to have optimal linear convergence under the additional assumption that the objective is strongly convex with known parameter of strong convexity.

Related work. It is quite hard to cover all the vast literature on accelerated gradient methods [23, 26, 3, 28, 24]. The versions adaptive to the Lipschitz constant of the gradient may be found in [28, 3]. Usually this type of adaptivity is called "line-

search” or ”backtracking”. Papers [4, 10, 9] consider the question of primal-duality of these methods in combination with ”backtracking” used for adaptivity to the Lipschitz constant. The authors of [11], by a special a priori choice of step sizes, construct a single method which works optimally for convex and non-convex problems. Universal gradient methods for convex problems were proposed in [29] and extended in [12] for non-convex problems and in [31] for primal-dual setting. Finally, there were some attempts to combine universality with small-dimension relaxation [8, 14]. Concerning conjugate gradient methods, we refer the reader to a good survey [2] and the classical book [30].

The structure of the paper is as follows. In section 2 we describe the main algorithm and analyze its convergence. We also show that this method admits a stopping criterion. The subsections of section 2 are dedicated to the modifications of the method applicable to γ -weakly-quasi-convex and strongly convex objectives. Then the dependence of the method on line-search accuracy is discussed briefly. In section 3 we present the universal version of the algorithm and establish its convergence for convex and non-convex objectives. Once again, a subsection is dedicated to strongly convex objectives. Section 4 contains the results concerning the primal-dual properties of our method. Finally, in section 5 one can find the results of numerical experiments.

Notation. Let E be a finite-dimensional real vector space and E^* be its dual. We denote the value of a linear function $g \in E^*$ at $x \in E$ by $\langle g, x \rangle$. Let $\|\cdot\|$ be some norm on E , $\|\cdot\|_*$ be its dual, defined by $\|g\|_* = \max_x \{\langle g, x \rangle, \|x\| \leq 1\}$. Given a vector $g \in E^*$, we denote by $(g)^\# = \arg \max_{\|s\| \leq 1} \langle g, s \rangle$. We use $\nabla f(x)$ to denote any subgradient of a function f at a point $x \in \text{dom} f$.

We choose a *prox-function* $d(x)$, which is continuous, convex on Q and

- (1) admits a continuous in $x \in Q^0$ selection of subgradients $\nabla d(x)$, where $Q^0 \subseteq Q$ is the set of all x , where $\nabla d(x)$ exists;
- (2) $d(x)$ is 1-strongly convex on Q with respect to $\|\cdot\|$, i.e., for any $x \in Q^0, y \in Q$

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

Without loss of generality, we assume that $\min_{x \in Q} d(x) = 0$.

We define also the corresponding *Bregman divergence* $V(x, z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$, $x \in Q, z \in Q^0$. Standard proximal setups, i.e. Euclidean, entropy, ℓ_1/ℓ_2 , simplex, nuclear norm, spectahedron can be found in [5].

2. Adaptive methods for smooth optimization

We consider the optimization problem

$$f(x) \rightarrow \min_{x \in E}, \tag{2}$$

and denote a solution to this problem as x_* . Our main assumption in this section is that the objective f is L -smooth, i.e. is continuously differentiable and has Lipschitz-continuous gradient

$$\|\nabla f(y) - \nabla f(x)\| \leq L \|x - y\|_*, \quad \forall x, y \in E. \tag{3}$$

Our main algorithm in this section is listed as Algorithm 1.

Algorithm 1 Accelerated Gradient Method with Small-Dimensional Relaxation (AGMsDR)

Output: x^k

- 1: Set $k = 0$, $A_0 = 0$, $x^0 = v^0$, $\psi_0(x) = V(x, x^0)$
- 2: **for** $k \geq 0$ **do**
- 3:

$$\beta_k = \arg \min_{\beta \in [0,1]} f \left(v^k + \beta(x^k - v^k) \right), \quad y^k = v^k + \beta_k(x^k - v^k). \quad (4)$$

- 4: Option a), L is known,

$$x^{k+1} = \arg \min_{x \in E} \left\{ f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{L}{2} \|x - y^k\|^2 \right\}. \quad (5)$$

Find a_{k+1} from equation $\frac{a_{k+1}^2}{A_k + a_{k+1}} = \frac{1}{L}$.

Option b),

$$h_{k+1} = \arg \min_{h \geq 0} f \left(y^k - h(\nabla f(y^k))^{\#} \right), \quad x^{k+1} = y^k - h_{k+1}(\nabla f(y^k))^{\#}. \quad (6)$$

Find a_{k+1} from equation $f(y^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})} \|\nabla f(y^k)\|_*^2 = f(x^{k+1})$.

- 5: Set $A_{k+1} = A_k + a_{k+1}$.
 - 6: Set $\psi_{k+1}(x) = \psi_k(x) + a_{k+1} \{ f(y^k) + \langle \nabla f(y^k), x - y^k \rangle \}$.
 - 7: $v^{k+1} = \arg \min_{x \in E} \psi_{k+1}(x)$
 - 8: $k = k + 1$
 - 9: **end for**
-

Here and for all the methods described further we assume that if the equation for a_{k+1} in step 4 admits multiple solutions, then the greater one is chosen.

Before we move to the theoretical results of this section, let us make some remarks. The main new element of the proposed method is in line 3. Unlike known methods [25, 26, 1], which use fixed $\beta_k = \frac{k}{k+2}$, we use minimization over the interval $\beta \in [0, 1]$. The choice of the fixed stepsize is motivated by the theoretical convergence analysis. Our goal is to choose best possible stepsize with the same convergence rate guarantees. Most of the results described further remain the same if the search over the unit interval $[0, 1]$ in line 3 is changed to line-search over any subset of \mathbb{R} containing said interval, for instance, the whole real line \mathbb{R} .

Theoretical analysis of Algorithm 1 is based on the following theorem. We underline that the convexity of the objective f is not required.

Theorem 1. *After k steps of Algorithm 1 for problem (2) it holds that*

$$A_k f(x^k) \leq \min_{x \in \mathbb{R}^n} \psi_k(x) = \psi_k(v^k). \quad (7)$$

Moreover, $A_k \geq \frac{k^2}{4L}$.

Proof. Denote

$$l_k(x) = \sum_{i=0}^k a_{i+1} \{f(y^i) + \langle \nabla f(y^i), x - y^i \rangle\}.$$

Then

$$\psi_{k+1}(x) = l_k(x) + \psi_0(x) = \psi_k(x) + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\}.$$

First, we prove inequality (7) by induction over k . For $k = 0$, the inequality holds. Assume that

$$A_k f(x^k) \leq \min_{x \in \mathbb{R}^n} \psi_k(x) = \psi_k(v^k).$$

Then

$$\begin{aligned} \psi_{k+1}(v^{k+1}) &= \min_{x \in \mathbb{R}^n} \left\{ \psi_k(x) + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\} \right\} \\ &\geq \min_{x \in \mathbb{R}^n} \left\{ \psi_k(v^k) + \frac{1}{2} \|x - v^k\|^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\} \right\} \\ &\geq \min_{x \in \mathbb{R}^n} \left\{ A_k f(x^k) + \frac{1}{2} \|x - v^k\|^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\} \right\}. \end{aligned}$$

Here we used that ψ_k is a strongly convex function with minimum at v^k .

By the definition of β_k and y^k in (4), we have $f(y^k) \leq f(x^k)$. By the optimality conditions in (4), either

- (1) $\beta_k = 0$, $\langle \nabla f(y^k), x^k - v^k \rangle \geq 0$, $y^k = v^k$;
- (2) $\beta_k \in (0, 1)$ and $\langle \nabla f(y^k), x^k - v^k \rangle = 0$, $y^k = v^k + \beta_k(x^k - v^k)$;
- (3) $\beta_k = 1$ and $\langle \nabla f(y^k), x^k - v^k \rangle \leq 0$, $y^k = x^k$.

In all three cases, $\langle \nabla f(y^k), v^k - y^k \rangle \geq 0$. Thus,

$$\begin{aligned} \psi_{k+1}(v^{k+1}) &\geq \min_{x \in \mathbb{R}^n} \left\{ A_{k+1} f(y^k) + a_{k+1} \langle \nabla f(y^k), x - v^k \rangle + \frac{1}{2} \|x - v^k\|^2 \right\} \\ &\geq A_{k+1} f(y^k) - \frac{a_{k+1}^2}{2} \|\nabla f(y^k)\|_*^2, \end{aligned}$$

where we used that for any $g \in E^*$, $s \in E$, $\zeta \geq 0$, it holds that $\langle g, s \rangle + \frac{\zeta}{2} \|s\|^2 \geq -\frac{1}{2\zeta} \|g\|_*^2$. Our next goal is to show that

$$A_{k+1} f(y^k) - \frac{a_{k+1}^2}{2} \|\nabla f(y^k)\|_*^2 \geq A_{k+1} f(x^{k+1}), \quad (8)$$

which proves the induction step.

For option a) in the step 4, using the L -smoothness of f and minimizing the r.h.s. of (5), we have

$$\begin{aligned} f(x^{k+1}) &\leq f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k \rangle + \frac{L}{2} \|x^{k+1} - y^k\|^2 = \min_{x \in E} \left(f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{L}{2} \|x - y^k\|^2 \right) \\ &= f(y^k) - \frac{1}{2L} \|\nabla f(y^k)\|_*^2. \end{aligned}$$

Since, for this option, $\frac{a_{k+1}^2}{A_{k+1}} = \frac{1}{L}$, inequality (8) holds. For option b) in the step 4, (8) holds by the choice of a_{k+1} from the equation

$$f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}} \|\nabla f(y^k)\|_*^2 = f(x^{k+1}). \quad (9)$$

It remains to show that this equation has a solution $a_{k+1} > 0$. By the L -smoothness of f , we have

$$\begin{aligned} f(x^{k+1}) &= \min_{h \geq 0} f\left(y^k - h(\nabla f(y^k))^\# \right) \leq \min_{h \geq 0} \left(f(y^k) - h \langle \nabla f(y^k), (\nabla f(y^k))^\# \rangle + \frac{Lh^2}{2} \|(\nabla f(y^k))^\#\|^2 \right) \\ &= f(y^k) - \frac{1}{2L} \|\nabla f(y^k)\|_*^2, \end{aligned} \quad (10)$$

where we used that $\langle \nabla f(y^k), (\nabla f(y^k))^\# \rangle = \|\nabla f(y^k)\|_*^2$ and $\|(\nabla f(y^k))^\#\|^2 = 1$ by definition of the vector $(\nabla f(y^k))^\#$. Since $A_{k+1} = A_k + a_{k+1}$, we can rewrite the equation (9) as

$$\frac{a_{k+1}^2}{2} \|\nabla f(y^k)\|_*^2 + a_{k+1}(f(x^{k+1}) - f(y^k)) + A_k(f(x^{k+1}) - f(y^k)) = 0.$$

Since, by (10), $f(x^{k+1}) - f(y^k) < 0$ (otherwise $\|\nabla f(y^k)\|_* = 0$ and y_k is a solution to the problem (2)),

$$a_{k+1} = \frac{f(y^k) - f(x^{k+1}) + \sqrt{(f(y^k) - f(x^{k+1}))^2 - 2A_k(f(x^{k+1}) - f(y^k))\|\nabla f(y^k)\|_*^2}}{\|\nabla f(y^k)\|_*^2} > 0.$$

Let us estimate the rate of the growth for A_k . If in the step 4 option a) is used, $\frac{a_{k+1}^2}{A_{k+1}} = \frac{1}{L}$. For the option b), using (9) and (10), we have $\frac{a_{k+1}^2}{A_{k+1}} \geq \frac{1}{L}$. Thus, for both options, $\frac{a_{k+1}^2}{A_k + a_{k+1}} = \frac{a_{k+1}^2}{A_{k+1}} \geq \frac{1}{L}$. Since $A_1 = a_1 \geq \frac{1}{L}$, we prove by induction that $\alpha_k \geq \frac{k}{2L}$ and $A_k \geq \frac{(k+1)^2}{4L} \geq \frac{k^2}{4L}$.

Indeed,

$$\begin{aligned} \alpha_{k+1} &\geq \frac{1 + \sqrt{1 + 4A_k L}}{2L} = \frac{1}{2L} + \sqrt{\frac{1}{4L^2} + \frac{A_k}{L}} \geq \frac{1}{2L} + \sqrt{\frac{A_k}{L}} \\ &\geq \frac{1}{2L} + \frac{1}{\sqrt{L}} \frac{k+1}{2\sqrt{L}} = \frac{k+2}{2L}. \end{aligned}$$

Hence,

$$A_{k+1} = A_k + \alpha_{k+1} \geq \frac{(k+1)^2}{4L} + \frac{k+2}{2L} \geq \frac{(k+2)^2}{4L}.$$

□

Next result is simple and standard for gradient methods, but we provide it for the sake of completeness of the paper.

Theorem 2. *Let function f be L -smooth and Algorithm 1 be run for N steps. Then*

$$\min_{k=0,\dots,N} \|\nabla f(y^k)\|_*^2 \leq \frac{2L(f(x^0) - f(x_*))}{N}.$$

Proof. We have that

$$f(x^{k+1}) \leq f(y^k) - \frac{1}{2L} \|\nabla f(y^k)\|_*^2 \leq f(x^k) - \frac{1}{2L} \|\nabla f(y^k)\|_*^2. \quad (11)$$

Summing this up for $k = 0, \dots, N$, we obtain

$$f(x^0) - f(x_*) \geq f(x^0) - f(x^{N+1}) \geq \frac{N}{2L} \min_{k=0,\dots,N} \|\nabla f(y^k)\|_*^2.$$

Consequently, we may guarantee

$$\min_{k=0,\dots,N} \|\nabla f(y^k)\|_2^2 \leq \frac{2L(f(x^0) - f(x_*))}{N}.$$

□

Before we move to the main results, we define γ -weakly-quasi-convex functions, which are unimodal, but generally non-convex. We say that $f(x)$ is γ -weakly-quasi-convex with $\gamma \in (0, 1]$ if for all $x \in \mathbb{R}^n$

$$\gamma(f(x) - f(x_*)) \leq \langle \nabla f(x), x - x_* \rangle.$$

Note that convex functions are 1-weakly-quasi-convex. The converse is generally not true.

Lemma 1. *Let function f be γ -weakly-quasi-convex and Algorithm 1 be run for N steps. Then*

$$A_k(f(x^k) - f(x_*)) \leq (1 - \gamma)A_k(f(x^0) - f(x_*)) + V(x_*, x^0).$$

Proof. According to the definition of γ -weak-quasi-convexity

$$l_k(x_*) = \sum_{i=0}^k a_{i+1} \{f(y^i) + \langle \nabla f(y^i), x_* - y^i \rangle\} \leq$$

$$\leq \sum_{i=0}^k a_{i+1} \{(1-\gamma)f(y^i) + \gamma f(x_*)\}.$$

By (11) and (4) we have $f(y^i) \leq f(x^i) \leq f(x^0)$, so

$$l_k(x_*) \leq \sum_{i=0}^k a_{i+1} \{(1-\gamma)f(x^0) + \gamma f(x_*)\}.$$

From this inequality and **Theorem 1** we have

$$\begin{aligned} A_k f(x_k) &\leq \min_{x \in \mathbb{R}^n} \psi_k(x) \leq \psi_k(x_*) = l_{k-1}(x_*) + V(x_*, x^0) \leq \\ &\leq \sum_{i=0}^{k-1} a_{i+1} \{(1-\gamma)f(x^0) + \gamma f(x_*)\} + V(x_*, x^0). \end{aligned}$$

From here, since $A_k = \sum_{i=0}^{k-1} a_{i+1}$, by rearranging the terms we obtain the statement of the theorem:

$$A_k(f(x^k) - f(x_*)) \leq (1-\gamma)A_k(f(x^0) - f(x_*)) + V(x_*, x^0).$$

□

Theorem 3. *Let function f be 1-weakly-quasi-convex and L -smooth and Algorithm 1 be run for N steps. Then*

$$\min_{k=\lceil N/2 \rceil, \dots, N} \|\nabla f(y^k)\|_*^2 \leq \frac{64L^2V(x_*, x^0)}{N^3}, \quad (12)$$

$$f(x^N) - f(x_*) \leq \frac{4LV(x_*, x^0)}{N^2}.$$

Proof. Applying **Lemma 1** with $\gamma = 1$, we get

$$f(x^N) - f(x_*) \leq \frac{V(x_*, x^0)}{A_N}.$$

Using the lower bound on A_N established in **Theorem 1** we obtain that for a convex (or 1-weakly-quasi-convex) objective

$$f(x^N) - f(x_*) \leq \frac{4LV(x_*, x^0)}{N^2}.$$

Summing up (11) for $k = \lceil N/2 \rceil, \dots, N$, we obtain

$$f(x^{[N/2]}) - f(x_*) \geq f(x^{[N/2]}) - f(x^{N+1}) \geq \sum_{k=[N/2]}^N \frac{\|\nabla f(y^k)\|_2^2}{2L} \geq [N/2] \min_{k=[N/2], \dots, N} \frac{\|\nabla f(y^k)\|_2^2}{2L}.$$

Finally, we have

$$\min_{k=[N/2], \dots, N} \|\nabla f(y^k)\|_2^2 \leq \frac{4L}{N} (f(x^{[N/2]}) - f(x_*)) \leq \frac{64L^2 V(x_*, x^0)}{N^3}.$$

□

Remark 1. Recently in [16] a special variant of accelerated gradient descent that converges at the rate

$$\|\nabla f(x^N)\|_2^2 = O\left(\frac{L(f(x^0) - f(x_*))}{N^2}\right). \quad (13)$$

was proposed.

This result seems to be weaker than (12), but actually from (13) one can obtain a much stronger result. Indeed, one can perform N iterations of common fast gradient descent and obtain

$$f(x^N) - f(x_*) = O\left(\frac{LR^2}{N^2}\right).$$

Then one can put $x^0 := x^N$ and perform N iterations of the method from [16]. Totally, we obtain

$$\|\nabla f(x^N)\|_2^2 = O\left(\frac{L^2 R^2}{N^4}\right).$$

This bound and the bound (13) are unimprovable, see [20, 27].

Remark 2. Note that our method does not require the knowledge about the convexity of the objective function and automatically works either with rate given by Theorem 2 or by Theorem 3.

2.0.1. Online stopping criterion

If the objective is smooth and convex, this method admits an efficient stopping criterion.

By rewriting the statement of **Theorem 1** we see that

$$f(x^k) \leq \frac{1}{A_k} \psi_k(v^k) = \frac{1}{A_k} \min_{x \in \mathbb{R}^n} \left[\frac{1}{2} \|x^0 - x\|_2^2 + \sum_{i=0}^{k-1} a_{i+1} \{f(y^i) + \langle \nabla f(y^i), x - y^i \rangle\} \right]$$

As it was before,

$$l^{k-1}(x) = \sum_{i=0}^{k-1} a_{i+1} \left\{ f(y^i) + \langle \nabla f(y^i), x - y^i \rangle + \frac{\mu}{2} \|x - y^i\|_2^2 \right\}$$

Denote $R = \|x^0 - x_*\|_2$ and

$$\hat{f}^k = \min_{x: \|x - x_0\| \leq R} \frac{1}{A_k} l^{k-1}(x).$$

The constraint may be rewritten equivalently as $\frac{1}{2} \|x - x_0\|^2 \leq \frac{R^2}{2}$. By strong duality we see that

$$\begin{aligned} \hat{f}^k &= \min_{x \in \mathbb{R}^n} \max_{\lambda \geq 0} \left\{ \frac{1}{A_k} l^{k-1}(x) + \lambda \left(\frac{1}{2} \|x_0 - x\|^2 - \frac{R^2}{2} \right) \right\} \\ &= \max_{\lambda \geq 0} \min_{x \in \mathbb{R}^n} \left\{ \frac{1}{A_k} l^{k-1}(x) + \lambda \left(\frac{1}{2} \|x_0 - x\|^2 - \frac{R^2}{2} \right) \right\}. \end{aligned}$$

Now we set $\lambda = \frac{1}{A_k}$ and obtain

$$\hat{f}^k \geq \frac{1}{A_k} \psi_k(v^k) - \frac{R^2}{2A_k}.$$

Then

$$f(x^k) - \hat{f}^k \leq \frac{R^2}{2A_k}.$$

But by convexity of $f(x)$ we have that $\forall k \frac{1}{A_k} l^{k-1}(x) \leq f(x)$, which implies that $\hat{f}^k \leq f(x_*)$. Finally, we have

$$f(x^k) - f(x_*) \leq f(x^k) - \hat{f}^k \leq \frac{R^2}{2A_k},$$

so the condition $f(x^k) - \hat{f}^k \leq \varepsilon$ is an efficient stopping criterion for the AGMsDR method.

2.1. γ -weakly-quasi-convex objectives

Next we describe a method for more general class of γ -weakly-quasi-convex functions. Algorithm 2 is obtained from Algorithm 1 by applying a restart technique.

Denote by \tilde{x}^i the sequence of all iterates x_i^j generated by the above method

Theorem 4. *If $f(x)$ is γ -weakly-quasi-convex and L -smooth function, then*

$$f(\tilde{x}^N) - f(x_*) = O\left(\frac{LR^2}{\gamma^3 N^2}\right),$$

Algorithm 2 Accelerated Gradient Method with Small-Dimensional Relaxation (AGMsDR)

Output: x_i^k

- 1: **for** $i \geq 0$ **do**
- 2: Set $k = 0$, $A_0 = 0$, $x_i^0 = v_i^0$, $\psi_0^i(x) = V(x, x_i^0)$
- 3: **for** $k \geq 0$ **do**
- 4:

$$\beta_k = \arg \min_{\beta \in [0,1]} f\left(v_i^k + \beta(x_i^k - v_i^k)\right), \quad y_i^k = v_i^k + \beta_k(x_i^k - v_i^k). \quad (14)$$

- 5: Option a), L is known,

$$x_i^{k+1} = \arg \min_{x \in E} \left\{ f(y_i^k) + \langle \nabla f(y_i^k), x - y_i^k \rangle + \frac{L}{2} \|x - y_i^k\|^2 \right\}. \quad (15)$$

Find a_{k+1} from equation $\frac{a_{k+1}^2}{A_k + a_{k+1}} = \frac{1}{L}$.

Option b),

$$h_{k+1} = \arg \min_{h \geq 0} f\left(y_i^k - h(\nabla f(y_i^k))^\# \right), \quad x_i^{k+1} = y_i^k - h_{k+1}(\nabla f(y_i^k))^\#. \quad (16)$$

Find a_{k+1} from equation $f(y_i^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})} \|\nabla f(y_i^k)\|_*^2 = f(x_i^{k+1})$.

- 6: Set $A_{k+1} = A_k + a_{k+1}$.
 - 7: Set $\psi_{k+1}(x) = \psi_k(x) + a_{k+1} \{f(y_i^k) + \langle \nabla f(y_i^k), x - y_i^k \rangle\}$.
 - 8: $v_i^{k+1} = \arg \min_{x \in E} \psi_{k+1}(x)$
 - 9: **if** $f(x_i^k) - f(x_*) \leq (1 - \gamma/2) (f(x_i^0) - f(x_*))$ **then**
 - 10: **break**
 - 11: **end if**
 - 12: $k = k + 1$
 - 13: **end for**
 - 14: Set $x_{i+1}^0 = x_i^N$.
 - 15: $i = i + 1$
 - 16: **end for**
-

where $R = \max_{x: f(x) \leq f(x_0)} \|x\|$.

Proof. Denote $\varepsilon_0 = f(x_0^0) - f_*$. From **Lemma 1** and **Theorem 1** we have that

$$f(x_0^k) - f(x_*) \leq (1 - \gamma)\varepsilon_0 + \frac{2LR_0^2}{k^2},$$

where $R_0 = \|x_0 - x_*\|_2$. We need to ensure

$$f(x_0^k) - f(x_*) \leq (1 - \gamma/2)\varepsilon_0.$$

That means that the method is first restarted no later than after $N_0 = \left\lceil 2\sqrt{\frac{LR^2}{\gamma\varepsilon_0}} \right\rceil$

iterations. Denote $R_1 = \|x_1^0 - x_*\|_2$. Again we apply **Lemma 1** and **Theorem 1**, we have

$$f(x_1^k) - f(x_*) \leq (1 - \gamma)(1 - \gamma/2)\varepsilon_0 + \frac{2LR_i^2}{k^2} \leq (1 - \gamma/2)^2\varepsilon_0,$$

which implies that the second restart happens no later than after $N_1 = \left\lceil 2\sqrt{\frac{LR_1^2}{\gamma(1-\gamma/2)\varepsilon_0}} \right\rceil$. By proceeding in the same way we show that no more than $N_i = \left\lceil 2\sqrt{\frac{LR_i^2}{\gamma(1-\gamma/2)^i\varepsilon_0}} \right\rceil$ iterations happen between the i -th and the $i+1$ -th restarts.

Let $d = \log_{1-\gamma/2} \frac{\varepsilon}{\varepsilon_0}$. Then an ε -solution is obtained in no more than $N = \sum_{i=0}^d N_i$ iterations. We also know that the sequence $f(x_i^0)$ is non-increasing, so $\forall i R_i \leq 2R = 2 \max_{x: f(x) \leq f(x_0)} \|x\|$. It follows from our restart rule that $\varepsilon < \varepsilon_0(1-\gamma/2)^{d-1}$. Then we have the following sequence of relations:

$$\begin{aligned} N &\leq \sum_{i=0}^d \left\lceil 2\sqrt{\frac{LR_i^2}{\gamma(1-\gamma/2)^i\varepsilon_0}} \right\rceil \leq d+1 + \sum_{i=0}^d 2\sqrt{\frac{4LR^2}{\gamma\varepsilon}} (1-\gamma/2)^{\frac{d-i+1}{2}} \leq \\ &\leq d+1 + 2\sqrt{\frac{4LR^2}{\gamma\varepsilon}} \sum_{i=-\infty}^d (1-\gamma/2)^{\frac{d-i+1}{2}} = d+1 + 2\sqrt{\frac{4LR^2}{\gamma\varepsilon}} \frac{\sqrt{1-\gamma/2}}{1-\sqrt{1-\gamma/2}} = \\ &= d+1 + 2\sqrt{4\frac{LR^2}{\gamma\varepsilon}} \frac{\sqrt{1-\gamma/2}(1+\sqrt{1-\gamma/2})}{\gamma/2} = d+1 + 3\sqrt{\frac{4LR^2}{\gamma^3\varepsilon}} = O\left(\sqrt{\frac{LR^2}{\gamma^3\varepsilon}}\right). \end{aligned}$$

□

Note, that using the Sequential Subspace Optimization Method [18] Guminov et al.[13] show that the last bound in theorem 4 can be improved under small γ to

$$f(\tilde{x}^N) - f(x_*) = O\left(\frac{LR^2}{\gamma^2 N^2}\right).$$

However, this requires solving a three-dimensional non-convex problem on each iteration. The method in this paper, on the other hand, only requires solving one minimization problem over an interval. If R is known, the stopping criterion may be used to restart the method.

2.2. Strongly convex objectives

Assume now that the objective function in problem (2) is μ -strongly convex with respect to the Euclidean norm:

$$\forall x, y \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \|y - x\|_2^2.$$

Next we describe two different ways to modify or method in order to deal with strongly convex objective functions.

The first way is to consider a slightly different estimating sequence:

$$\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2}\|x - y^k\|^2\}.$$

This leads us to the following method.

Algorithm 3 Accelerated Gradient Method with Small-Dimensional Relaxation (AGMsDR)

Output: x^k

- 1: Set $k = 0$, $A_0 = 0$, $x^0 = v^0$, $\psi_0(x) = \frac{1}{2}\|x - x_0\|_2^2$, $\tau_0 = 1$
- 2: **for** $k \geq 0$ **do**
- 3:

$$\beta_k = \arg \min_{\beta \in [0,1]} f\left(v^k + \beta(x^k - v^k)\right), \quad y^k = v^k + \beta_k(x^k - v^k). \quad (17)$$

- 4: Option a), L is known,

$$x^{k+1} = \arg \min_{x \in E} \left\{ f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{L}{2}\|x - y^k\|^2 \right\}. \quad (18)$$

Find a_{k+1} from equation $\frac{a_{k+1}^2}{(\tau_k + \mu a_{k+1})(A_k + a_{k+1})} = \frac{1}{L}$.

Option b),

$$h_{k+1} = \arg \min_{h \geq 0} f\left(y^k - h(\nabla f(y^k))^\#\right), \quad x^{k+1} = y^k - h_{k+1}(\nabla f(y^k))^\#. \quad (19)$$

Find a_{k+1} from equation $f(y^k) - \frac{a_{k+1}^2}{2(\tau_k + \mu a_{k+1})(A_k + a_{k+1})} \|\nabla f(y^k)\|_2^2 + \frac{\mu \tau_k a_{k+1}}{2(\tau_k + \mu a_{k+1})(A_k + a_{k+1})} \|v^k - y^k\|^2 = f(x^{k+1})$.

- 5: Set $A_{k+1} = A_k + a_{k+1}$, $\tau_{k+1} = \tau_k + \mu a_{k+1}$.
 - 6: Set $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2}\|x - y^k\|^2\}$.
 - 7: $v^{k+1} = \arg \min_{x \in E} \psi_{k+1}(x)$
 - 8: $k = k + 1$
 - 9: **end for**
-

Theorem 5. *After k steps of Algorithm 3 for problem (2) it holds that*

$$A_k f(x^k) \leq \min_{x \in \mathbb{R}^n} \psi_k(x) = \psi_k(v^k). \quad (20)$$

Moreover,

$$A_k \geq \max \left\{ \frac{k^2}{4L}, \frac{1}{L} \left(1 - \sqrt{\frac{\mu}{L}}\right)^{-(k-1)} \right\}.$$

Proof. Denote

$$l_k(x) = \sum_{i=0}^k a_{i+1} \{f(y^i) + \langle \nabla f(y^i), x - y^i \rangle + \frac{\mu}{2} \|x - y^i\|^2\}.$$

Then

$$\psi_{k+1}(x) = l_k(x) + \psi_0(x) = \psi_k(x) + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|^2\}.$$

Note that ψ_k is a sum of a 1-strongly convex function ψ_0 and μa_i -strongly convex functions for $i = 1, \dots, k$, which means that ψ_k is τ_k -strongly convex, where $\tau_k = 1 + \mu \sum_{i=1}^k a_i = 1 + \mu A_k$.

First, we prove inequality (20) by induction over k . For $k = 0$, the inequality holds. Assume that

$$A_k f(x^k) \leq \min_{x \in \mathbb{R}^n} \psi_k(x) = \psi_k(v^k).$$

Then

$$\begin{aligned} \psi_{k+1}(v^{k+1}) &= \min_{x \in \mathbb{R}^n} \left\{ \psi_k(x) + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|^2\} \right\} \\ &\geq \min_{x \in \mathbb{R}^n} \left\{ \psi_k(v^k) + \frac{\tau_k}{2} \|x - v^k\|^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|^2\} \right\} \\ &\geq \min_{x \in \mathbb{R}^n} \left\{ A_k f(x^k) + \frac{\tau_k}{2} \|x - v^k\|^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|^2\} \right\}. \end{aligned}$$

Here we used that ψ_k is a τ_k -strongly convex function with minimum at v^k .

By the definition of β_k and y^k in (17), we have $f(y^k) \leq f(x^k)$. By the optimality conditions in (17), either

- (1) $\beta_k = 0$, $\langle \nabla f(y^k), x^k - v^k \rangle \geq 0$, $y^k = v^k$;
- (2) $\beta_k \in (0, 1)$ and $\langle \nabla f(y^k), x^k - v^k \rangle = 0$, $y^k = v^k + \beta_k(x^k - v^k)$;
- (3) $\beta_k = 1$ and $\langle \nabla f(y^k), x^k - v^k \rangle \leq 0$, $y^k = x^k$.

In all three cases, $\langle \nabla f(y^k), v^k - y^k \rangle \geq 0$. Thus,

$$\psi_{k+1}(v^{k+1}) \geq \min_{x \in \mathbb{R}^n} \left\{ A_k f(y^k) + \frac{\tau_k}{2} \|x - v^k\|^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|^2\} \right\}.$$

The explicit solution to this quadratic minimization problem is

$$x = \frac{1}{\tau_{k+1}} (\tau_k v^k + \mu a_{k+1} y^k - a_{k+1} \nabla f(y^k)).$$

By plugging in the solution and using $\langle \nabla f(y^k), v^k - y^k \rangle \geq 0$, we obtain

$$\psi_{k+1}(v^{k+1}) \geq A_{k+1}f(y^k) - \frac{a_{k+1}^2}{2\tau_{k+1}}\|\nabla f(y^k)\|_2^2 + \frac{\mu\tau_k a_{k+1}}{2\tau_{k+1}}\|v^k - y^k\|^2.$$

Our next goal is to show that

$$A_{k+1}f(y^k) - \frac{a_{k+1}^2}{2\tau_{k+1}}\|\nabla f(y^k)\|_2^2 + \frac{\mu\tau_k a_{k+1}}{2\tau_{k+1}}\|v^k - y^k\|^2 \geq A_{k+1}f(x^{k+1}), \quad (21)$$

which proves the induction step.

For option a) in the step 4, (11) takes the form

$$f(x^{k+1}) \leq f(y^k) - \frac{1}{2L}\|\nabla f(y^k)\|^2. \quad (22)$$

Since, for this option, $\frac{a_{k+1}^2}{(\tau_k + \mu a_{k+1})(A_k + a_{k+1})} = \frac{1}{L}$, inequality (21) holds. For option b) in the step 4, (21) holds by the choice of a_{k+1} from the equation

$$f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}\tau_{k+1}}\|\nabla f(y^k)\|_2^2 + \frac{\mu\tau_k a_{k+1}}{2A_{k+1}\tau_{k+1}}\|v^k - y^k\|^2 = f(x^{k+1}). \quad (23)$$

It remains to show that this equation has a solution $a_{k+1} > 0$. Since $A_{k+1} = A_k + a_{k+1}$ and $\tau_{k+1} = \tau_k + \mu a_{k+1}$, we can rewrite the equation (23) as

$$(2\mu\delta_k + \|\nabla f(y^k)\|_2^2)a_{k+1}^2 + (2\delta_k(\tau_k + \mu A_k) - \mu\tau_k\|v^k - y^k\|_2^2)a_{k+1} + 2\tau_k A_k \delta_k = 0,$$

where $\delta_k = f(x^{k+1}) - f(y^k) < 0$. By strong convexity, $f(y^k) - f(x_*) \leq \frac{1}{2\mu}\|\nabla f(y^k)\|_2^2$, we have

$$2\mu\delta_k + \|\nabla f(y^k)\|_2^2 \geq 2\mu(f(x^{k+1}) - f(x_*)) \geq 0.$$

Therefore, a non-negative solution exists and may be written down as

$$a_{k+1} = \frac{-S_k + \sqrt{S_k^2 - 8A_k\delta_k\tau_k(2\delta_k\mu + \|\nabla f(y_k)\|^2)}}{4\delta_k\mu + 2\|\nabla f(y_k)\|^2},$$

where $S_k = 2\delta_k(\tau_k + \mu A_k) - \mu\tau_k\|v^k - y^k\|^2$

Let us estimate the rate of the growth for A_k . If in the step 4 option a) is used, $\frac{a_{k+1}^2}{\tau_{k+1}A_{k+1}} = \frac{1}{L}$. For the option b), using (23) and (22), we have

$$f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}\tau_{k+1}}\|\nabla f(y^k)\|_2^2 + \frac{\mu\tau_k a_{k+1}}{2A_{k+1}\tau_{k+1}}\|v^k - y^k\|^2 \leq f(y^k) - \frac{1}{2L}\|\nabla f(y^k)\|_2^2.$$

Thus, for both options, $\frac{a_{k+1}^2}{\tau_{k+1}A_{k+1}} \geq \frac{1}{L}$, or

$$a_i \geq \frac{1}{\sqrt{L}}\sqrt{A_i + \mu A_i^2} \geq \sqrt{\frac{\mu}{L}}A_i.$$

Using the left inequality, we obtain

$$\sqrt{A_i} - \sqrt{A_{i-1}} \geq \frac{A_i - A_{i-1}}{\sqrt{A_i} + \sqrt{A_{i-1}}} \geq \frac{a_i}{2\sqrt{A_i}} \geq \frac{1}{2\sqrt{L}} \sqrt{1 + \mu A_i}. \quad (24)$$

This in turn implies a weaker inequality

$$\sqrt{A_i} - \sqrt{A_{i-1}} \geq \frac{1}{2\sqrt{L}}.$$

Summing it up for $i = 1, \dots, k$ we get

$$A_k \geq \frac{k^2}{4L}.$$

We also have

$$A_{k+1} = A_k + a_{k+1} \geq A_k + \sqrt{\frac{\mu}{L}} A_{k+1},$$

which leads to

$$A_{k+1} \geq \left(1 - \sqrt{\frac{\mu}{L}}\right)^{-1} A_k.$$

To use this bound we only need to estimate A_1 , which we can do as follows:

$$A_1 = \frac{a_1^2}{A_1} \geq \frac{a_1^2}{(1 + \mu A_1)A_1} = \frac{a_1^2}{\tau_1 A_1} \geq \frac{1}{L}$$

By recursively applying the last bound we reach the desired result:

$$A_k \geq \max \left\{ \frac{k^2}{4L}, \frac{1}{L} \left(1 - \sqrt{\frac{\mu}{L}}\right)^{-(k-1)} \right\}$$

□

Theorem 6. *Let function f be μ -strongly convex and L -smooth and Algorithm 1 be run for N steps. Then*

$$f(x_k) - f(x_*) \leq \min \left\{ \frac{2LR^2}{k^2}, \left(1 - \sqrt{\frac{\mu}{L}}\right)^{-(k-1)} LR^2 \right\},$$

where $R = \|x_0 - x_*\|$

Proof. According to the definition of μ -strong convexity

$$l_k(x_*) = \sum_{i=0}^k a_{i+1} \{f(y^i) + \langle \nabla f(y^i), x_* - y^i \rangle + \frac{\mu}{2} \|x_* - y^i\|_2^2\} \leq \sum_{i=0}^k a_{i+1} f(x_*) = A_{k+1} f(x_*).$$

From this inequality and **Theorem 5** we have

$$A_k f(x_k) \leq \min_{x \in \mathbb{R}^n} \psi_k(x) \leq \psi_k(x_*) = l_{k-1}(x_*) + \frac{1}{2} \|x_0 - x_*\|_2^2 \leq A_k f(x_*) + \frac{1}{2} \|x_0 - x_*\|_2^2.$$

Finally, denoting $R = \|x_0 - x_*\|$, we have

$$f(x_k) - f(x_*) \leq \min \left\{ \frac{2LR^2}{k^2}, \left(1 - \sqrt{\frac{\mu}{L}}\right)^{(k-1)} LR^2 \right\}.$$

□

Another way to apply the algorithm to strongly convex objective is to use a restart procedure.

Of course, we have no direct way to check the inequality in step 9 of the algorithm. However, using strong convexity, we have that $\frac{\mu}{2} \|x^k - x_*\|_2^2 \leq f(x^k) - f(x_*) \leq \frac{R^2}{2A_k}$. Provided μ is known, it is sufficient to check whether the r.h.s. is smaller than $\frac{\mu}{4} R^2$, which would imply $\|x_i^k - x_*\|_2^2 \leq \frac{1}{2} \|x_i^0 - x_*\|_2^2$.

Theorem 7. *If $f(x)$ is a μ -strongly convex and L -smooth function, then*

$$\|\tilde{x}^N - x_*\|_2^2 = O(2^{-\sqrt{\mu N^2/L}} R^2),$$

$$f(\tilde{x}^N) - f(x_*) = O(2^{-\sqrt{\mu N^2/L}} LR^2).$$

where $R = \|x_0 - x_*\|$.

Proof. From the very definition of strong convexity we have

$$\frac{\mu}{2} \|x_i^k - x_*\|_2^2 \leq f(x_i^k) - f(x_*).$$

From **Theorem 3** we have that

$$f(x_i^k) - f(x_*) \leq \frac{2L \|x_i^0 - x_*\|_2^2}{k^2}.$$

To ensure $\|x_i^k - x_*\|_2^2 \leq \frac{1}{2} \|x_i^0 - x_*\|_2^2$ we then need to satisfy the following inequality:

$$\frac{4L \|x_i^0 - x_*\|_2^2}{\mu k^2} \leq \frac{1}{2} \|x_i^0 - x_*\|_2^2.$$

That means that no more than $N_i = \lceil \sqrt{8L/\mu} \rceil$ iterations happen between the i -th and the $i+1$ -th restarts.

Algorithm 4 Accelerated Gradient Method with Small-Dimensional Relaxation (AGMsDR)

Output: x_i^k

- 1: **for** $i \geq 0$ **do**
- 2: Set $k = 0$, $A_0 = 0$, $x_i^0 = v_i^0$, $\psi_0^i(x) = \frac{1}{2}\|x - x_0\|_2^2$
- 3: **for** $k \geq 0$ **do**
- 4:

$$\beta_k = \arg \min_{\beta \in [0,1]} f\left(v_i^k + \beta(x_i^k - v_i^k)\right), \quad y_i^k = v_i^k + \beta_k(x_i^k - v_i^k). \quad (25)$$

- 5: Option a), L is known,

$$x_i^{k+1} = \arg \min_{x \in E} \left\{ f(y_i^k) + \langle \nabla f(y_i^k), x - y_i^k \rangle + \frac{L}{2} \|x - y_i^k\|^2 \right\}. \quad (26)$$

Find a_{k+1} from equation $\frac{a_{k+1}^2}{A_k + a_{k+1}} = \frac{1}{L}$.

Option b),

$$h_{k+1} = \arg \min_{h \geq 0} f\left(y_i^k - h(\nabla f(y_i^k))^\# \right), \quad x_i^{k+1} = y_i^k - h_{k+1}(\nabla f(y_i^k))^\#. \quad (27)$$

Find a_{k+1} from equation $f(y_i^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})} \|\nabla f(y_i^k)\|_*^2 = f(x_i^{k+1})$.

- 6: Set $A_{k+1} = A_k + a_{k+1}$.
 - 7: Set $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\{f(y_i^k) + \langle \nabla f(y_i^k), x - y_i^k \rangle\}$.
 - 8: $v_i^{k+1} = \arg \min_{x \in E} \psi_{k+1}(x)$
 - 9: **if** $\|x_i^k - x_*\|_2^2 \leq \frac{1}{2}\|x_i^0 - x_*\|_2^2$ **then**
 - 10: **break**
 - 11: **end if**
 - 12: $k = k + 1$
 - 13: **end for**
 - 14: Set $x_{i+1}^0 = x_i^N$.
 - 15: $i = i + 1$
 - 16: **end for**
-

Hence,

$$\|\tilde{x}^N - x_*\|_2^2 = O(2^{-\sqrt{\mu N^2/L}} R^2),$$

$$f(\tilde{x}^N) - f(x_*) = O(2^{-\sqrt{\mu N^2/L}} L R^2).$$

□

2.3. Implementation details: line-search accuracy

In our methods the line search step is used to perform linear coupling and steepest descent. It will now be shown that in both cases performing the line search exactly is not critical for the methods' convergence.

- **Steepest descent.** In algorithms APDLSGD, UAPDLSGD and SCUAPDLSGD steepest descent is used to construct x^{k+1} . However, the convergence analysis of all the above methods only relies on $f(x^{k+1})$ being no greater than $f(y^k - \frac{1}{l}\nabla f(y^k))$, where $l = L$ for the APDLSGD method and $l = M(\frac{a_{k+1}}{A_{k+1}}\varepsilon, \nu, M_\nu)$ for the universal methods. This means that the accuracy of line search has no effect on the worst-case convergence bounds, as long as it is good enough to ensure that the result is no worse than one obtained by performing a gradient descent step. Since for most objectives using exact steepest descent should result in iterates different from ones obtained by gradient descent, it is reasonable to expect that performing steepest descent with some small error will still lead to iterates with low enough objective values.
- **Linear coupling.** The fact that the step

$$\beta_k = \arg \min_{\beta \in [0,1]} f\left(v^k + \beta(x^k - v^k)\right); \quad y^k = v^k + \beta_k(x^k - v^k)$$

guarantees $f(y^k) \leq f(x^k)$ and $\langle \nabla f(y^k), v^k - y^k \rangle \geq 0$ is used in the convergence analysis. Again, it is reasonable to expect the first inequality to hold true in the case of inexact line search. We will now show that allowing for some error in the second inequality does not lead to accumulating errors.

Lemma 2. *For the not necessarily convex problem (2) and the APDGD method with step 3 performed in a way that guarantees $f(y^k) \leq f(x^k)$ and $\langle \nabla f(y^k), v^k - y^k \rangle \geq -\tilde{\varepsilon}$,*

$$A_{k+1}f(x^{k+1}) \leq \min_{x \in \mathbb{R}^n} \psi_{k+1}(x) + A_{k+1}\tilde{\varepsilon} = \psi_{k+1}(v^{k+1}) + A_{k+1}\tilde{\varepsilon}$$

and

$$A_k = O\left(\frac{k^2}{L}\right).$$

Proof. Theorem can be prove by induction. Let's consider the step of induction. That is, assume that we've already proved that

$$A_k f(x^k) \leq \min_{x \in \mathbb{R}^n} \psi_k(x) + A_k \tilde{\varepsilon} = \psi_k(v^k) + A_k \tilde{\varepsilon}.$$

Then

$$\begin{aligned} \psi_{k+1}(v^{k+1}) &= \min_{x \in \mathbb{R}^n} \left\{ \psi_k(x) + a_{k+1} \{ f(y^k) + \langle \nabla f(y^k), x - y^k \rangle \} \right\} = \\ &= \min_{x \in \mathbb{R}^n} \left\{ \psi_k(v^k) + \frac{1}{2} \|x - v^k\|_2^2 + a_{k+1} \{ f(y^k) + \langle \nabla f(y^k), x - y^k \rangle \} \right\} \geq \end{aligned}$$

$$\geq \min_{x \in \mathbb{R}^n} \left\{ A_k f(x^k) + \frac{1}{2} \|x - v^k\|_2^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\} - A_k \tilde{\varepsilon} \right\}.$$

Due to the line 3 $f(y^k) \leq f(x^k)$ and $\langle \nabla f(y^k), v^k - y^k \rangle \geq -\tilde{\varepsilon}$. From this two inequalities and the fact that $A_{k+1} = A_k + a_{k+1}$ one can obtain

$$\begin{aligned} \psi_{k+1}(v^{k+1}) &\geq \min_{x \in \mathbb{R}^n} \left\{ A_{k+1} f(y^k) + a_{k+1} \langle \nabla f(y^k), x - v^k \rangle + \frac{1}{2} \|x - v^k\|_2^2 - A_{k+1} \tilde{\varepsilon} \right\} = \\ &= A_{k+1} f(y^k) - \frac{a_{k+1}^2}{2} \|\nabla f(y^k)\|_2^2 - A_{k+1} \tilde{\varepsilon} \end{aligned}$$

So let's choose a_{k+1} in such a way that guarantee

$$A_{k+1} f(y^k) - \frac{a_{k+1}^2}{2} \|\nabla f(y^k)\|_2^2 = A_{k+1} f(x^{k+1}). \quad (28)$$

This is quadratic equation on a_{k+1} . One can solve it explicitly. For the method APDGD (see line 4) this equation means that $\frac{a_{k+1}^2}{A_{k+1}} \geq \frac{1}{L}$, which, combined with $A_{k+1} = A_k + a_{k+1}$, means that $a_k = O\left(\frac{k}{L}\right)$, $A_k = O\left(\frac{k^2}{L}\right)$. \square

Of course, the same result applies to all the other version of the method presented further. This lemma leads to an additive term $\tilde{\varepsilon}$ in the convergence bounds.

3. Universal methods

We consider the optimization problem (2). By considering the class of objectives with Hölder continuous (sub)gradients we may generalize the above methods to non-smooth problems.

In this section we assume that the objective function has Hölder continuous (sub)gradients: for all $x, y \in \mathbb{R}^n$ and some $\nu \in [0, 1]$

$$\|\nabla f(y) - \nabla f(x)\|_* \leq M_\nu \|x - y\|^\nu. \quad (29)$$

Here if $\nu = 0$ $\nabla f(x)$ denotes some subgradient of $f(x)$.

Again, we will be using the sequence of estimating functions defined as

$$\psi_0(x) = V(x, x^0).$$

$$l_k(x) = \sum_{i=0}^k a_{i+1} \{f(y^i) + \langle \nabla f(y^i), x - y^i \rangle\},$$

$$\psi_{k+1}(x) = l_k(x) + \psi_0(x) = \psi_k(x) + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\},$$

$$A_{k+1} = A_k + a_{k+1}, \quad A_0 = 0.$$

Algorithm 5 Universal Accelerated Gradient Method with Small-Dimensional Relaxation (UAGMsDR)

Input: Accuracy ε

Output: x^k

- 1: Set $k = 0$, $A_0 = 0$, $x^0 = v^0$, $\psi_0(x) = V(x, x^0)$
- 2: **for** $k \geq 0$ **do**
- 3:

$$\beta_k = \arg \min_{\beta \in [0,1]} f\left(v^k + \beta(x^k - v^k)\right), \quad y^k = v^k + \beta_k(x^k - v^k). \quad (30)$$

4:

$$h_{k+1} = \arg \min_{h \geq 0} f\left(y^k - h(\nabla f(y^k))^\# \right), \quad x^{k+1} = y^k - h_{k+1}(\nabla f(y^k))^\#. \quad (31)$$

Find a_{k+1} from equation $f(y^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})} \|\nabla f(y^k)\|_*^2 + \frac{\varepsilon a_{k+1}}{2(A_k + a_{k+1})} = f(x^{k+1})$.

- 5: Set $A_{k+1} = A_k + a_{k+1}$.
 - 6: Set $\psi_{k+1}(x) = \psi_k(x) + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\}$.
 - 7: $v^{k+1} = \arg \min_{x \in E} \psi_{k+1}(x)$
 - 8: $k = k + 1$
 - 9: **end for**
-

In the analysis of the above method we will be using a particular choice of the subgradient in step 4. However, it seems that in practice this is not important for the method's convergence. All the results mentioned below remain correct if the line search domain $[0, 1]$ in line 3 is changed to any larger subset of \mathbb{R} . Note that unlike other universal methods, ([29, 14]) this method does not require estimating the step length in an inner cycle. This results in a slightly better complexity bound due to better step-lengths and lower iteration complexity.

The following lemma (the proof of which may be found in [29]) plays a major role in the convergence analysis of this method.

Lemma 3. *Let function $f(x)$ have Hölder continuous (sub)gradients for some $\nu \in [0, 1]$ and $M_\nu < +\infty$. Then for any $\delta > 0$ we have*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2} \|y - x\|^2 + \frac{\delta}{2},$$

where

$$M = M(\delta, \nu, M_\nu) = \left[\frac{1 - \nu}{1 + \nu} \frac{M_\nu}{\delta} \right]^{\frac{1-\nu}{1+\nu}} M_\nu.$$

If the subgradient is not Hölder continuous for some exponent ν it is convenient to consider the corresponding M_ν to be equal to $+\infty$.

Theorem 8. For the algorithm UAGMsDR and possibly non-convex (2), where $f(x)$ has Hölder continuous (sub)gradients,

$$A_k f(x^k) \leq \min_{x \in \mathbb{R}^n} \psi_k(x) + \frac{A_k \varepsilon}{2} = \psi_k(v^k) + \frac{A_k \varepsilon}{2}. \quad (32)$$

and

$$A_k \geq \sup_{\nu \in [0,1]} \left[\frac{1+\nu}{1-\nu} \right]^{\frac{1-\nu}{1+\nu}} \frac{k^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}$$

Proof. Denote

$$l_k(x) = \sum_{i=0}^k a_{i+1} \{f(y^i) + \langle \nabla f(y^i), x - y^i \rangle\}.$$

Then

$$\psi_{k+1}(x) = l_k(x) + \psi_0(x) = \psi_k(x) + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\}.$$

First, we prove inequality (32) by induction over k . For $k = 0$, the inequality holds. Assume that

$$A_k f(x^k) \leq \min_{x \in \mathbb{R}^n} \psi_k(x) + \frac{A_k \varepsilon}{2} = \psi_k(v^k) + \frac{A_k \varepsilon}{2}.$$

Then

$$\begin{aligned} \psi_{k+1}(v^{k+1}) &= \min_{x \in \mathbb{R}^n} \left\{ \psi_k(x) + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\} \right\} \\ &\geq \min_{x \in \mathbb{R}^n} \left\{ \psi_k(v^k) + \frac{1}{2} \|x - v^k\|^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\} \right\} \\ &\geq \min_{x \in \mathbb{R}^n} \left\{ A_k f(x^k) - \frac{A_k \varepsilon}{2} + \frac{1}{2} \|x - v^k\|^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\} \right\}. \end{aligned}$$

Here we used that ψ_k is a strongly convex function with minimum at v^k .

By the definition of β_k and y^k in (30), we have $f(y^k) \leq f(x^k)$. By the optimality conditions in (30), there exists such a subgradient $\nabla f(y^k)$ that either

- (1) $\beta_k = 0$, $\langle \nabla f(y^k), x^k - v^k \rangle \geq 0$, $y^k = v^k$;
- (2) $\beta_k \in (0, 1)$ and $\langle \nabla f(y^k), x^k - v^k \rangle = 0$, $y^k = v^k + \beta_k(x^k - v^k)$;
- (3) $\beta_k = 1$ and $\langle \nabla f(y^k), x^k - v^k \rangle \leq 0$, $y^k = x^k$.

In all three cases, $\langle \nabla f(y^k), v^k - y^k \rangle \geq 0$. Thus,

$$\psi_{k+1}(v^{k+1}) \geq \min_{x \in \mathbb{R}^n} \left\{ A_{k+1} f(y^k) - \frac{A_k \varepsilon}{2} + a_{k+1} \langle \nabla f(y^k), x - v^k \rangle + \frac{1}{2} \|x - v^k\|^2 \right\}$$

$$\geq A_{k+1}f(y^k) - \frac{A_k\varepsilon}{2} - \frac{a_{k+1}^2}{2}\|\nabla f(y^k)\|_*^2,$$

where we used that for any $g \in E^*$, $s \in E$, $\zeta \geq 0$, it holds that $\langle g, s \rangle + \frac{\zeta}{2}\|s\|^2 \geq -\frac{1}{2\zeta}\|g\|_*^2$.

The equation

$$A_{k+1}f(y^k) - \frac{a_{k+1}^2}{2}\|\nabla f(y^k)\|_*^2 \geq A_{k+1}f(x^{k+1}) - \frac{a_{k+1}\varepsilon}{2}, \quad (33)$$

holds by the choice of a_{k+1} from the equation

$$f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}}\|\nabla f(y^k)\|_*^2 + \frac{a_{k+1}\varepsilon}{2A_{k+1}} = f(x^{k+1}). \quad (34)$$

It remains to show that this equation has a solution $a_{k+1} > 0$. Applying **Lemma 3** with $\delta = \frac{a_{k+1}\varepsilon}{A_{k+1}}$, we have

$$f(x^{k+1}) = \min_{h \geq 0} f\left(y^k - h(\nabla f(y^k))^\# \right) \leq \quad (35)$$

$$\begin{aligned} &\leq \min_{h \geq 0} \left(f(y^k) - h\langle \nabla f(y^k), (\nabla f(y^k))^\# \rangle + \frac{Mh^2}{2}\|(\nabla f(y^k))^\#\|^2 + \frac{a_{k+1}\varepsilon}{2A_{k+1}} \right) \\ &= f(y^k) - \frac{1}{2M}\|\nabla f(y^k)\|_*^2 + \frac{a_{k+1}\varepsilon}{2A_{k+1}}, \end{aligned} \quad (36)$$

where $M = \left[\frac{1-\nu}{1+\nu} \frac{A_{k+1}M_\nu}{a_{k+1}\varepsilon} \right]^{\frac{1-\nu}{1+\nu}} M_\nu$, and we used that $\langle \nabla f(y^k), (\nabla f(y^k))^\# \rangle = \|\nabla f(y^k)\|_*^2$ and $\|(\nabla f(y^k))^\#\|^2 = 1$ by definition of the vector $(\nabla f(y^k))^\#$. Since $A_{k+1} = A_k + a_{k+1}$, we can rewrite the equation (34) as

$$\frac{a_{k+1}^2}{2}\|\nabla f(y^k)\|_*^2 + a_{k+1} \left(f(x^{k+1}) - f(y^k) - \frac{\varepsilon}{2} \right) + A_k(f(x^{k+1}) - f(y^k)) = 0.$$

Since, by (35), $f(x^{k+1}) - f(y^k) < \frac{a_{k+1}\varepsilon}{2A_{k+1}}$ (otherwise $\|\nabla f(y^k)\|_* = 0$ and y_k is a solution to the problem (2)), at least one solution exists, and the greater one is

$$a_{k+1} = \frac{-(f(x^{k+1}) - f(y^k) - \varepsilon/2) + \sqrt{(f(x^{k+1}) - f(y^k) - \varepsilon/2)^2 - 2A_k(f(x^{k+1}) - f(y^k))\|\nabla f(y_k)\|_*^2}}{\|\nabla f(y_k)\|_*^2}.$$

Let us estimate the rate of the growth for A_k . Using (34) and (35), we have

$$\frac{a_k^2}{A_k} \geq \frac{1}{M_\nu} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+\nu}} \left[\frac{a_k}{A_k} \right]^{\frac{1-\nu}{1+\nu}},$$

or

$$\frac{a_k^{\frac{1+3\nu}{1+\nu}}}{A_k^{\frac{2\nu}{1+\nu}}} \geq \frac{1}{M_\nu} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+\nu}}$$

Denote $\gamma = \frac{1+\nu}{1+3\nu} \geq \frac{1}{2}$. We have

$$\begin{aligned} (A_{k+1}^{1-\gamma} + A_k^{1-\gamma})(A_{k+1}^\gamma - A_k^\gamma) &= A_{k+1} - A_k + A_{k+1}^\gamma A_k^{1-\gamma} - A_k^\gamma A_{k+1}^{1-\gamma} = \\ &= A_{k+1} - A_k + (A_{k+1} A_k)^{1-\gamma} (A_{k+1}^{2\gamma-1} - A_k^{2\gamma-1}) \geq A_{k+1} - A_k. \end{aligned}$$

Since $A_{k+1} = A_k + a_{k+1}$,

$$A_{k+1}^\gamma - A_k^\gamma \geq \frac{A_{k+1} - A_k}{A_{k+1}^{1-\gamma} + A_k^{1-\gamma}} \geq \frac{a_{k+1}}{2A_{k+1}^{1-\gamma}} \geq \frac{1}{2M_\nu^{\frac{2}{1+3\nu}}} \left[\frac{1+\nu}{1-\nu} \varepsilon \right]^{\frac{1-\nu}{1+3\nu}} \quad (37)$$

Now we take a telescopic sum for $k = 0, \dots, N-1$ and get

$$A_N - A_0 = A_N \geq \left[\frac{1+\nu}{1-\nu} \right]^{\frac{1-\nu}{1+3\nu}} \frac{N^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+3\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+3\nu}}}. \quad (38)$$

To get the statement of the theorem it remains to notice that the algorithm is independent of the level of smoothness ν . Hence, if the objective has Hölder continuous gradient for multiple $\nu \in [0, 1]$, then A_k will grow according to the greatest lower bound. Thus, we have

$$A_k \geq \sup_{\nu \in [0,1]} \left[\frac{1+\nu}{1-\nu} \right]^{\frac{1-\nu}{1+3\nu}} \frac{k^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+3\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+3\nu}}}.$$

□

The convergence rate of the above algorithm is given by the following theorem.

Theorem 9. *If $f(x)$ is convex (or 1-weakly-quasi-convex) and has Hölder continuous (sub)gradients, method*

UAGMsDR generates x_k such that

$$f(x_k) - f(x_*) \leq \frac{1}{A_k} V(x_*, x^0) + \frac{\varepsilon}{2}$$

An ε -accurate iterate x_T is obtained in the number of iterations

$$N \leq \inf_{\nu \in [0,1]} 2^{\frac{2+4\nu}{1+3\nu}} \left[\frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+3\nu}} \left[\frac{M_\nu}{\varepsilon} \right]^{\frac{2}{1+3\nu}} \Theta^{\frac{1+\nu}{1+3\nu}},$$

where $V(x_, x^0) \leq \Theta$*

Proof. According to the definition of 1-weak-quasi-convexity

$$l_k(x_*) = \sum_{i=0}^k a_{i+1} \{f(y^i) + \langle \nabla f(y^i), x_* - y^i \rangle\} \leq$$

$$\leq \sum_{i=0}^k a_{i+1} f(x_*) = A_{k+1} f(x_*).$$

By (31) and (30) we have $f(y^i) \leq f(x^i) \leq f(y^{i-1}) \leq \dots \leq f(x^0)$, so

$$l_k(x_*) \leq \sum_{i=0}^k a_{i+1} \{(1 - \gamma)f(x^0) + \gamma f(x_*)\}.$$

From this inequality and **Theorem 8** we have

$$\begin{aligned} A_k f(x_k) &\leq \min_{x \in \mathbb{R}^n} \psi_k(x) + \frac{A_k \varepsilon}{2} \leq \psi_k(x_*) + \frac{A_k \varepsilon}{2} = \\ &= l_{k-1}(x_*) + V(x_*, x^0) + \frac{A_k \varepsilon}{2} \leq A_k f(x_*) + V(x_*, x^0) + \frac{A_k \varepsilon}{2}. \end{aligned}$$

From here by rearranging the terms we obtain the first part of the statement of the theorem:

$$f(x_k) - f(x_*) \leq \frac{1}{A_k} V(x_*, x^0) + \frac{\varepsilon}{2}.$$

To get the second part we need to find N such that the following bound holds:

$$\frac{1}{A_T} V(x_*, x^0) + \frac{\varepsilon}{2} \leq \varepsilon.$$

We have that $\forall \nu \in [0, 1]$ (with possibly infinite M_ν)

$$A_T \geq \left[\frac{1 + \nu}{1 - \nu} \right]^{\frac{1-\nu}{1+\nu}} \frac{N^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}},$$

so it is sufficient to guarantee

$$\left[\frac{1 + \nu}{1 - \nu} \right]^{\frac{1-\nu}{1+\nu}} \frac{N^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}} \geq \frac{2V(x_*, x^0)}{\varepsilon},$$

so after

$$N \geq 2^{\frac{2+4\nu}{1+3\nu}} \left[\frac{1 - \nu}{1 + \nu} \right]^{\frac{1-\nu}{1+3\nu}} \left[\frac{M_\nu}{\varepsilon} \right]^{\frac{2}{1+3\nu}} \Theta^{\frac{1+\nu}{1+3\nu}}$$

iterations accuracy ε is guaranteed. It remains to see that infimum over ν may be taken in this upper bound, since the method does not have ν as a parameter. \square

This method also converges to a stationary point for non-convex objectives.

Theorem 10. Let function f be L -smooth and algorithm UAGMsDR be run for N steps. Then

$$\min_{k=0,\dots,N-1} \|\nabla f(y^k)\|_2^2 \leq \frac{2L(f(x^0) - f(x_*))}{N} + L\varepsilon.$$

Proof. We have that

$$f(x^{k+1}) \leq f(y^k) - \frac{1}{2L} \|\nabla f(y^k)\|_*^2 + \frac{a_{k+1}\varepsilon}{2A_{k+1}} \leq f(x^k) - \frac{1}{2L} \|\nabla f(y^k)\|_*^2 + \frac{a_{k+1}\varepsilon}{2A_{k+1}}. \quad (39)$$

Summing this up for $k = 0, \dots, N$, we obtain

$$f(x^0) - f(x_*) \geq f(x^0) - f(x^{N+1}) \geq \frac{N}{2L} \min_{k=0,\dots,N} \|\nabla f(y^k)\|_*^2 - \sum_{k=0}^{N-1} \frac{a_{k+1}}{A_{k+1}} \frac{\varepsilon}{2} \geq \frac{N}{2L} \min_{k=0,\dots,N} \|\nabla f(y^k)\|_*^2 - \frac{N\varepsilon}{2}.$$

Consequently, we may guarantee

$$\min_{k=0,\dots,N-1} \|\nabla f(y^k)\|_2^2 \leq \frac{2L(f(x^0) - f(x_*))}{N} + L\varepsilon.$$

□

3.0.1. Online stopping criterion

This method also has an efficient stopping criterion, provided the objective is convex.

By rewriting the statement of **Theorem 9** we see that

$$f(x^k) \leq \frac{\varepsilon}{2} + \frac{1}{A_k} \psi_k(v^k) = \frac{\varepsilon}{2} + \frac{1}{A_k} \min_{x \in \mathbb{R}^n} \left[\frac{1}{2} \|x^0 - x\|_2^2 + \sum_{i=0}^{k-1} a_{i+1} \{f(y^i) + \langle \nabla f(y^i), x - y^i \rangle\} \right]$$

As it was before,

$$l^{k-1}(x) = \sum_{i=0}^{k-1} a_{i+1} \left\{ f(y^i) + \langle \nabla f(y^i), x - y^i \rangle + \frac{\mu}{2} \|x - y^i\|_2^2 \right\}$$

Denote $R = \|x^0 - x_*\|_2$ and

$$\hat{f}^k = \min_{x: \|x - x_0\| \leq R} \frac{1}{A_k} l^{k-1}(x).$$

The constraint may be rewritten equivalently as $\frac{1}{2} \|x - x_0\|_2^2 \leq \frac{R^2}{2}$. By strong duality we see that

$$\begin{aligned} \hat{f}^k &= \min_{x \in \mathbb{R}^n} \max_{\lambda \geq 0} \left\{ \frac{1}{A_k} l^{k-1}(x) + \lambda \left(\frac{1}{2} \|x_0 - x\|_2^2 - \frac{R^2}{2} \right) \right\} \\ &= \max_{\lambda \geq 0} \min_{x \in \mathbb{R}^n} \left\{ \frac{1}{A_k} l^{k-1}(x) + \lambda \left(\frac{1}{2} \|x_0 - x\|_2^2 - \frac{R^2}{2} \right) \right\}. \end{aligned}$$

Now we set $\lambda = \frac{1}{A_k}$ and obtain

$$\hat{f}^k \geq \frac{1}{A_k} \psi_k(v^k) - \frac{R^2}{2A_k}.$$

Then

$$f(x^k) - \hat{f}^k \leq \frac{R^2}{2A_k} + \frac{\varepsilon}{2}.$$

But by convexity of $f(x)$ we have that $\forall k \frac{1}{A_k} l^{k-1}(x) \leq f(x)$, which implies that $\hat{f}^k \leq f(x_*)$. Finally, we have

$$f(x^k) - f(x_*) \leq f(x^k) - \hat{f}^k \leq \frac{R^2}{2A_k} + \frac{\varepsilon}{2},$$

so the condition $f(x^k) - \hat{f}^k \leq \frac{\varepsilon}{2}$ implies $f(x^k) - f(x_*) \leq \varepsilon$ and is an efficient stopping criterion for the UAGMsDR method.

3.1. Non-smooth strongly convex objectives

It remains to combine the two ideas used previously into a method for the case of non-smooth strongly convex objective. However, while there exist functions which are globally both L -smooth and μ -strongly convex, this is not the case for objectives with Hölder continuous gradients. Indeed, no function satisfies

$$f(x) + \langle f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 \leq f(y) \leq f(x) + \langle f(x), y - x \rangle + \frac{M_\nu}{1 + \nu} \|y - x\|_2^{1+\nu}$$

for all $x, y \in \mathbb{R}^n$. However, we have already established that our method converges monotonously. Since strongly convex functions have bounded sublevel sets, we only need this pair of inequalities to hold true for $\forall x, y \in \mathcal{L}_f(f(x^0)) = \{x | f(x) \leq f(x^0)\}$.

Theorem 11. *For the algorithm UAGMsDR and μ -strongly convex $f(x)$ with Hölder continuous (sub)gradients,*

$$A_k f(x^k) \leq \min_{x \in \mathbb{R}^n} \psi_k(x) + \frac{A_k \varepsilon}{2} = \psi_k(v^k) + \frac{A_k \varepsilon}{2} \quad (42)$$

and

$$A_k \geq \max \left\{ \left[\frac{1 + \nu}{1 - \nu} \right]^{\frac{1-\nu}{1+\nu}} \frac{k^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}, \frac{1}{M_\nu} \left[\frac{1 + \nu}{1 - \nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+\nu}} \left(1 - M_\nu^{-\frac{1+\nu}{1+3\nu}} \left[\frac{1 + \nu}{1 - \nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+3\nu}} \mu^{\frac{1+\nu}{1+3\nu}} \right)^{-(k-1)} \right\}. \quad (43)$$

Algorithm 6 Universal Accelerated Gradient Method with Small-Dimensional Relaxation (UAGMsDR)

Input: Accuracy ε

Output: x^k

1: Set $k = 0$, $A_0 = 0$, $x^0 = v^0$, $\psi_0(x) = \frac{1}{2}\|x_0 - x_*\|_2^2$

2: **for** $k \geq 0$ **do**

3:

$$\beta_k = \arg \min_{\beta \in [0,1]} f\left(v^k + \beta(x^k - v^k)\right), \quad y^k = v^k + \beta_k(x^k - v^k). \quad (40)$$

4:

$$h_{k+1} = \arg \min_{h \geq 0} f\left(y^k - h(\nabla f(y^k))^\#\right), \quad x^{k+1} = y^k - h_{k+1}(\nabla f(y^k))^\#. \quad (41)$$

Find a_{k+1} from equation $f(y^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})} \|\nabla f(y^k)\|_2^2 + \frac{\mu\tau_k a_{k+1}}{2(\tau_k + \mu a_{k+1})(A_k + a_{k+1})} \|v^k - y^k\|_2^2 + \frac{\varepsilon a_{k+1}}{2(A_k + a_{k+1})} = f(x^{k+1})$.

5: Set $A_{k+1} = A_k + a_{k+1}$.

6: Set $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2}\|x - y^k\|_2^2\}$.

7: $v^{k+1} = \arg \min_{x \in E} \psi_{k+1}(x)$

8: $k = k + 1$

9: **end for**

Furthermore, an ε -accurate iterate x_T is obtained in the number of iterations

$$N \leq \inf_{\nu \in [0,1]} \min \left\{ 2 \left[\frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+3\nu}} \left[\frac{M_\nu}{\varepsilon} \right]^{\frac{2}{1+3\nu}} R^{\frac{2+2\nu}{1+3\nu}}, \frac{M_\nu^{\frac{2}{1+3\nu}}}{\varepsilon^{\frac{1-\nu}{1+3\nu}} \mu^{\frac{1+\nu}{1+3\nu}}} \ln \left(\left[\frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}} \frac{M_\nu^{\frac{2}{1+\nu}} R^2}{\varepsilon^{\frac{2}{1+\nu}}} \right) \right\}, \quad (44)$$

where $\|x^0 - x_*\| \leq R$.

Proof. Same as before, denote

$$l_k(x) = \sum_{i=0}^k a_{i+1} \{f(y^i) + \langle \nabla f(y^i), x - y^i \rangle + \frac{\mu}{2}\|x - y^i\|_2^2\}.$$

First, we prove inequality (42) by induction over k . For $k = 0$, the inequality holds. Assume that

$$A_k f(x^k) \leq \min_{x \in \mathbb{R}^n} \psi_k(x) + \frac{A_k \varepsilon}{2} = \psi_k(v^k) + \frac{A_k \varepsilon}{2}.$$

Then

$$\psi_{k+1}(v^{k+1}) = \min_{x \in \mathbb{R}^n} \left\{ \psi_k(x) + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2}\|x - y^k\|_2^2\} \right\}$$

$$\begin{aligned}
&\geq \min_{x \in \mathbb{R}^n} \left\{ \psi_k(v^k) + \frac{\tau_k}{2} \|x - v^k\|^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|^2\} \right\} \\
&\geq \min_{x \in \mathbb{R}^n} \left\{ A_k f(x^k) - \frac{A_k \varepsilon}{2} + \frac{\tau_k}{2} \|x - v^k\|^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|^2\} \right\}.
\end{aligned}$$

Here we used that ψ_k is a τ_k -strongly convex function with minimum at v^k .

By the definition of β_k and y_k in (17), we have $f(y^k) \leq f(x^k)$. By the optimality conditions in (17), there exists such a subgradient $\nabla f(y^k)$ that either

- (1) $\beta_k = 0$, $\langle \nabla f(y^k), x^k - v^k \rangle \geq 0$, $y^k = v^k$;
- (2) $\beta_k \in (0, 1)$ and $\langle \nabla f(y^k), x^k - v^k \rangle = 0$, $y^k = v^k + \beta_k(x^k - v^k)$;
- (3) $\beta_k = 1$ and $\langle \nabla f(y^k), x^k - v^k \rangle \leq 0$, $y^k = x^k$.

In all three cases, $\langle \nabla f(y^k), v^k - y^k \rangle \geq 0$. Thus,

$$\psi_{k+1}(v^{k+1}) \geq \min_{x \in \mathbb{R}^n} \left\{ A_k f(y^k) - \frac{A_k \varepsilon}{2} + \frac{\tau_k}{2} \|x - v^k\|^2 + a_{k+1} \{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|^2\} \right\}.$$

The explicit solution to this quadratic minimization problem is

$$x = \frac{1}{\tau_{k+1}} (\tau_k v^k + \mu a_{k+1} y^k - a_{k+1} \nabla f(y^k)).$$

By plugging in the solution and using $\langle \nabla f(y^k), v^k - y^k \rangle \geq 0$, we obtain

$$\psi_{k+1}(v^{k+1}) \geq A_{k+1} f(y^k) - \frac{A_k \varepsilon}{2} - \frac{a_{k+1}^2}{2\tau_{k+1}} \|\nabla f(y^k)\|_2^2 + \frac{\mu \tau_k a_{k+1}}{2\tau_{k+1}} \|v^k - y^k\|^2.$$

Our next goal is to show that

$$A_{k+1} f(y^k) - \frac{A_k \varepsilon}{2} - \frac{a_{k+1}^2}{2\tau_{k+1}} \|\nabla f(y^k)\|_2^2 + \frac{\mu \tau_k a_{k+1}}{2\tau_{k+1}} \|v^k - y^k\|^2 \geq A_{k+1} f(x^{k+1}) - \frac{A_{k+1} \varepsilon}{2}, \quad (45)$$

which proves the induction step. But this is guaranteed by the choice of a_{k+1} in the step 4 of the method as the solution of the equation

$$f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}\tau_{k+1}} \|\nabla f(y^k)\|_2^2 + \frac{\mu \tau_k a_{k+1}}{2A_{k+1}\tau_{k+1}} \|v^k - y^k\|^2 = f(x^{k+1}) - \frac{a_{k+1} \varepsilon}{2A_{k+1}}. \quad (46)$$

It remains to show that this equation has a solution $a_{k+1} > 0$. Again, this is a quadratic equation with the greatest solution given by

$$a_{k+1} = \frac{-S_{k,\varepsilon} + \sqrt{S_{k,\varepsilon}^2 - 8A_k \delta_k \tau_k (2\delta_{k,\varepsilon} \mu + \|\nabla f(y_k)\|_2^2)}}{4\delta_{k,\varepsilon} \mu + 2\|\nabla f(y_k)\|_2^2},$$

where $\delta_k = f(x^{k+1}) - f(y^k)$, $\delta_{k,\varepsilon} = \delta_k - \frac{\varepsilon}{2}$, $S_{k,\varepsilon} = 2\delta_{k,\varepsilon} \tau_k - 2\mu A_k \delta_k + \mu \tau_k \|v^k - y^k\|^2$

Note that if the objective is μ -strongly convex, it is true that $\forall x \in \mathbb{R}^n$ $f(x) - f(x_*) \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$. Hence,

$$2\delta_{k,\varepsilon}\mu + \|\nabla f(y_k)\|^2 \geq -\frac{\varepsilon}{2} + f(x^{k+1}) - f(x_*).$$

This means that a_{k+1} may only be negative or undefined if $f(x^{k+1}) - f(x_*) \leq \frac{\varepsilon}{2}$, which means that $f(x^{k+1})$ is already an $\frac{\varepsilon}{2}$ -accurate solution to the problem.

Let us estimate the rate of the growth for A_k . By the same argument as the one used in the proof of **Theorem 8** we have

$$\frac{a_k^2}{A_k \tau_k} \geq \frac{1}{M_\nu} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+\nu}} \left[\frac{a_k}{A_k} \right]^{\frac{1-\nu}{1+\nu}}$$

Using that we obtain

$$\frac{a_k^{\frac{1+3\nu}{1+\nu}}}{A_k^{\frac{2\nu}{1+\nu}}} \geq \frac{1}{M_\nu} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+\nu}} (1 + \mu A_k) \geq \frac{1}{M_\nu} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+\nu}} \mu A_k,$$

or, if we denote $\gamma = \frac{1+\nu}{1+3\nu}$,

$$a_k \geq \frac{1}{M_\nu^\gamma} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+3\nu}} \mu^\gamma A_k.$$

To get the left term in the stated lower bound on A_k we write

$$A_i^\gamma - A_{i-1}^\gamma \geq \frac{A_i - A_{i-1}}{A_i^{1-\gamma} + A_{i-1}^{1-\gamma}} \geq \frac{a_i}{2A_i^{1-\gamma}} \geq \frac{1}{2M_\nu^{\frac{2}{1+3\nu}}} \left[\frac{1+\nu}{1-\nu} \varepsilon \right]^{\frac{1-\nu}{1+3\nu}} (1 + \mu A_i)^\gamma. \quad (47)$$

From this we obtain a weaker inequality

$$A_i^\gamma - A_{i-1}^\gamma \geq \frac{1}{2M_\nu^{\frac{2}{1+3\nu}}} \left[\frac{1+\nu}{1-\nu} \varepsilon \right]^{\frac{1-\nu}{1+3\nu}}.$$

Now we telescope it for $i = 0, \dots, k$ and get

$$A_k - A_0 = A_k \geq \left[\frac{1+\nu}{1-\nu} \right]^{\frac{1-\nu}{1+\nu}} \frac{k^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}. \quad (48)$$

To get the right term we observe that

$$A_{k+1} = A_k + a_{k+1} \geq A_k + \frac{1}{M_\nu^\gamma} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+3\nu}} \mu^\gamma A_{k+1},$$

which leads to

$$A_{k+1} \geq \left(1 - \frac{1}{M_\nu^\gamma} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+3\nu}} \mu^\gamma \right)^{-1} A_k.$$

To use this bound we only need to estimate A_1 , which we can do as follows:

$$A_1 = \frac{a_1^2}{A_1} \geq \frac{a_1^2}{(1+\mu A_1)A_1} = \frac{a_1^2}{\tau_1 A_1} \geq \frac{1}{M_\nu} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+\nu}}$$

By recursively applying the last bound we reach the desired result:

$$A_k \geq \frac{1}{M_\nu} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+\nu}} \left(1 - \frac{1}{M_\nu^\gamma} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+3\nu}} \mu^\gamma \right)^{-(k-1)}$$

By combining both bounds we get the statement of the theorem:

$$A_k \geq \max \left\{ \left[\frac{1+\nu}{1-\nu} \right]^{\frac{1-\nu}{1+\nu}} \frac{k^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}, \frac{1}{M_\nu} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+\nu}} \left(1 - M_\nu^{-\frac{1+\nu}{1+3\nu}} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+3\nu}} \mu^{\frac{1+\nu}{1+3\nu}} \right)^{-(k-1)} \right\}$$

It has already been established in **Theorem 6** that

$$\min_{x \in \mathbb{R}^n} \psi_k(x) \leq \psi_k(x_*) = l_{k-1}(x_*) + \frac{1}{2} \|x_0 - x_*\|_2^2 \leq A_k f(x_*) + \frac{1}{2} \|x_0 - x_*\|_2^2.$$

In conjunction with (42) this gives us

$$f(x^N) - f(x_*) \leq \frac{R^2}{2A_T} + \frac{\varepsilon}{2}.$$

The first term in (44) has already been established previously. It remains to analyze rate of convergence if the maximum in (43) is achieved on the second ter,. We need to satisfy

$$\frac{R^2}{2A_T} + \frac{\varepsilon}{2} \leq \varepsilon.$$

$$M_\nu R^2 \left[\frac{1-\nu}{1+\nu} \frac{M_\nu}{\varepsilon} \right]^{\frac{1-\nu}{1+\nu}} \left(1 - M_\nu^{-\frac{1+\nu}{1+3\nu}} \mu^{\frac{1+\nu}{1+3\nu}} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+3\nu}} \right)^{N+1} \leq \varepsilon$$

Taking the natural logarithm of both sides, we obtain

$$(N+1) \ln \left(1 - M_\nu^{-\frac{1+\nu}{1+3\nu}} \mu^{\frac{1+\nu}{1+3\nu}} \left[\frac{1+\nu}{1-\nu} \frac{\varepsilon}{M_\nu} \right]^{\frac{1-\nu}{1+3\nu}} \right) \leq \ln \left(\frac{\varepsilon^{\frac{2}{1+\nu}}}{M_\nu^{\frac{2}{1+\nu}} R^2} \left[\frac{1+\nu}{1-\nu} \right]^{\frac{1-\nu}{1+\nu}} \right).$$

We now use $\ln(1-x) \leq -x$ and $N < N+1$ to get the final result: with such A_k $f(x^N) - f(x_*) \leq \varepsilon$ if

$$N \geq \frac{M_\nu^{\frac{2}{1+3\nu}}}{\varepsilon^{\frac{1-\nu}{1+3\nu}} \mu^{\frac{1+\nu}{1+3\nu}}} \ln \left(\frac{\varepsilon^{\frac{2}{1+\nu}}}{M_\nu^{\frac{2}{1+\nu}} R^2} \left[\frac{1+\nu}{1-\nu} \right]^{\frac{1-\nu}{1+\nu}} \right).$$

□

Note that if $\nu = 0$ we have

$$N \leq \frac{M_0^2}{\varepsilon \mu} \ln \left(\frac{M_0^2 R^2}{\varepsilon^2} \right),$$

which is optimal up to a logarithmic factor [21].

4. Application to problems with linear constraints

In this section, we consider a minimization problem with linear equality. The idea is to construct the dual problem and solve it by our UAGMsDR method endowed with a step in the primal space. Following [10, 9], we show that this modification solves simultaneously both primal and dual problems and allows to obtain convergence rate.

Specifically, we consider the following minimization problem

$$(P_1) \quad \min_{x \in Q \subseteq E} \{f(x) : \mathbf{A}x = b\},$$

where E is a finite-dimensional real vector space, Q is a simple closed convex set, \mathbf{A} is given linear operator from E to some finite-dimensional real vector space H , $b \in H$ is given. The Lagrange dual problem to Problem (P_1) is

$$(D_1) \quad \max_{\lambda \in \Lambda} \left\{ -\langle \lambda, b \rangle + \min_{x \in Q} (f(x) + \langle \mathbf{A}^T \lambda, x \rangle) \right\}.$$

Here we denote $\Lambda = H^*$. It is convenient to rewrite Problem (D_1) in the equivalent form of a minimization problem

$$(P_2) \quad \min_{\lambda \in \Lambda} \left\{ \langle \lambda, b \rangle + \max_{x \in Q} (-f(x) - \langle \mathbf{A}^T \lambda, x \rangle) \right\}.$$

We denote

$$\varphi(\lambda) = \langle \lambda, b \rangle + \max_{x \in Q} (-f(x) - \langle \mathbf{A}^T \lambda, x \rangle). \quad (49)$$

Since f is convex, $\varphi(\lambda)$ is a convex function and, by Danskin's theorem, its subgradient is equal to (see e.g. [26])

$$\nabla\varphi(\lambda) = b - \mathbf{A}x(\lambda) \quad (50)$$

where $x(\lambda)$ is some solution of the convex problem

$$\max_{x \in Q} (-f(x) - \langle \mathbf{A}^T \lambda, x \rangle). \quad (51)$$

In what follows, we make the following assumptions about the dual problem (D_1)

- Subgradient of the objective function $\varphi(\lambda)$ satisfies Hölder condition (29) with constant M_ν .
- The dual problem (D_1) has a solution λ^* and there exist some $R > 0$ such that

$$\|\lambda^*\|_2 \leq R < +\infty. \quad (52)$$

It is worth noting that the quantity R will be used only in the convergence analysis, but not in the algorithm itself. As it was pointed in [31], the first assumption is reasonable. Namely, if the set Q is bounded, then $\nabla\varphi(\lambda)$ is bounded and (3) holds with $\nu = 0$. If $f(x)$ is uniformly convex, i.e., for all $x, y \in Q$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^\rho$, for some $\mu > 0$, $\rho \geq 2$, then $\nabla\varphi(\lambda)$ satisfies (3) with $\nu = \frac{1}{\rho-1}$, $M_\nu = \left(\frac{\|\mathbf{A}\|_{E \rightarrow H}^2}{\mu} \right)^{\frac{1}{\rho-1}}$. Here the norm of an operator $\mathbf{A} : E_1 \rightarrow E_2$ is defined as follows

$$\|\mathbf{A}\|_{E_1 \rightarrow E_2} = \max_{x \in E_1, u \in E_2^*} \{\langle u, \mathbf{A}x \rangle : \|x\|_{E_1} = 1, \|u\|_{E_2^*} = 1\}.$$

We choose Euclidean proximal setup, which means that we introduce Euclidean norm $\|\cdot\|_2$ in the space of vectors λ and choose the prox-function $d(\lambda) = \frac{1}{2}\|\lambda\|_2^2$. Then, we have $V[\zeta](\lambda) = \frac{1}{2}\|\lambda - \zeta\|_2^2$. Our primal-dual algorithm for Problem (P_1) is listed below as Algorithm 7.

Algorithm 7 PDUGDsDR

Input: starting point $\lambda_0 = 0$, accuracy $\tilde{\varepsilon}_f, \tilde{\varepsilon}_{eq} > 0$.

1: Set $k = 0$, $A_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.

2: **repeat**

3: $\beta_k = \arg \min_{\beta \in [0,1]} \varphi(\zeta^k + \beta(\eta^k - \zeta^k)); \lambda^k = \zeta^k + \beta_k(\eta^k - \zeta^k)$

4: $h_{k+1} = \arg \min_{h \geq 0} \varphi(\lambda^k - h \nabla \varphi(\lambda^k)); \eta^{k+1} = \lambda^k - h_{k+1} \nabla \varphi(\lambda^k)$ // Choose $\nabla \varphi(\lambda^k) : \langle \nabla \varphi(\lambda^k), \zeta^k - \lambda^k \rangle \geq 0$

5: Choose a_{k+1} from $\varphi(\eta^{k+1}) = \varphi(\lambda^k) - \frac{a_{k+1}^2}{2A_{k+1}} \|\nabla \varphi(\lambda^k)\|_2^2 + \frac{\varepsilon a_{k+1}}{2A_{k+1}}$ // $A_{k+1} = A_k + a_{k+1}$

6: $\zeta^{k+1} = \zeta^k - a_{k+1} \nabla \varphi(\lambda^k)$

7: Set

$$\hat{x}^{k+1} = \frac{1}{A_{k+1}} \sum_{i=0}^k a_{i+1} x(\lambda^i) = \frac{a_{k+1} x(\lambda^k) + A_k \hat{x}^k}{A_{k+1}}.$$

8: Set $k = k + 1$.

9: **until** $|f(\hat{x}^{k+1}) + \varphi(\eta^{k+1})| \leq \tilde{\varepsilon}_f$, $\|\mathbf{A}\hat{x}^{k+1} - b\|_2 \leq \tilde{\varepsilon}_{eq}$.

Output: The points $\hat{x}^{k+1}, \eta^{k+1}$.

Theorem 12. Let the objective φ in the problem (P_2) have Hölder-continuous sub-gradient and the solution of this problem be bounded, i.e. $\|\lambda^*\|_2 \leq R$. Then, for the sequence $\hat{x}^{k+1}, \eta^{k+1}$, $k \geq 0$, generated by Algorithm 7,

$$\|\mathbf{A}\hat{x}^k - b\|_2 \leq \frac{2R}{A_k} + \frac{\varepsilon}{2R}, \quad |\varphi(\eta^k) + f(\hat{x}^k)| \leq \frac{2R^2}{A_k} + \frac{\varepsilon}{2}, \quad (53)$$

where $A_k \geq \left[\frac{1+\nu}{1-\nu} \right]^{\frac{1-\nu}{1+\nu}} \frac{k^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}$.

Proof. The proof mostly follows the steps of our previous work [6], but we give the proof for the reader's convenience. The main difference is that here we use universal method. From Theorem 1, since $\zeta_0 = 0$, we have, for all $k \geq 0$,

$$A_k \varphi(\eta^k) \leq \min_{\lambda \in \Lambda} \left\{ \sum_{i=0}^{k-1} a_{i+1} (\varphi(\lambda^i) + \langle \nabla \varphi(\lambda^i), \lambda - \lambda^i \rangle) + \frac{1}{2} \|\lambda\|_2^2 \right\} + \frac{A_k \varepsilon}{2} \quad (54)$$

Let us introduce a set $\Lambda_R = \{\lambda : \|\lambda\|_2 \leq 2R\}$ where R is given in (52). Then, from (54), we obtain

$$\begin{aligned} A_k \varphi(\eta^k) &\leq \min_{\lambda \in \Lambda} \left\{ \sum_{i=0}^{k-1} a_{i+1} (\varphi(\lambda^i) + \langle \nabla \varphi(\lambda^i), \lambda - \lambda^i \rangle) + \frac{1}{2} \|\lambda\|_2^2 \right\} + \frac{A_k \varepsilon}{2} \\ &\leq \min_{\lambda \in \Lambda_R} \left\{ \sum_{i=0}^{k-1} a_{i+1} (\varphi(\lambda^i) + \langle \nabla \varphi(\lambda^i), \lambda - \lambda^i \rangle) + \frac{1}{2} \|\lambda\|_2^2 \right\} + \frac{A_k \varepsilon}{2} \\ &\leq \min_{\lambda \in \Lambda_R} \left\{ \sum_{i=0}^{k-1} a_{i+1} (\varphi(\lambda^i) + \langle \nabla \varphi(\lambda^i), \lambda - \lambda^i \rangle) \right\} + 2R^2 + \frac{A_k \varepsilon}{2}. \end{aligned} \quad (55)$$

On the other hand, from the definition (49) of $\varphi(\lambda)$, we have

$$\begin{aligned} \varphi(\lambda^i) &= \langle \lambda^i, b \rangle + \max_{x \in Q} (-f(x) - \langle \mathbf{A}^T \lambda^i, x \rangle) \\ &= \langle \lambda^i, b \rangle - f(x(\lambda^i)) - \langle \mathbf{A}^T \lambda^i, x(\lambda^i) \rangle. \end{aligned}$$

Combining this equality with (50), we obtain

$$\begin{aligned} \varphi(\lambda^i) - \langle \nabla \varphi(\lambda^i), \lambda^i \rangle &= \langle \lambda^i, b \rangle - f(x(\lambda^i)) - \langle \mathbf{A}^T \lambda^i, x(\lambda^i) \rangle \\ &\quad - \langle b - \mathbf{A}x(\lambda^i), \lambda^i \rangle = -f(x(\lambda^i)). \end{aligned}$$

Summing these equalities from $i = 0$ to $i = k - 1$ with the weights $\{a_{i+1}\}_{i=0, \dots, k-1}$, we get, using the convexity of f

$$\begin{aligned} \sum_{i=0}^{k-1} a_{i+1} (\varphi(\lambda^i) + \langle \nabla \varphi(\lambda^i), \lambda - \lambda^i \rangle) &= - \sum_{i=0}^{k-1} a_{i+1} f(x(\lambda^i)) + \sum_{i=0}^{k-1} a_{i+1} \langle (b - \mathbf{A}x(\lambda^i)), \lambda \rangle \\ &\leq -A_k f(\hat{x}^k) + A_k \langle b - \mathbf{A}\hat{x}^k, \lambda \rangle. \end{aligned}$$

Substituting this inequality to (55), we obtain

$$A_k \varphi(\eta^k) \leq -A_k f(\hat{x}^k) + A_k \min_{\lambda \in \Lambda_R} \left\{ \langle b - \mathbf{A}\hat{x}^k, \lambda \rangle \right\} + 2R^2 + \frac{A_k \varepsilon}{2}.$$

Finally, since $\max_{\lambda \in \Lambda_R} \left\{ \langle -b + \mathbf{A}\hat{x}^k, \lambda \rangle \right\} = 2R \|\mathbf{A}\hat{x}^k - b\|_2$, we obtain

$$\varphi(\eta^k) + f(\hat{x}^k) + 2R \|\mathbf{A}\hat{x}^k - b\|_2 \leq \frac{2R^2}{A_k} + \frac{\varepsilon}{2}. \quad (56)$$

Since λ^* is an optimal solution of Problem (D_1) , we have, for any $x \in Q$

$$\text{Opt}[P_1] \leq f(x) + \langle \lambda^*, \mathbf{A}x - b \rangle.$$

Using the assumption (52), we get

$$f(\hat{x}^k) \geq \text{Opt}[P_1] - R \|\mathbf{A}\hat{x}^k - b\|_2. \quad (57)$$

Hence,

$$\begin{aligned} \varphi(\eta^k) + f(\hat{x}^k) &= \varphi(\eta^k) - \text{Opt}[P_2] + \text{Opt}[P_2] + \text{Opt}[P_1] - \text{Opt}[P_1] + f(\hat{x}^k) \\ &= \varphi(\eta^k) - \text{Opt}[P_2] - \text{Opt}[D_1] + \text{Opt}[P_1] - \text{Opt}[P_1] + f(\hat{x}^k) \\ &\geq -\text{Opt}[P_1] + f(\hat{x}^k) \stackrel{(57)}{\geq} -R \|\mathbf{A}\hat{x}^k - b\|_2. \end{aligned} \quad (58)$$

This and (56) give

$$R \|\mathbf{A}\hat{x}^k - b\|_2 \leq \frac{2R^2}{A_k} + \frac{\varepsilon}{2}. \quad (59)$$

Hence, we obtain

$$\varphi(\eta^k) + f(\hat{x}^k) \stackrel{(58), (59)}{\geq} -\frac{2R^2}{A_k} - \frac{\varepsilon}{2}. \quad (60)$$

On the other hand, we have

$$\varphi(\eta^k) + f(\hat{x}^k) \stackrel{(56)}{\leq} \frac{2R^2}{A_k} + \frac{\varepsilon}{2}. \quad (61)$$

Combining (59), (60), (61), we conclude

$$\begin{aligned} \|\mathbf{A}\hat{x}^k - b\|_2 &\leq \frac{2R}{A_k} + \frac{\varepsilon}{2R}, \\ |\varphi(\eta^k) + f(\hat{x}^k)| &\leq \frac{2R^2}{A_k} + \frac{\varepsilon}{2}. \end{aligned} \quad (62)$$

From Theorem 8, for any $k \geq 0$, $A_k \geq \left[\frac{1+\nu}{1-\nu} \right]^{\frac{1-\nu}{1+\nu}} \frac{k^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}$. Combining this and (53), we obtain the statement of the Theorem.

□

Let us make a remark on complexity. As it can be seen from Theorem 12, whenever $A_k \geq 2R^2/\varepsilon$, the error in the objective value and equality constraints is smaller than ε . At the same time, using the lower bound for A_k , we obtain that the number of iterations to achieve this accuracy is $O\left(\left(\frac{M_\nu^{\frac{2}{1+\nu}} R^2}{\varepsilon^{\frac{2}{1+\nu}}}\right)^{\frac{1+\nu}{1+3\nu}}\right)$.

5. Numerical experiments

5.1. Smooth convex problem

We tested the AGMsDR method on the problem of minimizing

$$f(x) = \frac{L}{8}x_1^2 + \sum_{i=1}^{n-1}(x_i - x_{i+1}^2) + x_n^2 - \frac{L}{4}x_1,$$

where $L = 10$ and the dimensionality $n = 1000$. This is the function used to derive the lower complexity bound for the class of L -smooth convex objectives in [24]. The results are presented below. The method was compared to the Linear Coupling method [1] and the Conjugate Gradients and BFGS methods implemented in the SciPy python package.

For all four methods the initial point was the origin. The results are presented below.

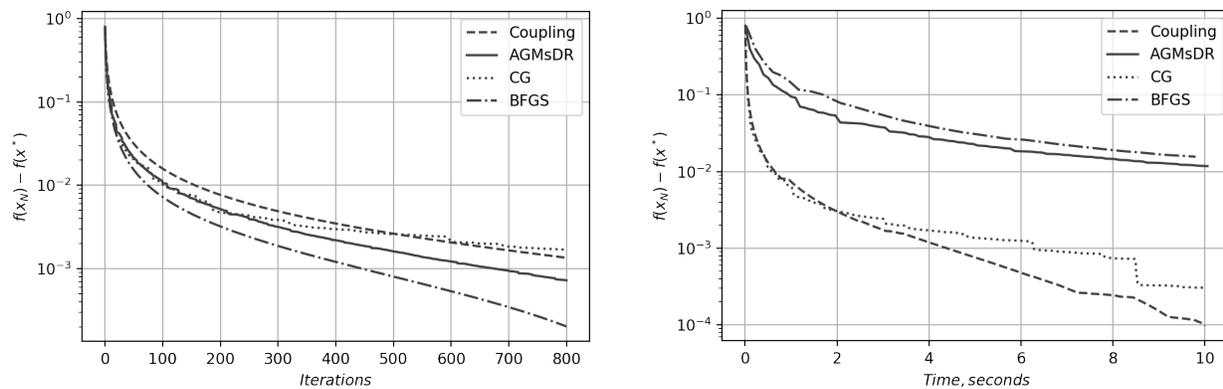


Figure 1.: Convergence for the smooth problem.

While utilizing line-search slightly improves the convergence rate in terms of required iterations compared to methods with fixed step sizes, the inherent complexity of line-search significantly increases the cost of each iteration. In the end, first-order methods with fixed step-sizes showed best results.

5.2. Non-smooth convex problem

We compared three different universal methods – UAGMsDR from this paper, Universal Linear Coupling Method from [14] and Universal Fast Gradient Method from

[29] – on the MAXQ problem [15]:

$$f(x) = \max_{1 \leq i \leq n} x_i^2 = \|x\|_\infty^2,$$

with the initial point

$$x_i^0 = \begin{cases} i, & \text{for } i = 1, \dots, n/2 \\ -i, & \text{for } i = n/2 + 1, \dots, n \end{cases}$$

and $n = 100$. For all methods the accuracy ε was set to $5 \cdot 10^{-4}$.

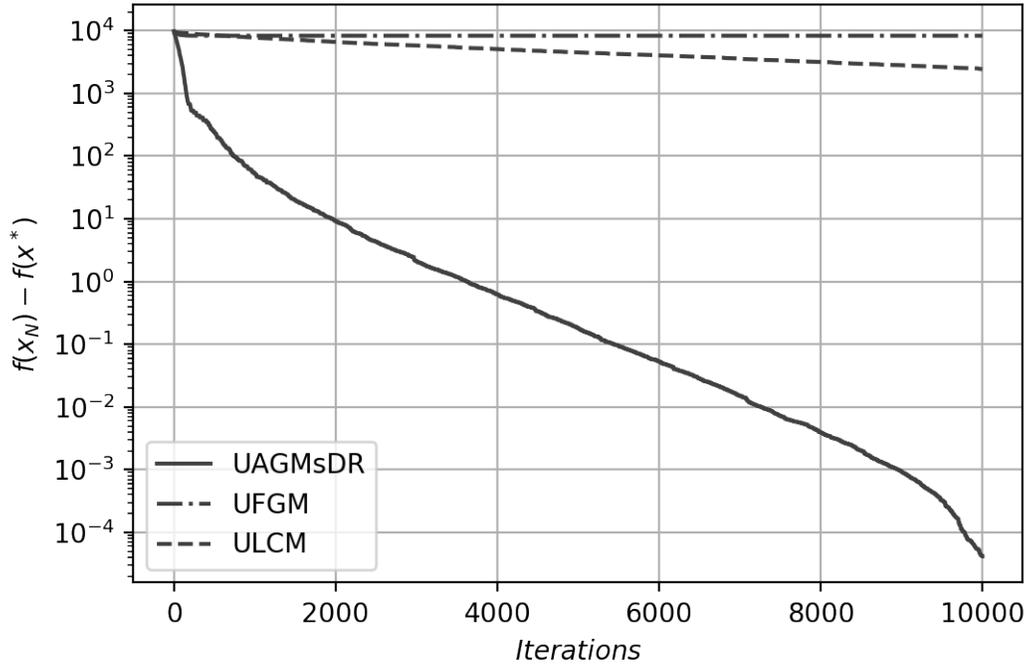


Figure 2.: Convergence for the non-smooth problem.

Even though all three methods have identical (up to a small constant multiplicative factor) theoretical convergence rates, for this problem the UAGMsDR method demonstrated practically linear convergence rate. It seems that using two line-searches in orthogonal directions helps the method use the fact that the graph of the function is, in a sense, similar to a quadratic function.

5.3. Non-convex problem

We consider the following non-convex objective with unique extremal point (which is the global minimum) at $(1, 1, 1, \dots, 1)$:

$$f(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^n (x_{i+1} - 2x_i^2 + 1)^2$$

and with the initial point $x^0 = (-1, -1, -1, \dots, -1)$. Even with low dimensionality $n \simeq 15$ this problem is very hard. There are points x such that $\|\nabla f(x)\|_2 \simeq 10^{-8}$, while at the same time typically $f(x) - f(x_*) = f(x) \simeq \frac{1}{2}$.

To perform exact line-search for this problem we utilized the fact that the objective is a polynomial. For example, to minimize the objective over the line $\{y^k - h\nabla f(y^k) | h \in \mathbb{R}\}$ it is then sufficient to find the roots of the third degree polynomial $\frac{\partial}{\partial h} f(y^k - h\nabla f(y^k))$ and choose the one which corresponds to the lesser value of $f(y^k - h\nabla f(y^k))$. For all universal methods the accuracy ε was set to $5 \cdot 10^{-4}$.

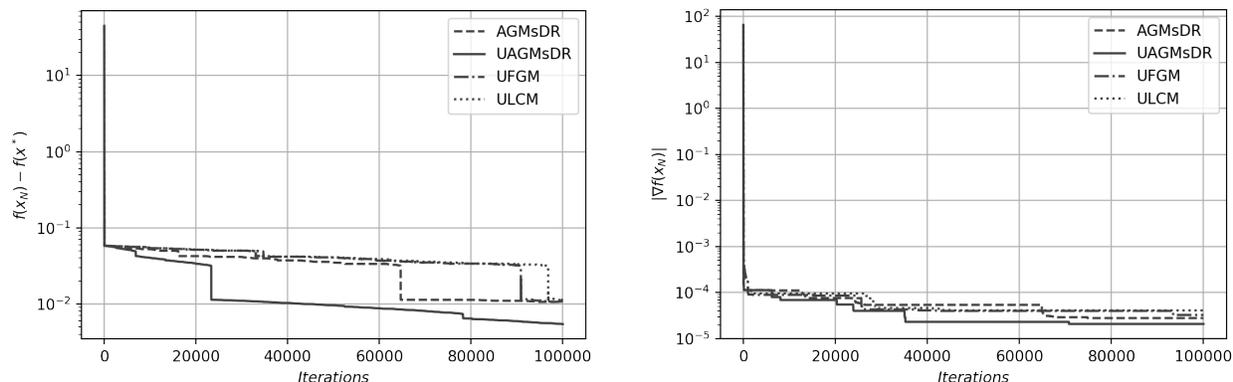


Figure 3.: Convergence for the smooth problem.

The universal UAGMsDR method demonstrates best performance, both in terms of convergence in gradient and the function’s value.

6. Funding

The work in Section 2 was funded by Russian Science Foundation (project 18-71-10108). The work in Section 3 was supported by grant RFBR 18-29-03071 mk. The work in Section 4 was supported Grant of the President of the Russian Federation MD-1320.2018.1 and RFBR 18-31-20005 mol_a_ved.

References

- [1] Z. Allen-Zhu and L. Orecchia, *Linear coupling: An ultimate unification of gradient and mirror descent*, arXiv:1407.1537 (2014).
- [2] N. Andrei, *40 conjugate gradient algorithms for unconstrained optimization. a survey on their definition*, 2008. <https://camo.ici.ro/neculai/p13a08.pdf>.
- [3] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences 2 (2009), pp. 183–202. Available at <https://doi.org/10.1137/080716542>.
- [4] A. Beck and M. Teboulle, *A fast dual proximal gradient algorithm for convex minimization and applications*, Operations Research Letters 42 (2014), pp. 1 – 6.
- [5] 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.

- [6] A. Chernov, P. Dvurechensky, and A. Gasnikov, *Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints*, in *Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings*, Y. Kochetov, M. Khachay, V. Beresnev, E. Nurminski, and P. Pardalos, eds. Springer International Publishing, 2016, pp. 391–403.
- [7] E. de Klerk, F. Glineur, and A.B. Taylor, *On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions*, *Optimization Letters* 11 (2017), pp. 1185–1199. Available at <https://doi.org/10.1007/s11590-016-1087-4>.
- [8] Y. Drori and A.B. Taylor, *Efficient first-order methods for convex minimization: a constructive approach*, arXiv:1803.05676 (2018).
- [9] P. Dvurechensky, A. Gasnikov, and A. Kroshnin, *Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm*, in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, eds., Proceedings of Machine Learning Research Vol. 80. 2018, pp. 1367–1376. arXiv:1802.04367.
- [10] P. Dvurechensky, A. Gasnikov, S. Omelchenko, and A. Tiurin, *Adaptive similar triangles method: a stable alternative to sinkhorn's algorithm for regularized optimal transport*, arXiv:1706.07622 (2017).
- [11] S. Ghadimi and G. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, *Mathematical Programming* 156 (2016), pp. 59–99. Available at <http://dx.doi.org/10.1007/s10107-015-0871-8>.
- [12] S. Ghadimi, G. Lan, and H. Zhang, *Generalized uniformly optimal methods for nonlinear programming*, arXiv:1508.07384 (2015). Available at <https://arxiv.org/abs/1508.07384>.
- [13] S. Guminov and A. Gasnikov, *Accelerated methods for α -weakly-quasi-convex problems*, arXiv preprint arXiv:1710.00797 (2017).
- [14] S. Guminov, A. Gasnikov, A. Anikin, and A. Gornov, *A universal modification of the linear coupling method*, *Optimization Methods and Software* 0 (2018), pp. 1–18. arXiv:1711.01850.
- [15] M. Haarala, K. Miettinen, and M.M. Mäkelä, *New limited memory bundle method for large-scale nonsmooth optimization*, *Optimization Methods and Software* 19 (2004), pp. 673–692.
- [16] D. Kim and J.A. Fessler, *Generalizing the optimized gradient method for smooth convex minimization*, *SIAM Journal on Optimization* 28 (2018), pp. 1920–1950.
- [17] G. Narkiss and M. Zibulevsky, *Sequential subspace optimization method for large-scale unconstrained problems*, Technion-IIT, Department of Electrical Engineering, 2005.
- [18] G. Narkiss and M. Zibulevsky, *Sequential subspace optimization method for large-scale unconstrained problems*, Technion-IIT, Department of Electrical Engineering, 2005.
- [19] A. Nemirovskii and D.B. Yudin, *Problem complexity and method efficiency in optimization* (1983).
- [20] A. Nemirovsky, *Information-based complexity of linear operator equations*, *Journal of Complexity* 8 (1992), pp. 153–175.
- [21] A. Nemirovsky and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, J. Wiley & Sons, New York, 1983.
- [22] Y.E. Nesterov, *Effective methods in nonlinear programming*, Moscow, Radio i Svyaz (1989).

- [23] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $o(1/k^2)$* , Soviet Mathematics Doklady 27 (1983), pp. 372–376.
- [24] Y. Nesterov, *Introductory Lectures on Convex Optimization: a basic course*, Kluwer Academic Publishers, Massachusetts, 2004.
- [25] Y. Nesterov, *Introductory lectures on convex optimization. applied optimization, vol. 87* (2004).
- [26] Y. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming 103 (2005), pp. 127–152.
- [27] Y. Nesterov, *How to make the gradients small*, Optima 88 (2012), pp. 10–11.
- [28] Y. Nesterov, *Gradient methods for minimizing composite functions*, Mathematical Programming 140 (2013), pp. 125–161. First appeared in 2007 as CORE discussion paper 2007/76.
- [29] Y. Nesterov, *Universal gradient methods for convex optimization problems*, Mathematical Programming 152 (2015), pp. 381–404. Available at <http://dx.doi.org/10.1007/s10107-014-0790-0>.
- [30] J. Nocedal and S.J. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, NY, USA, 2006.
- [31] A. Yurtsever, Q. Tran-Dinh, and V. Cevher, *A Universal Primal-dual Convex Optimization Framework*, in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA. MIT Press, NIPS’15, 2015, pp. 3150–3158.