Math. Program., Ser. A (2019) 175:69–107
https://doi.org/10.1007/s10107-018-1232-1

FULL LENGTH PAPER

Linear convergence of first order methods for non-strongly convex optimization

I. Necoara[1] · **Yu. Nesterov**[2] · **F. Glineur**[2]

Received: 26 July 2016 / Accepted: 4 January 2018 / Published online: 22 January 2018

Abstract The standard assumption for proving linear convergence of first order methods for smooth convex optimization is the strong convexity of the objective function, an assumption which does not hold for many practical applications. In this paper, we derive linear convergence rates of several first order methods for solving smooth non-strongly convex constrained optimization problems, i.e. involving an objective function with a Lipschitz continuous gradient that satisfies some relaxed strong convexity condition. In particular, in the case of smooth constrained convex optimization, we provide several relaxations of the strong convexity conditions and prove that they are sufficient for getting linear convergence for several first order methods such as projected gradient, fast gradient and feasible descent methods. We also provide examples of functional classes that satisfy our proposed relaxations of strong convexity conditions. Finally, we show that the proposed relaxed strong convexity conditions cover important applications ranging from solving linear systems, Linear Programming, and dual formulations of linearly constrained convex problems.

Mathematics Subject Classification 90C25 · 90C06 · 65K05

✉ I. Necoara
ion.necoara@acse.pub.ro

Yu. Nesterov
yurii.nesterov@uclouvain.be

F. Glineur
francois.glineur@uclouvain.be

[1] Automatic Control and Systems Engineering Department, University Politehnica Bucharest, 060042 Bucharest, Romania

[2] Center for Operations Research and Econometrics, Universite catholique de Louvain, 1348 Louvain-la-Neuve, Belgium

1 Introduction

Recently, there emerges a surge of interests in accelerating first order methods for difficult optimization problems, for example the ones without strongly convex objective function, arising in different applications such as data analysis [1] or machine learning [2]. Algorithms based on gradient information have proved to be effective in these settings, such as projected gradient and its fast variants [3], stochastic gradient descent [4] or coordinate gradient descent [5].

For smooth convex programming, i.e. optimization problems with convex objective function having a Lipschitz continuous gradient with constant $L_f > 0$, first order methods are converging sublinearly. In order to get an ϵ-optimal solution, we need to perform $\mathcal{O}\left(\frac{L_f}{\epsilon}\right)$ or even $\mathcal{O}\left(\sqrt{\frac{L_f}{\epsilon}}\right)$ calls to the oracle [3]. Typically, for proving linear convergence of the first order methods we also need to require strong convexity for the objective function. Unfortunately, many practical applications do not have strongly convex objective functions. A new line of analysis, that circumvents these difficulties, was developed using several notions. For example, sharp minima type condition for non-strongly convex optimization problems, i.e. the epigraph of the objective function is a polyhedron, has been proposed in [6–8]. An error bound property, that estimates the distance to the solution set from any feasible point by the norm of the proximal residual, has been analyzed in [9–11]. Finally, a restricted (also called essential) strong convexity inequality, which basically imposes a quadratic lower bound on the objective function, has been derived in [2,12]. For all these conditions (sharp minima, error bound or restricted strong convexity) several gradient-type methods are shown to converge linearly, see e.g. [2,9–12]. Several other papers on the linear convergence of first order methods for non-strongly convex optimization have appeared recently (after this paper was finalized) [8,13–15]. The main goal of this paper is to develop a framework for finding general functional conditions for smooth convex constrained optimization problems that allow us to prove linear convergence for a broad class of first order methods.

1.1 Contributions

For smooth constrained optimization, we show in this paper that some relaxations of the strong convexity conditions of the objective function are sufficient for obtaining linear convergence for several first order methods. The most general relaxation we introduce is a quadratic functional growth condition, which states that the objective function grows faster than the squared distance between any feasible point and the optimal set. We also propose other non-strongly convex conditions, which are more conservative than the quadratic functional growth condition, and establish relations between them. Further, we provide examples of functional classes that satisfy our proposed relaxations of strong convexity conditions. For all these smooth non-strongly convex constrained optimization problems, we prove that the corresponding relaxations are sufficient for getting linear convergence for several first order methods, such as projected gradient, fast gradient and feasible descent methods. We also show that the corresponding

linear rates are improved in some cases compared to the existing results. We also establish necessary and sufficient conditions for linear convergence of the gradient method. Finally, we show that the proposed relaxed strong convexity conditions cover important applications ranging from solving linear systems, Linear Programming, and dual formulations of linearly constrained convex problems.

1.2 Notations

We work in the space $\mathbb{R}^n$ composed of column vectors and $\mathbb{R}^n_+$ denotes the non-negative orthant. For $u, v \in \mathbb{R}^n$ we denote by $\langle u, v\rangle = u^T v$ the Euclidean inner product, $\|u\| = \sqrt{\langle u, u\rangle}$ the Euclidean norm and $[u]_X = \arg\min_{x\in X} \|x - u\|$ the projection of u onto convex set X. For matrix $A \in \mathbb{R}^{m\times n}$, we denote by $\sigma_{\min}(A)$ the smallest nonzero singular value and $\|A\|$ spectral norm.

2 Problem formulation

In this paper we consider the class of convex constrained optimization problems:

$$\textbf{(P)}: \quad f^* = \min_{x\in X} f(x),$$

where $X \subseteq \mathbb{R}^n$ is a simple closed convex set, that is the projection onto this set is easy, and $f : X \to \mathbb{R}$ is a closed convex function. We further denote by $X^* = \arg\min_{x\in X} f(x)$ the set of optimal solutions of problem (P). We assume throughout the paper that the optimal set X^* is nonempty and closed and the optimal value f^* is finite. Moreover, in this paper we assume that the objective function is smooth, that is f has Lipschitz continuous gradient with constant $L_f > 0$ on the set X:

$$\|\nabla f(x) - \nabla f(y)\| \le L_f \|x - y\| \quad \forall x, y \in X. \tag{1}$$

An immediate consequence of (1) is the following inequality [3]:

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_f}{2}\|x - y\|^2 \quad \forall x, y \in X, \tag{2}$$

while, under convexity of f, we also have:

$$0 \le \langle \nabla f(x) - \nabla f(y), x - y\rangle \le L_f \|x - y\|^2 \quad \forall x, y \in X. \tag{3}$$

It is well known that first order methods are converging sublinearly on the class of convex problems whose objective function f has Lipschitz continuous gradient with constant L_f on the set X, e.g. convergence rates in terms of function values of order [3]:

$$\begin{aligned} f(x^k) - f^* &\leq \frac{L_f \|x^0 - x^*\|^2}{2k} \quad \text{for projected gradient,} \\ f(x^k) - f^* &\leq \frac{2L_f \|x^0 - x^*\|^2}{(k+1)^2} \quad \text{for fast gradient,} \end{aligned} \tag{4}$$

where x^k is the kth iterate generated by the method. Typically, in order to show linear convergence of first order methods applied for solving smooth convex problems, we need to require strong convexity of the objective function. We recall that f is strongly convex function on the convex set X with constant $\sigma_f > 0$ if the following inequality holds [3]:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{\sigma_f \alpha (1-\alpha)}{2} \|x - y\|^2 \tag{5}$$

for all $x, y \in X$ and $\alpha \in [0,\ 1]$. Note that if $\sigma_f = 0$, then f is simply a convex function. We denote by $\mathcal{S}_{L_f,\sigma_f}(X)$ the class of σ_f-strongly convex functions with an L_f-Lipschitz continuous gradient on X. First order methods are converging linearly on the class of problems (P) whose objective function f is in $\mathcal{S}_{L_f,\sigma_f}(X)$, e.g. convergence rates of order [3]:

$$\begin{aligned} f(x^k) - f^* &\leq \frac{L_f \|x^0 - x^*\|^2}{2} \left(1 - \frac{\sigma_f}{L_f}\right)^k \quad \text{for projected gradient,} \\ f(x^k) - f^* &\leq 2\left(f(x^0) - f^*\right) \left(1 - \sqrt{\frac{\sigma_f}{L_f}}\right)^k \quad \text{for fast gradient.} \end{aligned} \tag{6}$$

In the case of a differentiable function f with L_f-Lipschitz continuous gradient, each of the following conditions below is equivalent to inclusion $f \in \mathcal{S}_{L_f,\sigma_f}(X)$ [3]:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma_f}{2} \|x - y\|^2 \qquad \forall x, y \in X, \tag{7}$$

$$\sigma_f \|x - y\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \qquad \forall x, y \in X. \tag{8}$$

Let us give some properties of smooth strongly convex functions from the class $\mathcal{S}_{L_f,\sigma_f}(X)$. Recall that the gradient mapping of a continuous differentiable function f with Lipschitz continuous gradient in a point $x \in \mathbb{R}^n$ is defined as [3]:

$$g(x) = L_f \left(x - [x - 1/L_f \nabla f(x)]_X\right),$$

Then, we obtain a first relation valid for any $f \in \mathcal{S}_{L_f,\sigma_f}(X)$ [11] [Lemma 22]:

$$\frac{\sigma_f}{2} \|x - y\| \leq \|g(x) - g(y)\| \qquad \forall x, y \in X. \tag{9}$$

Further, using the optimality conditions for (P), that is $\langle \nabla f(x^*), y - x^* \rangle \geq 0$ for all $y \in X$ and $x^* \in X^*$, in (7) we get a second relation:

$$f(x) - f^* \geq \frac{\sigma_f}{2}\|x - x^*\|^2 \qquad \forall x \in X. \tag{10}$$

However, in many applications the strong convexity condition (5) or equivalently one of the conditions (7)–(8) cannot be assumed to hold. Therefore, in the next sections we introduce some non-strongly convex conditions for the objective function f that are less conservative than strong convexity. These are based on relaxations of strong convexity relations (7)–(9).

3 Non-strongly convex conditions for a function

In this section we introduce several functional classes that are relaxing the strong convexity properties (7)–(9) of a function and derive relations between these classes. More precisely, we observe that strong convexity relations (7)–(9) are valid for all $x, y \in X$. We propose in this paper functional classes satisfying conditions of the form (7)–(9) that hold for some particular choices of x and y, or satisfying simply the condition (10).

3.1 Quasi-strong convexity

The first non-strongly convex relaxation we introduce is based on choosing a particular value for y in the strong convexity inequality (7), that is $y = \bar{x} \equiv [x]_{X^*}$ (recall that $[x]_{X^*}$ denotes the projection of x onto the optimal set X^* of convex problem (P)):

Definition 1 Continuously differentiable function f is called *quasi-strongly convex* on set X if there exists a constant $\kappa_f > 0$ such that for any $x \in X$ and $\bar{x} = [x]_{X^*}$ we have:

$$f^* \geq f(x) + \langle \nabla f(x), \bar{x} - x \rangle + \frac{\kappa_f}{2}\|x - \bar{x}\|^2 \quad \forall x \in X. \tag{11}$$

Note that inequality (11) alone does not even imply convexity of function f. Moreover, our definition of quasi-strongly convex functions does not ensure uniqueness of the optimal solution of problem (P) and does not require f to have Lipschitz continuous gradient. We denote the class of convex functions with Lipschitz continuous gradient with constant L_f in (1) and satisfying the quasi-strong convexity property with constant κ_f in (11) by $q\mathcal{S}_{L_f,\kappa_f}(X)$. Clearly, for strongly convex functions with constant κ_f, from the strong convexity condition (7) with $y = x^* \in X^*$, we observe that the following inclusion holds:

$$\mathcal{S}_{L_f,\kappa_f}(X) \subseteq q\mathcal{S}_{L_f,\kappa_f}(X). \tag{12}$$

Moreover, combining the inequalities (2) and (11), we obtain that the condition number of objective function $f \in q\mathcal{S}_{L_f,\kappa_f}(X)$, defined as $\mu_f = \kappa_f / L_f$, satisfies:

$$0 < \mu_f \leq 1. \tag{13}$$

We will derive below other functional classes that are related to our newly introduced class of quasi-strongly convex functions $q\mathcal{S}_{L_f,\kappa_f}(X)$.

3.2 Quadratic under-approximation

Let us define the class of functions satisfying a quadratic under-approximation on the set X, obtained from relaxing the strong convex inequality (7) by choosing $y = x$ and $x = \bar{x} \equiv [x]_{X^*}$:

Definition 2 Continuously differentiable function f has a *quadratic under-approximation* on X if there exists a constant $\kappa_f > 0$ such that for any $x \in X$ and $\bar{x} = [x]_{X^*}$ we have:

$$f(x) \geq f^* + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{\kappa_f}{2} \|x - \bar{x}\|^2 \quad \forall x \in X. \tag{14}$$

We denote the class of convex functions with Lipschitz continuous gradient and satisfying the quadratic under-approximation property (14) on X by $\mathcal{U}_{L_f,\kappa_f}(X)$. Then, we have the following inclusion:

Theorem 1 *Inequality* (11) *implies inequality* (14). *Therefore, the following inclusion holds:*

$$q\mathcal{S}_{L_f,\kappa_f}(X) \subseteq \mathcal{U}_{L_f,\kappa_f}(X). \tag{15}$$

Proof Let $f \in q\mathcal{S}_{L_f,\kappa_f}(X)$. Since f is convex function, it satisfies the inequality (14) with some constant $\kappa_f(0) \geq 0$, i.e.:

$$f(x) \geq f^* + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{\kappa_f(0)}{2} \|x - \bar{x}\|^2. \tag{16}$$

Using first order Taylor approximation in the integral form and the identity $[\bar{x} + \tau(x - \bar{x})]_{X^*} = \bar{x}$ for $\tau \in [0, 1]$, we have:

$$\begin{aligned} f(x) &= f(\bar{x}) + \int_0^1 \langle \nabla f(\bar{x} + \tau(x - \bar{x})), x - \bar{x} \rangle d\tau \\ &= f(\bar{x}) + \int_0^1 \frac{1}{\tau} \langle \nabla f(\bar{x} + \tau(x - \bar{x})), \tau(x - \bar{x}) \rangle d\tau \\ &\overset{(11) \text{ in } \bar{x}+\tau(x-\bar{x})}{\geq} f(\bar{x}) + \int_0^1 \frac{1}{\tau} \left(f(\bar{x} + \tau(x - \bar{x})) - f(\bar{x}) + \frac{\kappa_f}{2} \|\tau(x - \bar{x})\|^2 \right) d\tau \end{aligned}$$

$$\overset{(16)}{\geq} f(\bar{x}) + \int_0^1 \frac{1}{\tau}\left(\langle \nabla f(\bar{x}), \tau(x-\bar{x})\rangle + \frac{\kappa_f(0)}{2}\|\tau(x-\bar{x})\|^2\right) + \frac{1}{\tau}\frac{\kappa_f}{2}\|\tau(x-\bar{x})\|^2 d\tau$$
$$= f(\bar{x}) + \int_0^1 \langle \nabla f(\bar{x}), x-\bar{x}\rangle + \frac{\tau\kappa_f(0)}{2}\|x-\bar{x}\|^2 + \frac{\tau\kappa_f}{2}\|x-\bar{x}\|^2 d\tau$$
$$= f(\bar{x}) + \langle \nabla f(\bar{x}), x-\bar{x}\rangle + \frac{\kappa_f(0)+\kappa_f}{2}\cdot\frac{1}{2}\|x-\bar{x}\|^2.$$

If we denote $\kappa_f(1) = \frac{\kappa_f(0)+\kappa_f}{2}$, then we get that inequality (16) also holds for $\kappa_f(1)$. Repeating the same argument as above for $f \in qS_{L_f,\kappa_f}(X)$ and satisfying (16) for $\kappa_f(1)$ we get that inequality (16) also holds for $\kappa_f(2) = \frac{\kappa_f(1)+\kappa_f}{2} = \frac{\kappa_f(0)+3\kappa_f}{4}$. Iterating this procedure we obtain that after t steps:

$$\kappa_f(t) = \frac{\kappa_f(t-1)+\kappa_f}{2} = \frac{\kappa_f(0)+(2^t-1)\kappa_f}{2^t} \to \kappa_f \quad \text{as} \quad t \to \infty.$$

Since after any t steps the inequality (16) holds with $\kappa_f(t)$, using continuity of $\kappa_f(t)$ in (16) we obtain (14). This proves our statement. □

Moreover, combining the inequalities (2) and (14), we obtain that the condition number of objective function $f \in \mathcal{U}_{L_f,\kappa_f}(X)$, defined as $\mu_f = \kappa_f/L_f$, satisfies:

$$0 < \mu_f \leq 1. \tag{17}$$

3.3 Quadratic gradient growth

Let us define the class of functions satisfying a bound on the variation of gradients over the set X. It is obtained by relaxing the strong convex inequality (8) by choosing $y = \bar{x} \equiv [x]_{X^*}$:

Definition 3 Continuously differentiable function f has a *quadratic gradient growth* on set X if there exists a constant $\kappa_f > 0$ such that for any $x \in X$ and $\bar{x} = [x]_{X^*}$ we have:

$$\langle \nabla f(x) - \nabla f(\bar{x}), x-\bar{x}\rangle \geq \kappa_f\|x-\bar{x}\|^2 \quad \forall x \in X. \tag{18}$$

Now, let us denote the class of convex differentiable functions with Lipschitz gradient and satisfying the quadratic gradient growth (18) by $\mathcal{G}_{L_f,\kappa_f}(X)$. In [12] the authors analyzed a similar class of objective functions, but for unconstrained optimization problems, that is $X = \mathbb{R}^n$, which was called *restricted strong convexity* and was defined as: there exists a constant $\kappa_f > 0$ such that $\langle \nabla f(x), x-\bar{x}\rangle \geq \kappa_f\|x-\bar{x}\|^2$ for all $x \in \mathbb{R}^n$. An immediate consequence of Theorem 1 is the following inclusion:

Theorem 2 *Inequality* (11) *implies inequality* (18). *Therefore, the following inclusion holds:*

$$qS_{L_f,\kappa_f}(X) \subseteq \mathcal{G}_{L_f,\kappa_f}(X). \tag{19}$$

Proof If $f \in qS_{L_f,\kappa_f}(X)$, then f satisfies the inequality (11). From Theorem 1 we also have that f satisfies inequality (14). By adding the two inequalities (11) and (14) in x we get:

$$\langle \nabla f(x) - \nabla f(\bar{x}), x - \bar{x} \rangle \geq \kappa_f \|x - \bar{x}\|^2 \quad \forall x \in X, \tag{20}$$

which proves that inequality (18) holds. □

We prove below that (11) or (14) alone and convexity of f implies (18) with constant $\kappa_f/2$. Indeed, let us assume for example that (14) holds, then we have:

$$\begin{aligned} f(x) &\geq f^* + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{\kappa_f}{2}\|x - \bar{x}\|^2 \\ &\geq f(x) + \langle \nabla f(x), \bar{x} - x \rangle + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{\kappa_f}{2}\|x - \bar{x}\|^2 \\ &= f(x) + \langle \nabla f(\bar{x}) - \nabla f(x), x - \bar{x} \rangle + \frac{\kappa_f}{2}\|x - \bar{x}\|^2, \end{aligned}$$

which leads to (18) with constant $\kappa_f/2$. Combining the inequalities (3) and (18), we obtain that the condition number of objective function $f \in \mathcal{G}_{L_f,\kappa_f}(X)$, satisfies:

$$0 < \mu_f \leq 1. \tag{21}$$

Theorem 3 *Inequality* (18) *implies inequality* (14). *Therefore, the following inclusion holds:*

$$\mathcal{G}_{L_f,\kappa_f}(X) \subseteq \mathcal{U}_{L_f,\kappa_f}(X). \tag{22}$$

Proof Let $f \in \mathcal{G}_{L_f,\kappa_f}(X)$, then from first order Taylor approximation in the integral form we get:

$$\begin{aligned} f(x) &= f(\bar{x}) + \int_0^1 \langle \nabla f(\bar{x} + t(x - \bar{x})), x - \bar{x} \rangle dt \\ &= f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \int_0^1 \langle \nabla f(\bar{x} + t(x - \bar{x})) - \nabla f(\bar{x}), x - \bar{x} \rangle dt \\ &= f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \int_0^1 \frac{1}{t}\langle \nabla f(\bar{x} + t(x - \bar{x})) - \nabla f(\bar{x}), t(x - \bar{x}) \rangle dt \\ &\overset{(18)}{\geq} f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \int_0^1 \frac{1}{t}\kappa_f \|t(x - \bar{x})\|^2 dt \\ &= f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{\kappa_f}{2}\|x - \bar{x}\|^2, \end{aligned}$$

where we used that $[\bar{x} + t(x - \bar{x})]_{X^*} = \bar{x}$ for any $t \in [0,\ 1]$. This chain of inequalities proves that f satisfies inequality (14) with the same constant κ_f. □

3.4 Quadratic functional growth

We further define the class of functions satisfying a quadratic functional growth property on the set X. It shows that the objective function grows faster than the squared distance between any feasible point and the optimal set. More precisely, since $\langle \nabla f(x^*), y - x^* \rangle \geq 0$ for all $y \in X$ and $x^* \in X^*$, then using this relation and choosing $y = x$ and $x = \bar{x} \equiv [x]_{X^*}$ in the strong convex inequality (7), we get a relaxation of this condition similar to inequality (10):

Definition 4 Continuously differentiable function f has a *quadratic functional growth* on X if there exists a constant $\kappa_f > 0$ such that for any $x \in X$ and $\bar{x} = [x]_{X^*}$ we have:

$$f(x) - f^* \geq \frac{\kappa_f}{2}\|x - \bar{x}\|^2 \quad \forall x \in X. \tag{23}$$

Since the above quadratic functional growth inequality is given in $\bar{x}$, this does not mean that f grows everywhere faster than the quadratic function $\kappa_f/2\|x - \bar{x}\|^2$. We denote the class of convex differentiable functions with Lipschitz continuous gradient and satisfying the quadratic functional growth (23) by $\mathcal{F}_{L_f,\kappa_f}(X)$.

Note that each inequality (11), (14), (18) or (23) alone does not even imply convexity of function f. This observation open the possibility to extend our framework to the non-convex settings. We now derive inclusion relations between the functional classes we have introduced so far (see also Fig. 1):

Theorem 4 *The following chain of implications are valid:*

$$(7) \Rightarrow (11) \Rightarrow (18) \Rightarrow (14) \Rightarrow (23).$$

Therefore, the following inclusions hold:

$$\mathcal{S}_{L_f,\kappa_f}(X) \subseteq q\mathcal{S}_{L_f,\kappa_f}(X) \subseteq \mathcal{G}_{L_f,\kappa_f}(X) \subseteq \mathcal{U}_{L_f,\kappa_f}(X) \subseteq \mathcal{F}_{L_f,\kappa_f}(X). \tag{24}$$

Proof From the optimality conditions for problem (P) we have $\langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq 0$ for all $x \in X$. Then, for any objective function f satisfying (14), i.e. $f \in \mathcal{U}_{L_f,\kappa_f}(X)$, we also have (23). In conclusion, from previous derivations, (12) and Theorems 2 and 3 we obtain our chain of inclusions. □

Let us define the condition number of objective function $f \in \mathcal{F}_{L_f,\kappa_f}(X)$ as $\mu_f = \frac{\kappa_f}{L_f}$. If the feasible set X is unbounded, then combining (2) with (23) and considering $\|x - \bar{x}\| \to \infty$, we conclude that:

$$0 < \mu_f \leq 1. \tag{25}$$

However, if the feasible set X is bounded, we may have $\kappa_f \gg L_f$, provided that $\|\nabla f(\bar{x})\|$ is large, and thus the condition number might be greater than 1:

$$\mu_f \geq 1. \tag{26}$$

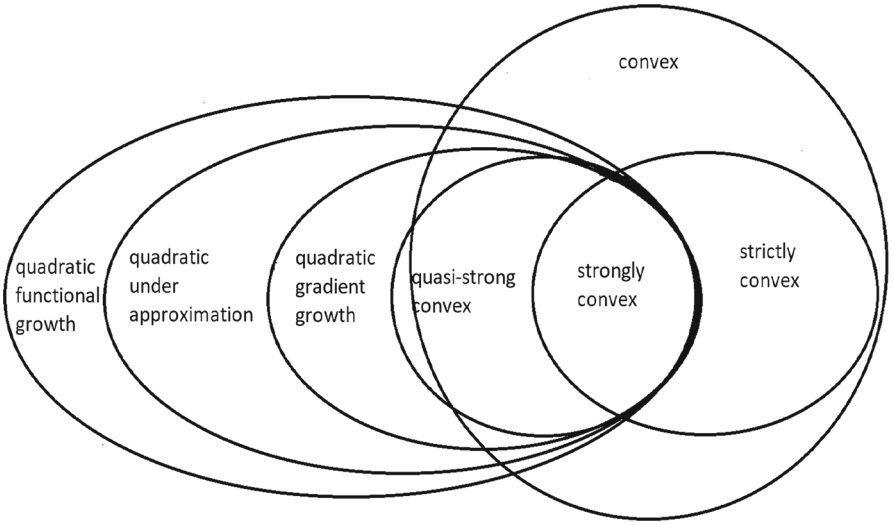

Fig. 1 Chain of inclusions between the functional classes: convex, strictly convex, strongly convex, quasi-strong convex, quadratic gradient growth, quadratic under-approximation, quadratic functional growth

Moreover, from the inclusions given by Theorem 4 we conclude that:

$$\mu_f(\mathcal{S}) \leq \mu_f(q\mathcal{S}) \leq \mu_f(\mathcal{G}) \leq \mu_f(\mathcal{U}) \leq \mu_f(\mathcal{F}).$$

Let us denote the projected gradient step from x with:

$$x^+ = [x - 1/L_f \nabla f(x)]_X,$$

and its projection onto the optimal set X^* with $\bar{x}^+ = [x^+]_{X^*}$. Then, we can prove that if x^+ is closer to X^* than x, then the objective function f must satisfy the quadratic functional growth (23). This result will be used later (see Theorem 13 below) to derive necessary and sufficient conditions for linear convergence of the gradient scheme.

Theorem 5 *Let f be a convex function with Lipschitz continuous gradient with constant L_f. If there exists some positive constant $\beta < 1$ such that the following inequality holds:*

$$\|x^+ - \bar{x}^+\| \leq \beta \|x - \bar{x}\| \quad \forall x \in X,$$

then f satisfies the quadratic functional growth (23) on X with the constant $\kappa_f = L_f(1-\beta)^2$.

Proof On the one hand, from triangle inequality for the projection we have:

$$\|x - \bar{x}\| \leq \|x - \bar{x}^+\| \leq \|x - x^+\| + \|x^+ - \bar{x}^+\|.$$

Combining this relation with the condition from the theorem, that is $\|x^+ - \bar{x}^+\| \leq \beta\|x - \bar{x}\|$, we get:

$$(1-\beta)\|x - \bar{x}\| \leq \|x - x^+\|. \tag{27}$$

On the other hand, we note that x^+ is the optimal solution of the problem:

$$x^+ = \arg\min_{z \in X} \left[f(x) + \langle \nabla f(x), z - x \rangle + \frac{L_f}{2}\|z - x\|^2 \right]. \tag{28}$$

From (2) we have:

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

and combining with the optimality conditions of (28) in x, that is $\langle \nabla f(x) + L_f(x^+ - x), x - x^+ \rangle \leq 0$, we get the following decrease in terms of the objective function:

$$f(x^+) \leq f(x) - \frac{L_f}{2}\|x^+ - x\|^2. \tag{29}$$

Finally, combining (27) with (29), and using $f(x^+) \geq f^*$, we get our statement. □

3.5 Error bound property

Let us recall the gradient mapping of a continuous differentiable function f with Lipschitz continuous gradient in a point $x \in \mathbb{R}^n$: $g(x) = L_f(x - x^+)$, where $x^+ = [x - 1/L_f \nabla f(x)]_X$ is the projected gradient step from x. Note that $g(x^*) = 0$ for all $x^* \in X^*$. Moreover, if $X = \mathbb{R}^n$, then $g(x) = \nabla f(x)$. Recall that the main property of the gradient mapping for convex objective functions with Lipschitz continuous gradient of constant L_f is given by the following inequality [3] [Theorem 2.2.7]:

$$f(y) \geq f(x^+) + \langle g(x), y - x \rangle + \frac{1}{2L_f}\|g(x)\|^2 \quad \forall y \in X \quad \text{and} \quad x \in \mathbb{R}^n. \tag{30}$$

Taking $y = \bar{x}$ in (30) and using that $f(x^+) \geq f^*$, we get the simpler inequality:

$$\langle g(x), x - \bar{x} \rangle \geq \frac{1}{2L_f}\|g(x)\|^2 \quad \forall x \in \mathbb{R}^n. \tag{31}$$

In [9] Tseng introduced an error bound condition that estimates the distance to the solution set from any feasible point by the norm of the proximal residual: there exists a constant $\kappa > 0$ such that $\|x - \bar{x}\| \leq \kappa\|x - [x - \nabla f(x)]_X\|$ for all $x \in X$. This notion was further extended and analyzed in [10,11]. Next, we define an error bound type condition, obtained from the relaxation of the strong convex inequality (9) for the particular choice $y = \bar{x} \equiv [x]_{X^*}$ and from the relation $g(\bar{x}) = 0$:

Definition 5 The continuously differentiable function f has a *global error bound* on X if there exists a constant $\kappa_f > 0$ such that for any $x \in X$ and $\bar{x} = [x]_{X^*}$ we have:

$$\|g(x)\| \geq \kappa_f \|x - \bar{x}\| \quad \forall x \in X. \tag{32}$$

We denote the class of convex functions with Lipschitz continuous gradient and satisfying the error bound (32) by $\mathcal{E}_{L_f,\kappa_f}$. Let us define the condition number of the objective function $f \in \mathcal{E}_{L_f,\kappa_f}(X)$ as $\mu_f = \frac{\kappa_f}{L_f}$. Combining inequality (31) and (32) we conclude that the condition number satisfies the inequality:

$$0 < \mu_f \leq 2. \tag{33}$$

However, for the unconstrained case, i.e. $X = \mathbb{R}^n$ and $\nabla f(\bar{x}) = 0$, from (1) and (32) we get $0 < \mu_f \leq 1$. We now determine relations between the quadratic functional growth condition and the error bound condition.

Theorem 6 *Inequality* (32) *implies inequality* (23) *with constant* $\mu_f \cdot \kappa_f$. *Therefore, the following inclusion holds for the functional class* $\mathcal{E}_{L_f,\kappa_f}(X)$*:*

$$\mathcal{E}_{L_f,\kappa_f}(X) \subseteq \mathcal{F}_{L_f,\mu_f\cdot\kappa_f}(X). \tag{34}$$

Proof Combining (29) and (32) we obtain:

$$\kappa_f^2 \|x - \bar{x}\|^2 \leq \|g(x)\|^2 \leq 2L_f(f(x) - f(x^+)) \leq 2L_f(f(x) - f^*) \quad \forall x \in X.$$

In conclusion, inequality (23) holds with the constant $\frac{\kappa_f^2}{L_f} = \mu_f \cdot \kappa_f$, where we recall $\mu_f = \kappa_f / L_f$. This also proves the inclusion: $\mathcal{E}_{L_f,\kappa_f}(X) \subseteq \mathcal{F}_{L_f,\mu_f\cdot\kappa_f}(X)$. □

Theorem 7 *Inequality* (23) *implies inequality* (32) *with constant* $\frac{1}{1+\mu_f+\sqrt{1+\mu_f}} \cdot \kappa_f$. *Therefore, the following inclusion holds for the functional class* $\mathcal{F}_{L_f,\kappa_f}(X)$*:*

$$\mathcal{F}_{L_f,\kappa_f}(X) \subseteq \mathcal{E}_{L_f,\frac{1}{1+\mu_f+\sqrt{1+\mu_f}}\cdot\kappa_f}(X). \tag{35}$$

Proof From the gradient mapping property (30) evaluated at the point $y = \bar{x}^+ \equiv [x^+]_{X^*}$, we get:

$$\begin{aligned} f^* &\geq f(x^+) + \langle g(x), \bar{x}^+ - x\rangle + \frac{1}{2L_f}\|g(x)\|^2 \\ &= f(x^+) + \langle g(x), \bar{x}^+ - x^+\rangle - \frac{1}{2L_f}\|g(x)\|^2. \end{aligned}$$

Further, combining the previous inequality and (23), we obtain:

$$\langle g(x), x^+ - \bar{x}^+ \rangle + \frac{1}{2L_f}\|g(x)\|^2 \geq f(x^+) - f^* \geq \frac{\kappa_f}{2}\|x^+ - \bar{x}^+\|^2.$$

Using Cauchy–Schwarz inequality for the scalar product and then rearranging the terms we obtain:

$$\frac{1}{2L_f}\left(\|g(x)\| + L_f\|x^+ - \bar{x}^+\|\right)^2 \geq \frac{\kappa_f + L_f}{2}\|x^+ - \bar{x}^+\|^2$$

or equivalently

$$\|g(x)\| + L_f\|x^+ - \bar{x}^+\| \geq \sqrt{L_f(\kappa_f + L_f)}\|x^+ - \bar{x}^+\|.$$

We conclude that:

$$\|g(x)\| \geq \left(\sqrt{L_f(\kappa_f + L_f)} - L_f\right)\|x^+ - \bar{x}^+\|.$$

Since

$$\|x - \bar{x}\| \leq \|x - \bar{x}^+\| \leq \|x - x^+\| + \|x^+ - \bar{x}^+\| = \frac{1}{L_f}\|g(x)\| + \|x^+ - \bar{x}^+\|,$$

then we obtain:

$$\|g(x)\| \geq \left(\sqrt{L_f(\kappa_f + L_f)} - L_f\right)\left(\|x - \bar{x}\| - \frac{1}{L_f}\|g(x)\|\right).$$

After simple manipulations and using that $\mu_f = \kappa_f / L_f$, we arrive at:

$$\|g(x)\| \geq \frac{\kappa_f}{1 + \mu_f + \sqrt{1 + \mu_f}}\|x - \bar{x}\|,$$

which shows that inequality (32) is valid for the constant $\frac{1}{1+\mu_f+\sqrt{1+\mu_f}} \cdot \kappa_f$. □

Note that the functional classes we have introduced previously were obtained by relaxing the strong convexity inequalities (7)–(9) for some particular choices of x and y. The reader can find other favorable examples of relaxations of strong convexity inequalities and we believe that this paper opens a window of opportunity for algorithmic research in non-strongly convex optimization settings. In the next section we provide concrete examples of objective functions that can be found in the functional classes introduced above.

4 Functional classes in $q\mathcal{S}_{L_f,\kappa_f}(X)$, $\mathcal{G}_{L_f,\kappa_f}(X)$ and $\mathcal{F}_{L_f,\kappa_f}(X)$

We now provide examples of structured convex optimization problems whose objective function satisfies one of our relaxations of strong convexity conditions that we have introduced in the previous sections. We start first recalling some error bounds for the solutions of a system of linear equalities and inequalities. Let $A \in \mathbb{R}^{p\times n}$, $C \in \mathbb{R}^{m\times n}$ and the arbitrary norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ in $\mathbb{R}^{m+p}$ and $\mathbb{R}^n$. Given the nonempty polyhedron:

$$\mathcal{P} = \{x \in \mathbb{R}^n : \ Ax = b, \ Cx \leq d\},$$

then there exists a constant $\theta(A, C) > 0$ such that Hoffman inequality holds (for a proof of the Hoffman inequality see e.g. [11,16]):

$$\|x - \bar{x}\|_\beta \leq \theta(A, C) \left\| \begin{matrix} Ax - b \\ [Cx - d]_+ \end{matrix} \right\|_\alpha \quad \forall x \in \mathbb{R}^n,$$

where $\bar{x} = [x]_{\mathcal{P}} \equiv \arg\min_{z\in\mathcal{P}} \|z - x\|_\beta$. The constant $\theta(A, C)$ is the Hoffman constant for the polyhedron $\mathcal{P}$ with respect to the pair of norms $(\|\cdot\|_\alpha, \|\cdot\|_\beta)$. We denote $\|\cdot\|_{\alpha^*}$ the dual norm of $\|\cdot\|_\alpha$, i.e. $\|x\|_{\alpha^*} = \max\{x^T y : \ \|y\|_\alpha = 1\}$. In [17], the authors provide several estimates for the Hoffman constant. Assume that A has full row rank and define the following quantity:

$$\zeta_{\alpha,\beta}(A, C) := \min_{I\in\mathcal{J}} \min_{u,v} \left\{ \|A^T u + C^T v\|_{\beta^*} : \left\| \begin{matrix} u \\ v \end{matrix} \right\|_{\alpha^*} = 1, v_I \geq 0, v_{[m]\setminus I} = 0 \right\},$$

where $\mathcal{J} = \{I \in 2^{[m]} : \text{card } I = r - p, \ \text{rank}[A^T, \ C_I^T] = r\}$ and $r = \text{rank}[A^T, \ C^T]$. An alternative formulation of the above quantity is:

$$\frac{1}{\zeta_{\alpha,\beta}(A, C)} = \sup \left\{ \left\| \begin{matrix} u \\ v \end{matrix} \right\|_{\alpha^*} : \begin{matrix} \|A^T u + C^T v\|_{\beta^*} = 1, \text{rows of } C \\ \text{corresponding to nonzero components of } v \\ \text{and rows of } A \text{ are linearly independent} \end{matrix} \right\}. \tag{36}$$

In [17] it was proved that $\zeta_{\alpha,\beta}(A, C)^{-1}$, where $\zeta_{\alpha,\beta}(A, C)$ is defined in (36), is the Hoffman constant for the polyhedral set $\mathcal{P}$ w.r.t. the norms $(\|\cdot\|_\alpha, \|\cdot\|_\beta)$. Note that $\theta(A, C) = \zeta_{\alpha,\beta}(A, C)^{-1}$ can be arbitrarily large. For example, let us consider $\alpha = \beta = 2$ and the polyhedral set $\mathcal{P} = \{x \in \mathbb{R}^n : \ Ax = b, \ x \geq 0\}$, where A is a general matrix having two rows highly correlated, e.g. for the rows ith and jth there exists $\varepsilon > 0$ sufficiently small such that $a_i = a_j + \varepsilon e_1$, where e_1 denotes the vector with first entry 1 and the rest of entries are zero. Then, the pair $u = [1/\varepsilon \ -1/\varepsilon \ 0 \cdots 0]^T$ and $v = 0$ is feasible for (36) and hence we obtain a lower bound on the Hoffman constant $\theta(A, C) \geq \sqrt{2}/\varepsilon$. Considering the Euclidean setting ($\alpha = \beta = 2$) and the

above assumptions, then from previous discussion we have:

$$\theta(A, C) = \max_{I \in \mathcal{J}} \frac{1}{\sigma_{\min}([A^T,\ C_I^T]^T)}.$$

Under some regularity condition we can state a simpler form for $\zeta_{\alpha,2}(A, C)$. Assume that A has full row rank and that the set $\{h \in \mathbb{R}^n :\ Ah = 0,\ Ch < 0\} \neq \emptyset$, then, we have [17]:

$$\zeta_{\alpha,2}(A, C) := \min \left\{ \|A^T u + C^T v\|_2 : \left\| \begin{matrix} u \\ v \end{matrix} \right\|_{\alpha^*} = 1, v \geq 0 \right\}. \tag{37}$$

Thus, for the special case $m = 0$, i.e. there are no inequalities, we have $\zeta_{2,2}(A, 0) = \sigma_{\min}(A)$, where $\sigma_{\min}(A)$ denotes the smallest nonzero singular value of A, and the Hoffman constant is:[1]

$$\theta(A, 0) = \frac{1}{\sigma_{\min}(A)}. \tag{38}$$

We observe again that for this special case the Hoffman constant can be very large, since we can easily construct matrices having $\sigma_{\min}$ arbitrarily small: e.g. for a given $\varepsilon > 0$ sufficiently small take U and V orthogonal matrices of appropriate dimensions and Σ a diagonal matrix having the smallest diagonal entry ε; then the matrix $A = U\Sigma V^T$ has $\sigma_{\min}(A) = \varepsilon$ and $\theta(A, 0) = 1/\varepsilon$. Based on the Hoffman bound we derive below several functional classes that satisfy our proposed relaxations of strong convexity conditions. Similar results have appeared independently around the time this paper was finalized, e.g in [8,12,14,15]. Older papers that derive functional classes satisfying a local/generalized error bound type condition using the Hoffman bound are e.g. [9–11].

4.1 Composition of strongly convex function with linear map is in $q\mathcal{S}_{L_f,\kappa_f}(X)$

Let us consider the class of optimization problems (P) having the following structured form:

$$\begin{aligned} f^* = \min_x\ & f(x) \ \equiv g(Ax) \\ \text{s.t.}:\quad & x \in X \equiv \{x \in \mathbb{R}^n :\ Cx \leq d\}, \end{aligned} \tag{39}$$

[1] This result can be also proved using simple algebraic arguments. More precisely, from Courant-Fischer theorem we know that $\|Ax\| \geq \sigma_{\min}(A)\|x\|$ for all $x \in \mathrm{Im}(A^T)$. Since we assume that our polyhedron $\mathcal{P} = \{x :\ Ax = b\}$ is non-empty, then $x - [x]_{\mathcal{P}} \in \mathrm{Im}(A^T)$ for all $x \in \mathbb{R}^n$ (from KKT conditions of $\min_{z:Az=b} \|x - z\|^2$ we have that there exists μ such that $x - [x]_{\mathcal{P}} + A^T\mu = 0$). In conclusion, we get:

$$\|Ax - b\| = \|Ax - A[x]_{\mathcal{P}}\| \geq \sigma_{\min}(A)\|x - [x]_{\mathcal{P}}\| = \sigma_{\min}(A)\mathrm{dist}_2(x, \mathcal{P}) \quad \forall x \in \mathbb{R}^n.$$

i.e. the objective function is in the form $f(x) = g(Ax)$, where g is a smooth and strongly convex function and $A \in \mathbb{R}^{m\times n}$ is a nonzero general matrix. Problems of this form arise in various applications including dual formulations of linearly constrained convex problems, convex quadratic problems, routing problems in data networks, statistical regression and many others. Note that if A has full column rank, then $g(Ax)$ is strongly convex function. However, if A is rank deficient, then $g(Ax)$ is not strongly convex. We prove in the next theorem that the objective function of problem (39) belongs to the class $q\mathcal{S}_{L_f,\kappa_f}$.

Theorem 8 *Let $X = \{x \in \mathbb{R}^n : Cx \leq d\}$ be a polyhedral set, function $g : \mathbb{R}^m \to \mathbb{R}$ be σ_g-strongly convex with L_g-Lipschitz continuous gradient on X, and $A \in \mathbb{R}^{m\times n}$ be a nonzero matrix. Then, the convex function $f(x) = g(Ax)$ belongs to the class $q\mathcal{S}_{L_f,\kappa_f}(X)$, with constants $L_f = L_g\|A\|^2$ and $\kappa_f = \frac{\sigma_g}{\theta^2(A,C)}$, where $\theta(A, C)$ is the Hoffman constant for the polyhedral optimal set X^*.*

Proof The fact that f has Lipschitz continuous gradient follows immediately from the definition (1). Indeed,

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\| &= \|A^T\nabla g(Ax) - A^T\nabla g(Ay)\| \leq \|A\|\|\nabla g(Ax) - \nabla g(Ay)\| \\ &\leq \|A\|L_g\|Ax - Ay\| \leq \|A\|^2 L_g\|x - y\|.\end{aligned}$$

Thus, $L_f = L_g\|A\|^2$. Further, under assumptions of the theorem, there exists a unique pair $(t^*, T^*) \in \mathbb{R}^m \times \mathbb{R}^n$ such that the following relations hold:

$$Ax^* = t^*, \quad \nabla f(x^*) = T^* \quad \forall x^* \in X^*. \tag{40}$$

For completeness, we give a short proof of this well known fact (see also [9]): let x_1^*, x_2^* be two optimal points for the optimization problem (39). Then, from convexity of f and definition of optimal points, it follows that:

$$f\left(\frac{x_1^* + x_2^*}{2}\right) = \frac{f(x_1^*) + f(x_2^*)}{2}.$$

Since $f(x) = g(Ax)$ we get from previous relation that:

$$g\left(\frac{Ax_1^* + Ax_2^*}{2}\right) = \frac{g(Ax_1^*) + g(Ax_2^*)}{2}.$$

On the other hand using the definition of strong convexity (5) for g we have:

$$g\left(\frac{Ax_1^* + Ax_2^*}{2}\right) \leq \frac{g(Ax_1^*) + g(Ax_2^*)}{2} - \frac{\sigma_g}{8}\|Ax_1^* - Ax_2^*\|^2.$$

Combining the previous two relations, we obtain that $Ax_1^* = Ax_2^*$. Moreover, $\nabla f(x^*) = A^T\nabla g(Ax^*)$. In conclusion, Ax and the gradient of f are constant over the set of optimal solutions X^* for (39), i.e. the relations (40) hold. Moreover, we have

that $f^* = f(x^*) = g(Ax^*) = g(t^*)$ for all $x^* \in X^*$. In conclusion, the set of optimal solutions X^* is described by the following polyhedral set:

$$X^* = \{x^* : \quad Ax^* = t^*, \quad Cx^* \leq d\}.$$

Since we assume that our optimization problem (P) has at least one solution, i.e. the optimal polyhedral set X^* is non-empty, then from Hoffman inequality we have that there exists some positive constant depending on the matrices A and C describing the polyhedral set X^*, i.e. $\theta(A, C) > 0$, such that:

$$\|x - \bar{x}\| \leq \theta(A, C) \left\| \begin{bmatrix} Ax - t^* \\ [Cx - d]_+ \end{bmatrix} \right\| \quad \forall x \in \mathbb{R}^n,$$

where $\bar{x} = [x]_{X^*}$ (the projection of the vector x onto the optimal set X^*). Then, for any feasible x, i.e. x satisfying $Cx \leq d$, we have:

$$\|x - \bar{x}\| \leq \theta(A, C)\|Ax - A\bar{x}\| \quad \forall x \in X.$$

On the other hand, since g is strongly convex, it follows that:

$$g(A\bar{x}) \overset{(7)}{\geq} g(Ax) + \langle \nabla g(Ax), A\bar{x} - Ax \rangle + \frac{\sigma_g}{2}\|Ax - A\bar{x}\|^2.$$

Combining the previous two relations and keeping in mind that $f(x) = g(Ax)$ and $\nabla f(x) = A^T \nabla g(Ax)$, we obtain:

$$f^* \geq f(x) + \langle \nabla f(x), \bar{x} - x \rangle + \frac{\sigma_g}{2\theta^2(A, C)}\|x - \bar{x}\|^2 \quad \forall x \in X,$$

which proves that the quasi-strong convex inequality (11) holds for the constant $\kappa_f = \sigma_g/\theta^2(A, C)$. □

Note that we can relax the requirements for g in Theorem 8. For example, we can replace the strong convexity assumption on g with the conditions that g has unique minimizer t^* and it satisfies the quasi-strong convex condition (11) with constant $\kappa_g > 0$. Then, using the same arguments as in the proof of Theorem 8, we can show that for objective functions $f(x) = g(Ax)$ of problem (P), the optimal set is $X^* = \{x^* : \quad Ax^* = t^*, \quad Cx^* \leq d\}$ and f satisfies (11) with constant $\kappa_f = \frac{\kappa_g}{\theta^2(A,C)}$, provided that the corresponding optimal set X^* is nonempty.

Moreover, in the unconstrained case, that is $X = \mathbb{R}^n$, and for objective function $f(x) = g(Ax)$, we get from (38) the following expression for the quasi-strong convexity constant:

$$\kappa_f = \sigma_g \sigma_{\min}^2(A). \tag{41}$$

Below we prove two extensions that belong to other functional classes we have introduced in this paper.

4.2 Composition of strongly convex function with linear map plus a linear term for $X = \mathbb{R}^n$ is in $\mathcal{G}_{L_f,\kappa_f}(X)$

Let us now consider the class of unconstrained optimization problems (P), i.e. $X = \mathbb{R}^n$, having the form:

$$f^* = \min_{x \in \mathbb{R}^n} \; f(x) \; \equiv g(Ax) + c^T x, \tag{42}$$

i.e. the objective function is in the form $f(x) = g(Ax) + c^T x$, where g is a smooth and strongly convex function, $A \in \mathbb{R}^{m \times n}$ is a nonzero general matrix and $c \in \mathbb{R}^n$. We prove in the next theorem that this type of objective function for problem (42) belongs to the class $\mathcal{G}_{L_f,\kappa_f}$:

Theorem 9 *Under the same assumptions as in Theorem 8 with $X = \mathbb{R}^n$, the objective function of the form $f(x) = g(Ax) + c^T x$ belongs to the class $\mathcal{G}_{L_f,\kappa_f}(X)$, with constants $L_f = L_g \|A\|^2$ and $\kappa_f = \frac{\sigma_g}{\theta^2(A,0)}$, where $\theta(A, 0)$ is the Hoffman constant for the optimal set X^*.*

Proof Since g is σ_g-strongly convex and with L_g-Lipschitz continuous gradient, then by the same reasoning as in the proof of Theorem 8 we get that there exists unique vector t^* such that $Ax^* = t^*$ for all $x^* \in X^*$. Similarly, there exists unique scalar s^* such that $c^T x^* = s^*$ for all $x^* \in X^*$. Indeed, for $x_1^*, x_2^* \in X^*$ we have:

$$f^* = g(t^*) + c^T x_1^* = g(t^*) + c^T x_2^*,$$

which implies that $c^T x_1^* = c^T x_2^*$. On the other hand, since problem (P) is unconstrained, for any $x^* \in X^*$ we have:

$$0 = \nabla f(x^*) = A^T \nabla g(t^*) + c,$$

which implies that $c^T x^* = -(\nabla g(t^*))^T A x^* = -(\nabla g(t^*))^T t^*$. Therefore, the set of optimal solutions X^* is described in this case by the following polyhedron:

$$X^* = \{x^* : \; Ax^* = t^*\}.$$

Then, there exists $\theta(A, 0) > 0$ such that the Hoffman inequality holds:

$$\|x - \bar{x}\| \leq \theta(A, 0) \|Ax - A\bar{x}\| \quad \forall x \in \mathbb{R}^n.$$

From the previous inequality and strong convexity of g, we have:

$$\begin{aligned} \frac{\sigma_g}{\theta^2(A,0)} \|x - \bar{x}\|^2 \leq \sigma_g \|Ax - A\bar{x}\|^2 &\overset{(8)}{\leq} \langle \nabla g(Ax) - \nabla g(A\bar{x}), Ax - A\bar{x} \rangle \\ &= \langle A^T \nabla g(Ax) + c - A^T \nabla g(A\bar{x}) - c, x - \bar{x} \rangle \\ &= \langle \nabla f(x) - \nabla f(\bar{x}), x - \bar{x} \rangle. \end{aligned}$$

Finally, we conclude that the inequality on the variation of gradients (18) holds with constant $\kappa_f = \frac{\sigma_g}{\theta^2(A,0)}$. □

4.3 Composition of strongly convex function with linear map plus a linear term is in $\mathcal{F}_{L_f,\kappa_f}(X_M)$

Finally, let us now consider the class of optimization problems (P) of the form:

$$\begin{aligned} f^* = \min_x f(x) \ &\equiv g(Ax) + c^T x \\ \text{s.t.:} \quad x \in X &\equiv \{x \in \mathbb{R}^n : \ Cx \leq d\}, \end{aligned} \tag{43}$$

i.e. the objective function is in the form $f(x) = g(Ax) + c^T x$, where g is a smooth and strongly convex function, $A \in \mathbb{R}^{m\times n}$ is a nonzero matrix and $c \in \mathbb{R}^n$. We now prove that the objective function of problem (43) belongs to class $\mathcal{F}_{L_f,\kappa_f}$, provided that some boundedness assumption is imposed on f.

Theorem 10 *Under the same assumptions as in Theorem 8, the objective function $f(x) = g(Ax) + c^T x$ belongs to the class $\mathcal{F}_{L_f,\kappa_f}(X_M)$ for any constant $M > 0$ such that $X_M = \{x : \ x \in X, \ f(x) - f^* \leq M\}$, with constants $L_f = L_g\|A\|^2$ and $\kappa_f = \frac{\sigma_g}{\theta^2(A,c,C)\left(1+M\sigma_g+2c_g^2\right)}$, where $\theta(A, c, C)$ is the Hoffman constant for the polyhedral optimal set X^* and $c_g = \|\nabla g(Ax^*)\|$, with $x^* \in X^*$.*

Proof From the proof of Theorem 9 it follows that there exist unique t^* and s^* such that the optimal set of (43) is given as follows:

$$X^* = \{x^* : \ Ax^* = t^*, \ c^T x^* = s^*, \ Cx^* \leq d\}.$$

From Hoffman inequality we have that there exists some positive constant depending on the matrices A, C and c describing the polyhedral set X^*, i.e. $\theta(A, C, c) > 0$, such that:

$$\|x - \bar{x}\| \leq \theta(A, c, C) \left\| \begin{bmatrix} Ax - t^* \\ c^T x - s^* \\ [Cx - d]_+ \end{bmatrix} \right\| \quad \forall x \in \mathbb{R}^n,$$

where recall that $\bar{x} = [x]_{X^*}$. Then, for any feasible x, i.e. satisfying $Cx \leq d$, we have:

$$\|x - \bar{x}\|^2 \leq \theta^2(A, c, C) \left(\|Ax - A\bar{x}\|^2 + (c^T x - c^T \bar{x})^2 \right) \quad \forall x \in X. \tag{44}$$

Since $f(x) = g(Ax) + c^T x$ and g is strongly convex, it follows from (7) that:

$$g(Ax) - g(A\bar{x}) \geq \langle \nabla g(A\bar{x}), Ax - A\bar{x} \rangle + \frac{\sigma_g}{2}\|Ax - A\bar{x}\|^2$$
$$= \langle A^T \nabla g(A\bar{x}) + c, x - \bar{x} \rangle - \langle c, x - \bar{x} \rangle + \frac{\sigma_g}{2}\|Ax - A\bar{x}\|^2$$
$$= \langle \nabla f(\bar{x}), x - \bar{x} \rangle - \langle c, x - \bar{x} \rangle + \frac{\sigma_g}{2}\|Ax - A\bar{x}\|^2.$$

Using that $\langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq 0$ for all $x \in X$, and definition of f, we obtain:

$$f(x) - f^* \geq \frac{\sigma_g}{2}\|Ax - A\bar{x}\|^2 \quad \forall x \in X. \tag{45}$$

It remains to bound $(c^T x - c^T \bar{x})^2$. It is easy to notice that $\theta(A, c, C) \geq 1/\|c\|$. We also observe that:

$$c^T x - c^T \bar{x} = \langle \nabla f(\bar{x}), x - \bar{x} \rangle - \langle \nabla g(A\bar{x}), Ax - A\bar{x} \rangle.$$

Since $f(x) - f^* \geq \langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq 0$ for all $x \in X$, then we obtain:

$$|c^T x - c^T \bar{x}| \leq f(x) - f^* + \|\nabla g(A\bar{x})\| \, \|Ax - A\bar{x}\|,$$

and then using inequality $(\alpha + \beta)^2 \leq 2\alpha^2 + 2\beta^2$ and considering $f(x) - f^* \leq M$, $c_g = \|\nabla g(t^*)\|$ and (45), we get:

$$(c^T x - c^T \bar{x})^2 \leq 2(f(x) - f^*)^2 + 2c_g^2\|Ax - A\bar{x}\|^2$$
$$\leq \left(2M + \frac{4c_g^2}{\sigma_g}\right)\left(f(x) - f^*\right) \quad \forall x \in X, \ f(x) - f^* \leq M.$$

Finally, we conclude that:

$$\|x - \bar{x}\|^2 \leq \frac{2\theta^2(A, c, C)}{\sigma_g}\left(1 + M\sigma_g + 2c_g^2\right)\left(f(x) - f^*\right) \quad \forall x \in X, \ f(x) - f^* \leq M.$$

This proves the statement of the theorem. □

Typically, for feasible descent methods we take $M = f(x^0) - f^*$ in the previous theorem, where x^0 is the starting point of the method. Moreover, if X is bounded, then there exists always M such that $f(x) - f^* \leq M$ for all $x \in X$. Note that the requirement $f(x) - f^* \leq M$ for having a second order growth inequality (23) for f is necessary, as shown in the following example:

Example 1 Let us consider problem (P) in the form (43) given by:

$$\min_{x \in \mathbb{R}^2_+} \frac{1}{2}x_1^2 + x_2,$$

which has $X^* = \{0\}$ and $f^* = 0$. Clearly, there is no constant $\kappa_f < \infty$ such that the following inequality to be valid:

$$f(x) \geq \frac{\kappa_f}{2}\|x\|^2 \quad \forall x \geq 0,$$

for example we can take $x_1 = 0$ and $x_2 \to +\infty$. However, for any $M > 0$ there exists $\kappa_f(M) < \infty$ satisfying the above inequality for all $x \geq 0$ with $f(x) \leq M$. For example, we can take:

$$\kappa_f(M) = \min\left(1, \frac{1}{M}\right) \Rightarrow \mu_f(M) = \frac{1}{M} \text{ for } M \geq 1, \text{ otherwise } \mu_f(M) = 1.$$

Note that for this example the Hoffman constant is $\theta(A, c, C) = \frac{1}{\|c\|} = 1$.

□

In the sequel we analyze the convergence rate of several first order methods for solving convex constrained optimization problem (P) having the objective function in one of the functional classes introduced in this paper.

5 Linear convergence of first order methods

We show in the next sections that a broad class of first order methods, covering important particular algorithms, such as projected gradient, fast gradient, random/cyclic coordinate descent, extragradient descent and matrix splitting, have linear convergence rates on optimization problems (P), whose objective function satisfies one of the non-strongly convex conditions given above.

5.1 Projected gradient method (GM)

In this section we consider the projected gradient algorithm with variable step size:

Algorithm (GM)

Given $x^0 \in X$ for $k \geq 1$ do:

1. Compute $x^{k+1} = \left[x^k - \alpha_k \nabla f(x^k)\right]_X$

where α_k is a step size such that $\alpha_k \in [\bar{L}_f^{-1}, L_f^{-1}]$, with $\bar{L}_f \geq L_f$.

5.1.1 Linear convergence of (GM) for $q\mathcal{S}_{L_f,\kappa_f}$

Let us show that the projected gradient method converges linearly on optimization problems (P) whose objective functions belong to the class $q\mathcal{S}_{L_f,\kappa_f}$.

Theorem 11 *Let the optimization problem (P) have the objective function belonging to the class qS_{L_f,κ_f}. Then, the sequence x^k generated by the projected gradient method (GM) with constant step size $\alpha_k = 1/L_f$ on (P) converges linearly to some optimal point in X^* with the rate:*

$$\|x^k - \bar{x}^k\|^2 \leq \left(\frac{1-\mu_f}{1+\mu_f}\right)^k \|x^0 - \bar{x}^0\|^2, \quad \text{where} \quad \mu_f = \frac{\kappa_f}{L_f}. \tag{46}$$

Proof From Lipschitz continuity of the gradient of f given in (2) we have:

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L_f}{2}\|x^{k+1} - x^k\|^2. \tag{47}$$

The optimality conditions for x^{k+1} are:

$$\langle x^{k+1} - x^k + \alpha_k \nabla f(x^k), x - x^{k+1} \rangle \geq 0 \quad \forall x \in X. \tag{48}$$

Taking $x = x^k$ in (48) and replacing the corresponding expression in (47), we get:

$$f(x^{k+1}) \leq f(x^k) + (\frac{L_f}{2} - \frac{1}{\alpha_k})\|x^{k+1} - x^k\|^2 \overset{\alpha_k \leq L_f^{-1}}{\leq} f(x^k) - \frac{L_f}{2}\|x^{k+1} - x^k\|^2.$$

Further, we have:

$$\begin{aligned}
\|x^{k+1} - \bar{x}^k\|^2 &= \|x^k - \bar{x}^k\|^2 + 2\langle x^k - \bar{x}^k, x^{k+1} - x^k \rangle + \|x^{k+1} - x^k\|^2 \\
&= \|x^k - \bar{x}^k\|^2 + 2\langle x^{k+1} - \bar{x}^k, x^{k+1} - x^k \rangle - \|x^{k+1} - x^k\|^2 \\
&\overset{(48)}{\leq} \|x^k - \bar{x}^k\|^2 + 2\alpha_k \langle \nabla f(x^k), \bar{x}^k - x^{k+1} \rangle - \|x^{k+1} - x^k\|^2 \\
&= \|x^k - \bar{x}^k\|^2 + 2\alpha_k \langle \nabla f(x^k), \bar{x}^k - x^k \rangle + 2\alpha_k \langle \nabla f(x^k), x^k - x^{k+1} \rangle \\
&\qquad - \|x^{k+1} - x^k\|^2 \\
&\overset{(11)}{\leq} \|x^k - \bar{x}^k\|^2 + 2\alpha_k \Big(f^* - f(x^k) - \frac{\kappa_f}{2}\|x^k - \bar{x}^k\|^2 \Big) \\
&\quad - 2\alpha_k \Big(\langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{2\alpha_k}\|x^{k+1} - x^k\|^2 \Big) \\
&= (1 - \alpha_k \kappa_f)\|x^k - \bar{x}^k\|^2 + 2\alpha_k f^* \\
&\quad - 2\alpha_k \Big(f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{2\alpha_k}\|x^{k+1} - x^k\|^2 \Big) \\
&\overset{L_f \leq 1/\alpha_k}{\leq} (1 - \alpha_k \kappa_f)\|x^k - \bar{x}^k\|^2 + 2\alpha_k f^* \\
&\quad - 2\alpha_k \Big(f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L_f}{2}\|x^{k+1} - x^k\|^2 \Big) \\
&\overset{(47)}{\leq} (1 - \alpha_k \kappa_f)\|x^k - \bar{x}^k\|^2 - 2\alpha_k (f(x^{k+1}) - f^*).
\end{aligned}$$

Since (11) holds for the function f, then from Theorem 4 we also have that (23) holds and therefore $f(x^{k+1}) - f^* \geq \frac{\kappa_f}{2}\|x^{k+1} - \bar{x}^{k+1}\|^2$. Combining the last inequality with the previous one and taking into account that $\|x^{k+1} - \bar{x}^{k+1}\| \leq \|x^{k+1} - \bar{x}^k\|$, we get:

$$\|x^{k+1} - \bar{x}^{k+1}\|^2 \leq (1 - \alpha_k \kappa_f)\|x^k - \bar{x}^k\|^2 - \alpha_k \kappa_f \|x^{k+1} - \bar{x}^{k+1}\|^2,$$

or equivalently

$$\|x^{k+1} - \bar{x}^{k+1}\|^2 \leq \frac{1 - \alpha_k \kappa_f}{1 + \alpha_k \kappa_f} \cdot \|x^k - \bar{x}^k\|^2. \tag{49}$$

However, the best decrease is obtained for the constant step size $\alpha_k = 1/L_f$ and using the definition of the condition number $\mu_f = \kappa_f / L_f$, we get:

$$\|x^{k+1} - \bar{x}^{k+1}\|^2 \leq \frac{1 - \mu_f}{1 + \mu_f} \cdot \|x^k - \bar{x}^k\|^2.$$

This proves our statement. □

Based on Theorem 11 we can easily derive linear convergence for the projected gradient algorithm (GM) in terms of the function values:

$$\begin{aligned}
f(x^{k+1}) &\overset{(47)}{\leq} f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L_f}{2}\|x^{k+1} - x^k\|^2 \\
&\overset{L_f \leq 1/\alpha_k}{\leq} \min_{x \in X} f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2\alpha_k}\|x^k - x\|^2 \\
&\leq \min_{x \in X} f(x) + \frac{1}{2\alpha_k}\|x^k - x\|^2 \leq f(\bar{x}^k) + \frac{1}{2\alpha_k}\|x^k - \bar{x}^k\|^2 \\
&\overset{(46)}{\leq} f^* + \frac{\bar{L}_f}{2}\left(\frac{1 - \mu_f}{1 + \mu_f}\right)^k \|x^0 - \bar{x}^0\|^2.
\end{aligned}$$

Finally, the best convergence rate is obtained for constant step size $\alpha_k = 1/L_f$:

$$f(x^k) - f^* \overset{(52)}{\leq} \frac{L_f \|x^0 - \bar{x}^0\|^2}{2}\left(\frac{1 - \mu_f}{1 + \mu_f}\right)^{k-1} \qquad \forall k \geq 1. \tag{50}$$

However, this rate of convergence vanishes as $\mu_f \to 0$. For simplicity, let us assume constant step size $\alpha_k = 1/L_f$, and then, using that (GM) is a descent method, i.e. $f(x^k) - f^* \leq f(x^{k-j}) - f^*$ for all $j < k$ and iterating the main inequality from the proof of Theorem 11, we obtain:

$$\begin{aligned}
\|x^k - \bar{x}^k\|^2 &\leq (1 - \mu_f)\|x^{k-1} - \bar{x}^{k-1}\|^2 - \frac{2}{L_f}\left(f(x^k) - f^*\right) \\
&\leq (1 - \mu_f)^k \|x^0 - \bar{x}^0\|^2 - \frac{2}{L_f}\sum_{j=0}^{k}(1 - \mu_f)^j (f(x^{k-j}) - f^*)
\end{aligned}$$

$$\leq (1-\mu_f)^k \|x^0 - \bar{x}^0\|^2 - \frac{2}{L_f}\left(f(x^k) - f^*\right)\sum_{j=0}^{k}(1-\mu_f)^j.$$

Finally, we get linear convergence in terms of the function values:

$$f(x^k) - f^* \leq \frac{L_f\|x^0 - \bar{x}^0\|^2}{2} \cdot \frac{\mu_f}{(1-\mu_f)^{-k} - 1}. \tag{51}$$

Since $(1+\alpha)^k \to 1 + \alpha k$ as $\alpha \to 0$, then we see that:

$$\frac{\mu_f}{(1-\mu_f)^{-k} - 1} \leq \frac{1}{k} \quad \text{as} \quad \mu_f \to 0.$$

Thus, from (51), as $\mu_f \to 0$, we recover the classical sublinear rate of convergence for (GM) when the objective function is only L_f-smooth:

$$f(x^k) - f^* \leq \frac{L_f\|x^0 - \bar{x}^0\|^2}{2k} \quad \text{as} \quad \mu_f \to 0.$$

5.1.2 Linear convergence of (GM) for $\mathcal{F}_{L_f,\kappa_f}$

We now show that the projected gradient method converges linearly on optimization problems (P) whose objective functions belong to the class $\mathcal{F}_{L_f,\kappa_f}$.

Theorem 12 *Let optimization problem (P) have objective function belonging to the class $\mathcal{F}_{L_f,\kappa_f}$. Then, the sequence x^k generated by the projected gradient method (GM) with constant step size $\alpha_k = 1/L_f$ on (P) converges linearly to some optimal point in X^* with the rate:*

$$\|x^k - \bar{x}^k\|^2 \leq \left(\frac{1}{1+\mu_f}\right)^k \|x^0 - \bar{x}^0\|^2, \quad \textit{where} \quad \mu_f = \frac{\kappa_f}{L_f}. \tag{52}$$

Proof Using similar arguments as in the previous Theorem 11, we have:

$$\begin{aligned}
\|x^{k+1} - x\|^2 &= \|x^k - x\|^2 + 2\langle x^k - x, x^{k+1} - x^k\rangle + \|x^{k+1} - x^k\|^2 \\
&= \|x^k - x\|^2 + 2\langle x^{k+1} - x, x^{k+1} - x^k\rangle - \|x^{k+1} - x^k\|^2 \\
&\overset{(48)}{\leq} \|x^k - x\|^2 + 2\alpha_k\langle \nabla f(x^k), x - x^{k+1}\rangle - \|x^{k+1} - x^k\|^2 \\
&\leq \|x^k - x\|^2 - 2\alpha_k\Big(\langle \nabla f(x^k), x^{k+1} - x\rangle + \frac{L_f}{2}\|x^{k+1} - x^k\|^2 \\
&\qquad + (\frac{1}{2\alpha_k} - \frac{L_f}{2})\|x^{k+1} - x^k\|^2\Big) \\
&= \|x^k - x\|^2 + (L_f\alpha_k - 1)\|x^{k+1} - x^k\|^2
\end{aligned}$$

$$
\begin{aligned}
&- 2\alpha_k\left(\langle \nabla f(x^k), x_k - x\rangle + \langle \nabla f(x^k), x^{k+1} - x_k\rangle + \frac{L_f}{2}\|x^{k+1} - x^k\|^2\right)\\
&\overset{(47)}{\leq} \|x^k - x\|^2 + (L_f\alpha_k - 1)\|x^{k+1} - x^k\|^2\\
&\qquad + 2\alpha_k(f(x) - f(x^k)) + 2\alpha_k(f(x^k) - f(x^{k+1}))\\
&\overset{\alpha_k \leq L_f^{-1}}{\leq} \|x^k - x\|^2 - 2\alpha_k(f(x^{k+1}) - f(x)) \qquad \forall x \in X.
\end{aligned}
$$

Taking now in the previous relations $x = \bar{x}^k$, using $\|x^{k+1} - \bar{x}^{k+1}\| \leq \|x^{k+1} - \bar{x}^k\|$ and the quadratic functional growth of f (23), we get:

$$
\|x^{k+1} - \bar{x}^{k+1}\|^2 \overset{(23)}{\leq} \|x^k - \bar{x}^k\|^2 - \kappa_f\alpha_k\|x^{k+1} - \bar{x}^{k+1}\|^2
$$

or equivalently

$$
\|x^{k+1} - \bar{x}^{k+1}\|^2 \leq \frac{1}{1 + \kappa_f\alpha_k}\|x^k - \bar{x}^k\|^2. \tag{53}
$$

However, the best decrease is obtained for the constant step size $\alpha_k = 1/L_f$ and using the definition of the condition number $\mu_f = \kappa_f/L_f$, we get:

$$
\|x^{k+1} - \bar{x}^{k+1}\|^2 \leq \frac{1}{1 + \mu_f}\|x^k - \bar{x}^k\|^2.
$$

Thus, we have obtained the linear convergence rate for (GM) with constant step size $\alpha_k = 1/L_f$ from the theorem. □

Using similar arguments as for (50) and combining with (53) we can also derive linear convergence of (GM) in terms of the function values:

$$
f(x^{k+1}) - f^* \leq \frac{1}{2\alpha_k}\|x^k - \bar{x}^k\|^2 \overset{(53)}{\leq} \frac{1}{2\alpha_k}\left(\frac{1}{1 + \kappa_f\alpha_k}\right)\|x^{k-1} - \bar{x}^{k-1}\|^2,
$$

and the best convergence rate is obtained for constant step size $\alpha_k = 1/L_f$:

$$
f(x^k) - f^* \overset{(52)}{\leq} \frac{L_f\|x^0 - \bar{x}^0\|^2}{2}\left(\frac{1}{1 + \mu_f}\right)^{k-1} \qquad \forall k \geq 1. \tag{54}
$$

However, this rate of convergence vanishes as $\mu_f \to 0$. We can interpolate between the right hand side terms in (4) and (54) to obtain convergence rates in terms of function values of the form:

$$
f(x^k) - f^* \leq \frac{L_f\|x^t - \bar{x}^t\|^2}{2(k-t)} \leq \frac{L_f\|x^0 - \bar{x}^0\|^2}{2(k-t)}\frac{1}{(1 + \mu_f)^t} \quad \forall t = 0 : k-1,
$$

or equivalently

$$f(x^k) - f^* \leq \frac{L_f \|x^0 - \bar{x}^0\|^2}{2} \min_{t=0:k-1} \frac{1}{(1+\mu_f)^t (k-t)}.$$

Thus, we recover again the classical sublinear rate of convergence for (GM) as $\mu_f \to 0$, when the objective function is only L_f-smooth. Finally, in the next theorem we establish necessary and sufficient conditions for linear convergence of the gradient method (GM).

Theorem 13 *On the class of optimization problems (P) the sequence generated by the gradient method (GM) with constant step size is converging linearly to some optimal point in X^* if and only if the objective function f satisfies the quadratic functional growth (23), i.e f belongs to the functional class $\mathcal{F}_{L_f,\kappa_f}$.*

Proof The fact that linear convergence of the gradient method implies f satisfying the second order growth property (23) follows from Theorem 5. The other implication follows from Theorem 12, Eq. (53). □

5.2 Fast gradient method (FGM)

In this section we consider the following fast gradient algorithm, which is a version of Nesterov's optimal gradient method [3]:

Algorithm (FGM)

Given $x^0 = y^0 \in X$, for $k \geq 1$ do:

1. Compute $x^{k+1} = \left[y^k - \frac{1}{L_f} \nabla f(y^k) \right]_X$ and
2. $y^{k+1} = x^{k+1} + \beta_k \left(x^{k+1} - x^k \right)$

for appropriate choice of the parameter $\beta_k > 0$ for all $k \geq 0$.

5.2.1 Linear convergence of (FGM) for $q\mathcal{S}_{L_f,\kappa_f}$.

When the objective function $f \in q\mathcal{S}_{L_f,\kappa_f}(X)$ we take the following expression for the parameter β_k:

$$\beta_k = \frac{\sqrt{L_f} - \sqrt{\kappa_f}}{\sqrt{L_f} + \sqrt{\kappa_f}} \quad \forall k \geq 0.$$

First of all we can easily observe that if $f \in q\mathcal{S}_{L_f,\kappa_f}(X)$, then the gradient mapping $g(x)$ satisfies the following inequality:

$$f^* \geq f(x^+) + \langle g(x), \bar{x} - x \rangle + \frac{1}{2L_f} \|g(x)\|^2 + \frac{\kappa_f}{2} \|\bar{x} - x\| \equiv q_{L_f,\kappa_f}(\bar{x}, x) \quad (55)$$

for all $x \in \mathbb{R}^n$ (recall that $\bar{x} = [x]_{X^*}$ and $x^+ = [x - 1/L_f \nabla f(x)]_X$). The convergence proof follows similar steps as in [3] [Section 2.2.4].

Lemma 1 *Let optimization problem (P) have the objective function f belonging to the class $q\mathcal{S}_{L_f,\kappa_f}$ and an arbitrary sequence $\{y^k\}_{k\geq 0}$ satisfying $\bar{y}^k = [y^k]_{X^*} = y^*$ for all $k \geq 0$. Define an initial function:*

$$\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}\|x - v^0\|^2, \quad \textit{where} \quad \gamma_0 = \kappa_f, \quad v^0 = y^0 \textit{ and } \phi_0^* = f(y^0),$$

and a sequence $\{\alpha_k\}_{k\geq 0}$ satisfying $\alpha_k \in (0,\ 1)$. Then, the following two sequences, iteratively defined as:

$$\begin{aligned}
&\lambda_{k+1} = (1-\alpha_k)\lambda_k, \quad \textit{with} \quad \lambda_0 = 1,\\
&\phi_{k+1}(x) = (1-\alpha_k)\phi_k(x)\\
&\qquad + \alpha_k\left(f(x^{k+1}) + \frac{1}{2L_f}\|g(y^k)\|^2 + \langle g(y^k), x - y^k\rangle + \frac{\kappa_f}{2}\|x - y^k\|^2\right),
\end{aligned} \tag{56}$$

where $x^0 = y^0$ and $x^{k+1} = \left[y^k - \frac{1}{L_f}\nabla f(y^k)\right]_X$, satisfy the following property:

$$\phi_k(y^*) \leq (1-\lambda_k) f^* + \lambda_k \phi_0(y^*) \quad \forall k \geq 0. \tag{57}$$

Proof We prove this statement by induction. Since $\lambda_0 = 1$, we observe that:

$$\phi_0(y^*) = (1-\lambda_0) f^* + \lambda_0 \phi_0(y^*).$$

Assume that the following inequality is valid:

$$\phi_k(y^*) \leq (1-\lambda_k) f^* + \lambda_k \phi_0(y^*), \tag{58}$$

then we have:

$$\begin{aligned}
\phi_{k+1}(y^*) = \phi_{k+1}(\bar{y}^k) &= (1-\alpha_k)\phi_k(\bar{y}^k) + \alpha_k q_{L_f,\kappa_f}(\bar{y}^k, y^k)\\
&\overset{(55)}{\leq} (1-\alpha_k)\phi_k(\bar{y}^k) + \alpha_k f^*\\
&= [1 - (1-\alpha_k)\lambda_k] f^* + (1-\alpha_k)\left(\phi_k(\bar{y}^k) - (1-\lambda_k) f^*\right)\\
&\overset{\bar{y}^k = y^* + (58)}{\leq} (1-\lambda_{k+1}) f^* + \lambda_{k+1}\phi_0(y^*).
\end{aligned}$$

which proves our statement. □

Lemma 2 *Under the same assumptions as in Lemma 1 and assuming also that the sequence* $\{x_k\}_{k\geq 0}$, *defined as* $x^0 = y^0$ *and* $x^{k+1} = \left[y^k - \frac{1}{L_f}\nabla f(y^k)\right]_X$, *satisfies:*

$$f(x^k) \leq \phi_k^* = \min_{x\in\mathbb{R}^n} \phi_k(x) \qquad \forall k \geq 0, \tag{59}$$

then we obtain the following convergence:

$$f(x^k) - f^* \leq \lambda_k \left(f(x^0) - f^* + \frac{\gamma_0}{2}\|y^* - y^0\|\right). \tag{60}$$

Proof Indeed we have:

$$\begin{aligned} f(x^k) - f^* \leq \phi_k^* - f^* &= \min_{x\in\mathbb{R}^n} \phi_k(x) - f^* \leq \phi_k(y^*) - f^* \\ &\overset{(57)}{\leq} (1-\lambda_k) f^* + \lambda_k \phi_0(y^*) - f^* = \lambda_k \left(\phi_0(y^*) - f^*\right), \end{aligned}$$

which proves the statement of the lemma. □

Theorem 14 *Under the same assumptions as in Lemma 1, the sequence* x^k *generated by fast gradient method (FGM) with constant parameter* $\beta_k = (\sqrt{L_f} - \sqrt{\kappa_f})/(\sqrt{L_f} + \sqrt{\kappa_f})$ *converges linearly in terms of function values with the rate:*

$$f(x^k) - f^* \leq \left(1 - \sqrt{\mu_f}\right)^k \cdot 2\left(f(x^0) - f^*\right), \quad \text{where} \quad \mu_f = \frac{\kappa_f}{L_f}, \tag{61}$$

provided that all iterates y^k *produce the same projection*[2] *onto optimal set* X^*.

Proof Let us consider $x^0 = y^0 = v^0 \in X$. Further, for the sequence of functions $\phi_k(x)$ as defined in (56) take $\alpha_k = \sqrt{\mu_f} \in (0,\ 1)$ for all $k \geq 0$ and denote $\alpha = \sqrt{\mu_f}$. First, we need to show that the method (FGM) defined above generates a sequence x^k satisfying $\phi_k^* \geq f(x^k)$. Assuming that $\phi_k(x)$ has the following two properties:

$$\phi_k(x) = \phi_k^* + \frac{\kappa_f}{2}\|x - v^k\|^2 \quad \text{and} \quad \phi_k^* \geq f(x^k),$$

where $\phi_k^* = \min_{x\in\mathbb{R}^n} \phi_k(x)$ and $v^k = \arg\min_{x\in\mathbb{R}^n} \phi_k(x)$, then we will show that $\phi_{k+1}(x)$ has similar properties. First of all, from the definition of $\phi_{k+1}(x)$, we get:

$$\nabla^2 \phi_{k+1}(x) = \left((1-\alpha)\kappa_f + \alpha\kappa_f\right) I_n = \kappa_f I_n,$$

i.e. $\phi_{k+1}(x)$ is also a quadratic function of the same form as $\phi_k(x)$:

$$\phi_{k+1}(x) = \phi_{k+1}^* + \frac{\kappa_f}{2}\|x - v^{k+1}\|^2,$$

[2] See Remark 1 below for an example satisfying this condition.

where the expression of $v^{k+1} = \arg\min_{x\in\mathbb{R}^n} \phi_{k+1}(x)$ is obtained from the equation $\nabla\phi_{k+1}(x) = 0$, which leads to:

$$v^{k+1} = \frac{1}{\kappa_f}\left((1-\alpha)\kappa_f v^k + \alpha\kappa_f y^k - \alpha g(y^k)\right).$$

Evaluating ϕ_{k+1} in y^k leads to:

$$\begin{aligned}\phi^*_{k+1} + \frac{\kappa_f}{2}\|y^k - v^{k+1}\|^2 =& (1-\alpha)\left(\phi^*_k + \frac{\kappa_f}{2}\|y^k - v^k\|^2\right)\\ &+ \alpha\left(f(x^{k+1}) + \frac{1}{2L_f}\|g(y^k)\|^2\right).\end{aligned}$$

On the other hand, we have:

$$v^{k+1} - y^k = \frac{1}{\kappa_f}\left(\kappa_f(1-\alpha)(v^k - y^k) - \alpha g(y^k)\right).$$

If we substitute this expression above, we obtain:

$$\begin{aligned}\phi^*_{k+1} = (1-\alpha)\phi^*_k + \alpha f(x^{k+1}) + \left(\frac{\alpha}{2L_f} - \frac{\alpha^2}{2\kappa_f}\right)\|g(y^k)\|^2\\ + \alpha(1-\alpha)\left(\frac{\kappa_f}{2}\|y^k - v^k\|^2 + \langle g(y^k), v^k - y^k\rangle\right).\end{aligned}$$

Using the main property of the gradient mapping (30), valid for functions with Lipschitz continuous gradient, we have:

$$\phi^*_k \geq f(x^k) \geq f(x^{k+1}) + \langle g(y^k), x^k - y^k\rangle + \frac{1}{2L_f}\|g(y^k)\|^2.$$

Substituting this inequality in the previous one we get:

$$\phi^*_{k+1} \geq f(x^{k+1}) + \left(\frac{1}{2L_f} - \frac{\alpha^2}{2\kappa_f}\right)\|g(y^k)\|^2 + (1-\alpha)\langle g(y^k), \alpha(v^k - y^k) + x^k - y^k\rangle.$$

Since $\alpha = \sqrt{\mu_f}$, then $\frac{1}{2L_f} - \frac{\alpha^2}{2\kappa_f} = 0$. Moreover, we have the freedom to choose y^k, which is obtained from the condition $\alpha(v^k - y^k) + x^k - y^k = 0$:

$$y^k = \frac{1}{1+\alpha}(\alpha v^k + x^k).$$

Then, we can conclude that $\phi^*_{k+1} \geq f(x^{k+1})$. Moreover, replacing the expression of y^k in v^{k+1} leads to the conclusion that we can eliminate the sequence v^k since it can be expressed as: $v^{k+1} = x^k + \frac{1}{\alpha}(x^{k+1} - x^k)$. Then, we find that y^{k+1} has the

expression as in our scheme (FGM) above with $\beta_k = (\sqrt{L_f} - \sqrt{\kappa_f})/(\sqrt{L_f} + \sqrt{\kappa_f})$. Using, now Lemmas 1 and 2 we get the convergence rate from (61) (we also use that $\frac{\kappa_f}{2}\|x^0 - \bar{x}^0\|^2 \leq f(x^0) - f^*$). □

Remark 1 For unconstrained problem $\min_{x\in\mathbb{R}^n} g(Ax)$, the gradient in some point y is given by $A^T\nabla g(Ay) \in \text{Range}(A^T)$. Then, the method (FGM) generates in this case a sequence y^k of the form:

$$y^k = y^0 + A^T z^k, \qquad z^k \in \mathbb{R}^m \quad \forall k \geq 0.$$

Moreover, for this problem the optimal set $X^* = \{x : \ Ax = t^*\}$ and the projection onto this affine subspace is given by:

$$[\,\cdot\,]_{X^*} = \left(I_n - A^T(AA^T)^{-1}A\right)(\cdot) + A^T(AA^T)^{-1}t^*.$$

In conclusion, all vectors y^k generated by algorithm (FGM) produce the same projection onto the optimal set X^*:

$$\bar{y}^k = y^0 - A^T(AA^T)^{-1}Ay^0 + A^T(AA^T)^{-1}t^* \quad \forall k \geq 0,$$

i.e. the assumptions of Theorem 14 are valid for this optimization problem. □

Note that the fast gradient method (FGM) requires knowledge of the condition number, in particular κ_f. However, a search procedure for κ_f can be implemented here as in the paper [18]. We skip the details of the complexity analysis of such strategy and refer to [18] [Section 5.3] for a detailed analysis.

5.2.2 Linear convergence of restart (FGM) for $\mathcal{F}_{L_f,\kappa_f}$

It is known that for the convex optimization problem (P), whose objective function f has Lipschitz continuous gradient, and for the choice:

$$\beta_k = \frac{\theta_k - 1}{\theta_{k+1}}, \quad \text{with} \quad \theta_1 = 1 \text{ and } \theta_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k^2}}{2},$$

the algorithm (FGM) has the following convergence rate [3,19]:

$$f(x^k) - f^* \leq \frac{2L_f\|x^0 - \bar{x}^0\|^2}{(k+1)^2} \qquad \forall k > 0. \tag{62}$$

We will show next that on the optimization problem (P) whose objective function satisfies additionally the quadratic functional growth (23), i.e. $f \in \mathcal{F}_{L_f,\kappa_f}$, a restarting version of algorithm (FGM) with the above choice of β_k has linear convergence without the assumption $\bar{y}^k = y^*$ for all $k \geq 0$. Restarting variants of (FGM) have been also

considered in other contexts, see e.g. [19]. By fixing a positive constant $c \in (0,\ 1)$ and then combining (62) and (23), we get:

$$f(x^k) - f^* \leq \frac{2L_f}{(k+1)^2}\|x^0 - \bar{x}^0\|^2 \leq \frac{4L_f}{\kappa_f(k+1)^2}(f(x^0) - f^*) \leq c(f(x^0) - f^*),$$

which leads to the following expression:

$$c = \frac{4L_f}{\kappa_f k^2}.$$

Then, for fixed c, the number of iterations K_c that we need to perform in order to obtain $f(x^{K_c}) - f^* \leq c(f(x^0) - f^*)$ is given by:

$$K_c = \left\lceil \sqrt{\frac{4L_f}{c\kappa_f}} \right\rceil = \left\lceil \sqrt{\frac{4}{c\mu_f}} \right\rceil.$$

Therefore, after each K_c steps of Algorithm (FGM) we restart it obtaining the following scheme:

Algorithm (R-FGM)

Given $x^{0,0} = y^{0,0} = x^0 \in X$ and restart interval K_c. For $j \geq 0$ do:

1. Run Algorithm (FGM) for K_c iterations to get $x^{K_c,j}$
2. Restart: $x^{0,j+1} = x^{K_c,j}$, $y^{0,j+1} = x^{K_c,j}$ and $\theta_1 = 1$.

Then, after p restarts of Algorithm (R-FGM) we obtain the linear convergence:

$$\begin{aligned} f(x^{0,p}) - f^* = f(x^{K_c,p-1}) - f^* &\leq \frac{2L_f\|x^{0,p-1} - \bar{x}^{0,p-1}\|^2}{(K_c+1)^2} \\ &\leq c(f(x^{0,p-1}) - f^*) \leq \cdots \leq c^p(f(x^{0,0}) - f^*) = c^p(f(x^0) - f^*). \end{aligned}$$

Thus, total number of iterations is $k = pK_c$ and denote $x^k = x^{0,p}$. Then, we have:

$$f(x^k) - f^* \leq \left(c^{\frac{1}{K_c}}\right)^k (f(x^0) - f^*).$$

We want to optimize e.g. the number of iteration K_c:

$$\min_{K_c} c^{\frac{1}{K_c}} \quad \Leftrightarrow \quad \min_{K_c} \frac{1}{K_c}\log c \quad \Leftrightarrow \quad \min_{K_c} \frac{1}{K_c}\log\frac{4}{\mu_f K_c^2},$$

which leads to

$$K_c^* = \frac{2e}{\sqrt{\mu_f}} \quad \text{and} \quad c = e^{-2}.$$

In conclusion, we get the following convergence rate for (R-FGM) method:

$$f(x^k) - f^* \leq \left(e^{-2\frac{\sqrt{\mu_f}}{2e}}\right)^k (f(x^0) - f^*) = \left(e^{-\frac{\sqrt{\mu_f}}{e}}\right)^k (f(x^0) - f^*), \quad (63)$$

and since $e^\alpha \approx 1 + \alpha$ as $\alpha \approx 0$, then for $\frac{\sqrt{\mu_f}}{e} \approx 0$ we get:

$$f(x^k) - f^* \leq \left(e^{-\frac{\sqrt{\mu_f}}{e}}\right)^k (f(x^0) - f^*) \approx \left(1 - \frac{\sqrt{\mu_f}}{e}\right)^k (f(x^0) - f^*). \quad (64)$$

Note that if the optimal value f^* is known in advance, then we just need to restart algorithm (R-FGM) at the iteration $\bar{K}_c \leq K_c^*$ when the following condition holds:

$$f(x^{\bar{K}_c, j}) - f^* \leq c(f(x^{0,j}) - f^*),$$

which can be practically verified. Using the second order growth property (23) we can also obtain easily linear convergence of the generated sequence x^k to some optimal point in X^*.

5.3 Feasible descent methods (FDM)

We now consider a more general descent version of Algorithm (GM) where the gradients are perturbed:

Algorithm (FDM)

Given $x^0 \in X$ and $\beta, L > 0$ for $k \geq 0$ do:

Compute $x^{k+1} = \left[x^k - \alpha_k \nabla f(x^k) + e^k\right]_X$
such that
$\|e^k\| \leq \beta \|x^{k+1} - x^k\|$ and $f(x^{k+1}) \leq f(x^k) - \frac{L}{2}\|x^{k+1} - x^k\|^2$,

where the stepsize α_k is chosen such that $\alpha_k \geq \bar{L}_f^{-1} > 0$ for all k. It has been showed in [9,11] that algorithm (FDM) covers important particular schemes: e.g. proximal point minimization, random/cyclic coordinate descent, extragradient descent and matrix splitting methods are all feasible descent methods. Moreover, linear convergence of algorithm (FDM) under the error bound assumption (32), i.e. $f \in \mathcal{E}_{L_f, \kappa_f}$, is proved e.g. in [9,11]. Note that if the error bound condition (32) holds for a constant κ_f, then according to Theorem 6 we also have that the quadratic functional growth condition (23) is valid with a constant $\mathcal{O}(\kappa_f^2)$. In [10,11] it has been proved that for the class of functions (43) the error bound constant κ_f depends on the Hoffman constant θ as $\kappa_f = \mathcal{O}(1/\theta^2)$, and thus the quadratic functional growth constant depends on the Hoffman constant θ as $\mathcal{O}(\kappa_f^2) = \mathcal{O}(1/\theta^4)$ (recall that the Hoffman constant θ can be very large, see the discussion from Sect. 4). Therefore, in the next theorem we prove

that the feasible descent method (FDM) converges linearly in terms of function values on optimization problems (P) whose objective functions belong to the class $\mathcal{F}_{L_f,\kappa_f}$, with better rates in terms of the Hoffman constant.

Theorem 15 *Let the optimization problem (P) have the objective function belonging to the class $\mathcal{F}_{L_f,\kappa_f}$. Then, the sequence x_k generated by the feasible descent method (FDM) on (P) converges linearly in terms of function values with the rate:*

$$f(x^k) - f^* \leq \left(\frac{1}{1 + \frac{L\kappa_f}{4(L_f+\bar{L}_f+\beta\bar{L}_f)^2}} \right)^k (f(x^0) - f^*). \tag{65}$$

Proof The optimality conditions for computing x^{k+1} are:

$$\langle x^{k+1} - x^k + \alpha_k \nabla f(x^k) - e^k, x - x^{k+1} \rangle \geq 0 \quad \forall x \in X. \tag{66}$$

Then, using convexity of f and Cauchy–Schwarz inequality, we get:

$$\begin{aligned} f(x^{k+1}) - f(\bar{x}^{k+1}) &\leq \langle \nabla f(x^{k+1}), x^{k+1} - \bar{x}^{k+1} \rangle \\ &= \langle \nabla f(x^{k+1}) - \nabla f(x^k) + \nabla f(x^k), x^{k+1} - \bar{x}^{k+1} \rangle \\ &\overset{(1)+(66)}{\leq} L_f \|x^{k+1} - x^k\| \|x^{k+1} - \bar{x}^{k+1}\| + \frac{1}{\alpha_k} \langle x^{k+1} - x^k - e^k, \bar{x}^{k+1} - x^{k+1} \rangle \\ &\leq (L_f + \bar{L}_f) \|x^{k+1} - x^k\| \|x^{k+1} - \bar{x}^{k+1}\| + \bar{L}_f \|e^k\| \|x^{k+1} - \bar{x}^{k+1}\| \\ &\leq (L_f + \bar{L}_f + \beta\bar{L}_f) \|x^{k+1} - x^k\| \|x^{k+1} - \bar{x}^{k+1}\|. \end{aligned}$$

Since $f \in \mathcal{F}_{L_f,\kappa_f}$ then it satisfies the second order growth property, i.e. $f(x^{k+1}) - f(\bar{x}^{k+1}) \geq \frac{\kappa_f}{2} \|x^{k+1} - \bar{x}^{k+1}\|^2$, and using it in the previous derivations we obtain:

$$f(x^{k+1}) - f(\bar{x}^{k+1}) \leq \frac{2(L_f + \bar{L}_f + \beta\bar{L}_f)^2}{\kappa_f} \|x^{k+1} - x^k\|^2. \tag{67}$$

Combining (67) with the descent property of the algorithm (FDM), i.e. $\|x^{k+1} - x^k\|^2 \leq \frac{2}{L} \left(f(x^k) - f(x^{k+1}) \right)$, we get:

$$f(x^{k+1}) - f(\bar{x}^{k+1}) \leq \frac{4(L_f + \bar{L}_f + \beta\bar{L}_f)^2}{L\kappa_f} \left(f(x^k) - f(x^{k+1}) \right),$$

which leads to

$$f(x^{k+1}) - f(\bar{x}^{k+1}) \leq \frac{1}{1 + \frac{L\kappa_f}{4(L_f+\bar{L}_f+\beta\bar{L}_f)^2}} \left(f(x^k) - f(\bar{x}^k) \right).$$

Using an inductive argument we get the statement of the theorem. □

Note that, once we have obtained linear convergence in terms of function values for the algorithm (FDM), we can also obtain linear convergence of the generated sequence x^k to some optimal point in X^* by using the second order growth property (23).

5.4 Discussion

From previous sections we can conclude that for some classes of problems improved linear convergence rates are obtained as compared to existing results.

For example, in the recent paper [12] the authors show that the objective function of the convex unconstrained problem $\min_{x\in\mathbb{R}^n} g(Ax)$, with g strongly convex function having Lipschitz continuous gradient, satisfies a particular version of our more general quadratic gradient growth inequality (18). However, in this paper we prove that the objective function of this particular class of optimization problems belongs to a more restricted functional class, namely $q\mathcal{S}_{L_f,\kappa_f}(X)$, i.e. it satisfies (11). Thus, for this class of problems we provide better linear rates for the gradient method as compared to [12]. More precisely, on the functional class $q\mathcal{S}_{L_f,\kappa_f}(X)$, for the gradient method we derived in Theorem 11 linear convergence rate of order $(1-\mu_f)/(1+\mu_f)$, while [12] proved linear convergence rate of order $1-\mu_f$. Similarly, in [9–11,14] the authors prove that the objective function of the convex problem $\min_{Cx\leq d} g(Ax)$, with g strongly convex function having Lipschitz continuous gradient, satisfies an error bound type condition (32) and derive convergence rates for (proximal) gradient and/or coordinate descent methods of order $1/(1+\mu_f)$, which are worse than our linear rates $(1-\mu_f)/(1+\mu_f)$ derived for the same class (see Theorem 11). Finally, in [9,11] the objective function of the constrained problem $\min_{Cx\leq d} g(Ax)+c^Tx$, with g strongly convex function having Lipschitz continuous gradient, is shown to also satisfy an error bound condition of the form (32) on any compact set and the convergence analysis of the gradient method follows usually from the one given for the feasible descent method. However, this approach provides worse convergence estimates. For example, for the choices $\alpha_k = 1/L_f$, $\beta = 0$ and $L = L_f$ we recover from the feasible descent method the gradient scheme, but the linear convergence from (54), given by $\frac{1}{1+\mu_f}$, is better than the one obtained in Theorem 15, given by $\frac{1}{1+\mu_f/16}$.

Finally,, from our best knowledge, this paper proves for the first time linear convergence of order $1-\sqrt{\mu_f}$ of the usual fast gradient method on the class of convex problems $\min_{Cx\leq d} g(Ax)$, provided that some projection property holds, while for example [12] derives a worse rate of convergence and for a restarting variant of the fast gradient method. Other papers analyzing error bound type conditions, such as [2,9–11,13–15], derive convergence rates only for feasible descent type methods and do not provide any convergence rates for the fast gradient type algorithms.

6 Applications

In this section we present several applications having the objective function in one of the structured functional classes of Sect. 4.

6.1 Solution of linear systems

It is well known that finding a solution of a symmetric linear system $Qx + q = 0$, where $Q \succeq 0$ (notation for positive semi-definite matrix), is equivalent to solving a convex quadratic program (QP):

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \left(= \frac{1}{2} x^T Q x + q^T x \right).$$

Let $Q = L_Q^T L_Q$ be the Cholesky decomposition of Q. For simplicity, let us assume that our symmetric linear system has a solution, e.g. x_s, then q is in the range of Q, i.e. $q = -Qx_s = -L_Q^T L_Q x_s$. Therefore, if we define the strongly convex function $g(z) = \frac{1}{2}\|z\|^2 - (L_Q x_s)^T z$, having $L_g = \sigma_g = 1$, then our objective function is the composition of g with the linear map $L_Q x$:

$$f(x) = \frac{1}{2}\|L_Q x\|^2 - (L_Q^T L_Q x_s)^T x = g(L_Q x).$$

Thus, our convex quadratic problem is in the form of unconstrained structured optimization problem (39) and from Sect. 4 we conclude that the objective function of this QP is in the class $q\mathcal{S}_{L_f,\kappa_f}$ with:

$$L_f = \lambda_{\max}(Q) \text{ and } \kappa_f = \sigma_{\min}^2(L_Q) = \lambda_{\min}(Q) \quad \Rightarrow \quad \mu_f = \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)} \equiv \frac{1}{\text{cond}(Q)},$$

where $\lambda_{\min}(Q)$ denotes the smallest non-zero eigenvalue of Q and $\lambda_{\max}(Q)$ is the largest eigenvalue of Q. Since we assume that our symmetric linear system has a solution, i.e. $f^* = 0$, from Theorem 14 and Remark 1 we conclude that when solving this convex QP with the algorithm (FGM) we get the convergence rate in terms of function values:

$$f(x^k) \leq \left(1 - \sqrt{\frac{1}{\text{cond}(Q)}}\right)^k \cdot 2 f(x^0)$$

or in terms of residual (gradient) or distance to the solution:

$$\begin{aligned} \|Qx^k + q\|^2 = \|\nabla f(x^k)\|^2 \leq L_f^2 \|x^k - \bar{x}^k\|^2 \leq \frac{2L_f^2}{\kappa_f} \left(f(x^k) - f^*\right) \\ \leq \left(1 - \sqrt{\frac{1}{\text{cond}(Q)}}\right)^k \cdot \lambda_{\max}(Q) \cdot \text{cond}(Q) \left(\frac{1}{2}(x^0)^T Q x^0 + q^T x^0\right). \end{aligned}$$

Therefore, the usual (FGM) algorithm without restart attains an ϵ-optimal solution in a number of iterations of order $\sqrt{\text{cond}(Q)} \log \frac{1}{\epsilon}$, i.e. the condition number $\text{cond}(Q)$ of

the matrix Q is square rooted. From our knowledge, this is one of the first results showing linear convergence depending on the square root of the condition number for the fast gradient method on solving a symmetric linear system with positive semi-definite matrix $Q \succeq 0$. Note that the linear conjugate gradient method can also attain an ϵ-optimal solution in much fewer than n steps, i.e. the same $\sqrt{\text{cond}(Q)} \log \frac{1}{\epsilon}$ iterations [20]. Usually, in the literature the condition number appears linearly in the convergence rate of first order methods for solving linear systems with positive semi-definite matrices. For example, the coordinate descent method from [1] requires $\text{cond}(Q) \log \frac{1}{\epsilon}$ iterations for obtaining an ϵ-optimal solution.

Our results can be extended for solving general linear systems $Ax + b = 0$, where $A \in \mathbb{R}^{m \times n}$. In this case we can formulate the equivalent unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^n} \|Ax + b\|^2,$$

which is a particular case of (39) and from Sect. 4 we can also conclude that the objective function of this QP is in the class $q\mathcal{S}_{L_f,\kappa_f}$ with:

$$L_f = \sigma^2_{\max}(A) \text{ and } \kappa_f = \sigma^2_{\min}(A) \quad \Rightarrow \quad \mu_f = \frac{\sigma^2_{\min}(A)}{\sigma^2_{\max}(A)},$$

where $\sigma_{\min}(A)$ denotes the smallest non-zero singular value of A and $\sigma_{\max}(A)$ is the largest singular value of A. In this case the usual (FGM) algorithm attains and ϵ-optimal solution in a number of iterations of order $\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \log \frac{1}{\epsilon}$.

6.2 Dual of linearly constrained convex problems

Let us consider the following linearly constrained convex problem:

$$\begin{aligned} &\min_u \tilde{g}(u) \\ &\text{s.t. : } c - A^T u \in \mathcal{K} = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}_+. \end{aligned}$$

Then, the dual of this optimization problem can be written in the form of structured problem (43), where g is the convex conjugate of $\tilde{g}$. From duality theory we know that g is strongly convex and with Lipschitz gradient, provided that $\tilde{g}$ is strongly convex and with Lipschitz gradient. Thus, the requirements of Theorem 10 hold for this problem when $\tilde{g}$ is smooth strongly convex function. Therefore, the convergence rates derived in this paper for the first order methods on optimization problems whose objective function belongs to the functional class $\mathcal{F}_{L_f,\kappa_f}$ are valid when solving the dual of the linearly constrained convex problem defined above.

6.3 Lasso problem

Consider the following optimization problem:

$$\min_{x: Cx \leq d} f(x) + \lambda \|x\|_1,$$

where $\lambda \geq 0$ and f has the special structure $f(x) = g(Ax) + c^T x$, where A is some data matrix, c is a vector and g is a smooth strongly convex function. It is easy to see that if we double the dimension of x using the new decision variable $\mathbf{x} = [x_+^T \; x_-^T]^T$, we can replace the 1-norm term $\lambda \|x\|_1$ with $\lambda c_1^T x_+ + \lambda c_1^T x_-$, where c_1 is the vector with all entries 1, and impose the additional polyhedral constraints $x_+, x_- \geq 0$. Then, the new reformulation of the Lasso problem, using the decision variable $\mathbf{x} = [x_+^T \; x_-^T]^T$, is a particular case of the structured optimization problem (43). Moreover, the requirements of Theorem 10 hold provided that g is a smooth strongly convex function. In conclusion, the convergence rates derived in this paper for the first order methods on optimization problems whose objective function belongs to the functional class $\mathcal{F}_{L_f,\kappa_f}$ are valid when solving the new reformulation of the Lasso problem.

6.4 Linear programming

Finding a primal-dual solution of a linear cone program can also be written in the form of a structured optimization problem (39). Indeed, let $c \in \mathbb{R}^N$, $b \in \mathbb{R}^m$ and $\mathcal{K} \subseteq \mathbb{R}^N$ be a closed convex cone, then we define the linear cone programming:

$$\min_u \langle c, u \rangle \qquad \text{s.t. } Eu = b, \quad u \in \mathcal{K}, \tag{68}$$

and its associated dual problem

$$\min_{v,s} \langle b, v \rangle \qquad \text{s.t. } E^T v + s = c, \quad s \in \mathcal{K}^*, \tag{69}$$

where $\mathcal{K}^*$ denotes the dual cone. We assume that the pair of cone programming (68)–(69) have optimal solutions and their associated duality gap is zero. Therefore, a primal-dual solution of (68)–(69) can be found by solving the following convex feasibility problem:

$$\text{find } (u, v, s) \text{ such that } \begin{cases} E^T v + s = c, \quad Eu = b, \quad \langle c, u \rangle = \langle b, v \rangle \\ u \in \mathcal{K}, \quad s \in \mathcal{K}^*, \quad v \in \mathbb{R}^m, \end{cases} \tag{70}$$

or, in a more compact formulation:

$$\text{find } x \text{ such that } \begin{cases} Ax = d \\ x \in \mathbf{K}, \end{cases}$$

where $x = \begin{bmatrix} u \\ v \\ s \end{bmatrix}$, $A = \begin{bmatrix} 0 & E^T & I_n \\ E & 0 & 0 \\ c^T & -b^T & 0 \end{bmatrix}$, $d = \begin{bmatrix} c \\ b \\ 0 \end{bmatrix}$, $\mathbf{K} = \mathcal{K} \times \mathbb{R}^m \times \mathcal{K}^*$. The authors in [21] proposed solving conic optimization problems given in the previous form using ADMM. In this paper we propose solving a Linear Program using the

first order methods presented above. A simple reformulation of this constrained linear system as an optimization problem is:

$$\min_{x \in \mathbf{K}} \|Ax - d\|^2. \tag{71}$$

Denote the dimension of the variable x as $n = 2N+m$. Let us note that the optimization problem (71) is a particular case of (39) with objective function of the form $f(x) = g(Ax)$, with $g(\cdot) = \|\cdot - d\|^2$. Moreover, the conditions of Theorem 8 hold provided that $\mathcal{K} = \mathbb{R}^N_+$. We conclude that we can always solve a linear program in linear time using the first order methods described in the present paper.

7 Conclusions

In this paper, we have derived linear convergence rates of several first order methods for solving smooth non-strongly convex constrained problems, i.e. involving an objective function satisfying some relaxed strong convexity condition. Our paper opens a large stream of new questions. For example, we have derived linear convergence rate of the fast gradient method for quasi-strongly-convex objectives, provided that the auxiliary iterates produce the same projection onto the optimal set. Although we present an example satisfying this condition, it will be interesting to investigate if alternative extrapolation schemes allow to lift this condition for constrained optimization problems. Moreover, we have proved that the class of functions satisfying the quadratic functional growth is the largest one for which gradient method is converging linearly. It will be interesting to determine also the largest class of functions for which fast gradient methods enjoy a (near-optimal) linear convergence rate. It is also worth to investigate the complexity bounds obtained when wrapping linearly-convergent algorithms under the proposed conditions in an outer-loop using a generic acceleration scheme such as Catalyst [22]. This would allow to extend to non-strongly convex settings a wide range of incremental optimization algorithms designed for large finite-sum problems arising in machine learning and statistics.

Acknowledgements The research leading to these results has received funding from the Executive Agency for Higher Education, Research and Innovation Funding (UEFISCDI), Romania: PN-III-P4-PCE-2016-0731, project ScaleFreeNet, No. 39/2017.

References

1. Leventhal, D., Lewis, A.S.: Randomized methods for linear constraints: convergence rates and conditioning. Math. Oper. Res. **35**(3), 641–654 (2010)
2. Liu, J., Wright, S., Re, C., Bittorf, V., Sridhar, S.: An asynchronous parallel stochastic coordinate descent algorithm. J. Mach. Learn. Res. **16**(1), 285–322 (2015)
3. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer, Dordrecht (2004)
4. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
5. Wright, S.: Coordinate descent algorithms. Math. Program. **151**(1), 3–34 (2015)

6. Burke, J.V., Deng, S.: Weak sharp minima revisited Part III: error bounds for differentiable convex inclusions. Math. Program. **116**(1–2), 37–56 (2009)
7. Lewis, A.S., Pang, J.S.: Error bounds for convex inequality systems. In: Chapter In: Generalized Convexity, Generalized Monotonicity–Recent Results. Springer, Berlin (1998)
8. Yangy, T., Lin, Q.: A stochastic gradient method with linear convergence rate for a class of non-smooth non-strongly convex optimization. Tech. rep. (2015). www.arxiv.org
9. Luo, Z.-Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46**(1), 157–178 (1993)
10. Necoara, I., Clipici, D.: Parallel random coordinate descent method for composite minimization: convergence analysis and error bounds. SIAM J. Optim. **26**(1), 197–226 (2016)
11. Wang, P.W., Lin, C.J.: Iteration complexity of feasible descent methods for convex optimization. J. Mach. Learn. Res. **15**(4), 1523–1548 (2014)
12. Zhang, H., Cheng, L.: Restricted strong convexity and its applications to convergence analysis of gradient type methods in convex optimization. Optim. Lett. **9**(5), 961–979 (2015)
13. Beck, A., Shtern, S.: Linearly convergent away-step conditional gradient for non-strongly convex functions. Math. Program. **164**(1–2), 1–27 (2017)
14. Drusvyatskiy, D., Lewis, A.: Error bounds, quadratic growth, and linear convergence of proximal methods. Tech. rep., (2016). (arXiv:1602.06661)
15. Zhou, Z., So, A.: A unified approach to error bounds for structured convex optimization problems. Math. Program. **165**(2), 689–728 (2017)
16. Hoffman, A.J.: On approximate solutions of systems of linear inequalities. J. Res. Natl. Bur. Stand. **49**(4), 263–265 (1952)
17. Klatte, D., Thiere, G.: Error bounds for solutions of linear equations and inequalities. Math. Methods Oper. Res. **41**(2), 191–214 (1995)
18. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
19. O'Donoghue, B., Candes, E.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**(3), 715–732 (2013)
20. Bubeck, S.: Convex optimization: algorithms and complexity. Found. Trends Mach. Learn. **8**(3–4), 231–357 (2015)
21. O'Donoghue, B., Chu, E., Parikh, N., Boyd, S.: Conic optimization via operator splitting and homogeneous self-dual embedding. J. Optim. Theory Appl. **169**(3), 1042–1068 (2016)
22. Lin, H., Mairal, J., Harchaoui, Z.: A universal Catalyst for first-order optimization. In: Advances in neural information processing systems, 3384–3392 (2015)