Worst-case convergence analysis of gradient and Newton methods through semidefinite programming performance estimation

Etienne de Klerk* François Glineur† Adrien B. Taylor†

September 15, 2017

Abstract

We provide new tools for worst-case performance analysis of the gradient (or steepest descent) method of Cauchy for smooth strongly convex functions, and Newton's method for self-concordant functions. The analysis uses semidefinite programming performance estimation, as pioneered by Drori en Teboulle [*Mathematical Programming*, 145(1-2):451–482, 2014], and extends recent performance estimation results for the method of Cauchy by the authors [*Optimization Letters*, to appear]. To illustrate the applicability of the tools, we sketch how to give a rigorous worst-case complexity analysis of a recent interior point method by Abernethy and Hazan [arXiv 1507.02528, 2015]. The algorithm of Abernethy and Hazan has sparked much recent interest, since it demonstrates the formal equivalence between an interior point method and a simulated annealing algorithm for convex optimization, but several details of the worst-case analysis are omitted in their paper.

Keywords: simulated annealing, interior point methods, performance estimation problems, semidefinite programming
AMS classification: 90C22; 90C26; 90C30

1 Introduction

We consider the worst-case convergence of the gradient and Newton methods (with or without exact linesearch, and with possibly inexact search directions) for certain smooth and strongly convex functions.

Our analysis is computer-assisted and relies on semidefinite programming performance estimation problems, as introduced by Drori and Teboulle [6]. As a result, we develop a set of tools that may be used to design or analyse a wide range of interior point algorithms. Our analysis is in fact an extension of the worst-case analysis of the gradient method in [8], combined with the fact that Newton's method may be viewed as a gradient method with respect to a suitable local (intrinsic) inner product. This is similar in spirit to recent analysis by Li et al [9] of inexact proximal Newton methods for self-concordant functions, but our approach is different.

As an illustration of the tools we develop, we show how one may give a rigorous analysis of a recent interior point method by Abernethy and Hazan [1], where the search direction is approximated through sampling. This particular method has sparked much recent interest, since it demonstrates the links between simulated annealing and interior point methods. However, a rigorous analysis of the method is not given in [1], and we supply crucial details that are missing in [1]. Polynomial-time complexity of certain simulated annealing methods for convex optimization was first shown by Kalai and Vempala [7], and the link with interior point methods casts light on their result.

2 Preliminaries

Throughout f denotes a differentiable convex function with convex domain $D_f \subset \mathbb{R}^n$. We will indicate additional assumptions on f as needed. We will mostly use the notation from the book by Renegar [13], for easy reference.

*Tilburg University and Delft University of Technology, The Netherlands, E.deKlerk@uvt.nl

†UCL / CORE and ICTEAM, Louvain-la-Neuve, Belgium, Francois.Glineur@uclouvain.be, Adrien.Taylor@uclouvain.be.
The UCL/CORE authors are supported by the Belgian Interuniversity Attraction Poles, and by the ARC grant 13/18-054 (Communauté française de Belgique).

2.1 Gradients and Hessians

In what follows we fix a reference inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^n$ with induced norm $\| \cdot \|$.

Definition 2.1 (Gradient of differentiable f)**.** *The gradient of f at $x \in D_f$ with respect to $\langle \cdot, \cdot \rangle$ is the unique vector $g(x)$ such that*

$$\lim_{\|\Delta x\| \to 0} \frac{f(x + \Delta x) - f(x) - \langle g(x), \Delta x \rangle}{\|\Delta x\|} = 0.$$

Note that $g(x)$ depends on the reference inner product. If $\langle \cdot, \cdot \rangle$ is the Euclidean dot product then $g(x) = \nabla f(x) = \left[\frac{\partial f(x)}{\partial x_i} \right]_{i=1,\ldots,n}$.

If $B : \mathbb{R}^n \to \mathbb{R}^n$ is a self-adjoint positive definite linear operator, we may define a new inner product in terms of the reference inner product as follows: $\langle \cdot, \cdot \rangle_B$ via $\langle x, y \rangle_B = \langle x, By \rangle \ \forall x, y \in \mathbb{R}^n$. (Recall that all inner products in $\mathbb{R}^n$ arise in this way.)

If we change the inner product in this way, then the gradient changes as follows:

$$\langle \cdot, \cdot \rangle \to \langle \cdot, \cdot \rangle_B \Rightarrow g(x) \to B^{-1} g(x).$$

Thus one always has $g(x) = B^{-1} \nabla f(x)$ for some positive definite B, that depends on the choice of reference inner product.

If f is twice differentiable, we define its Hessian as follows.

Definition 2.2 (Second derivative of twice differentiable f)**.** *The second derivative (or Hessian) of f at x is defined as the (unique) linear operator $H(x)$ that satisfies*

$$\lim_{\|\Delta x\| \to 0} \frac{\|g(x + \Delta x) - g(x) - H(x)\Delta x\|}{\|\Delta x\|} = 0.$$

The second derivative also depends on the reference inner product, since $g(x)$ does. Recall that $H(x)$ is self-adjoint with respect to the reference inner product if f is twice continuously differentiable.

Assuming that $H(x)$ is positive definite and self-adjoint at a given x, define the intrinsic (w.r.t. f at x) inner product

$$\langle u, v \rangle_x := \langle u, v \rangle_{H(x)} \equiv \langle u, H(x)v \rangle.$$

The definition is *independent of the reference inner product* $\langle \cdot, \cdot \rangle$, e.g. without loss of generality we may set the reference inner product to be the Euclidean one, to obtain:

$$\langle u, v \rangle_x = u^{\mathsf{T}} \nabla^2 f(x) v,$$

where $\nabla^2 f(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right)$ is the usual $n \times n$ Hessian matrix of second partial derivatives at x. (The Hessian matrix is obtained by writing $H(x)$ as a matrix in the standard basis, when the reference inner product is the Euclidean inner product.)

The induced norm for the intrinsic inner product is denoted by: $\|u\|_x = \sqrt{\langle u, u \rangle_x}$. For the intrinsic inner product

$$\langle u, v \rangle_x := \langle u, v \rangle_{H(x)},$$

the gradient at y is denoted by $g_x(y) := H(x)^{-1} g(y)$, and the Hessian at y by $H_x(y) := H(x)^{-1} H(y)$.

2.2 Fundamental theorem of calculus

In what follows, we will recall coordinate-free versions of the fundamental theorem of calculus. Our review follows Renegar [13], and all proofs may be found there.

Theorem 2.3 (Theorem 1.5.1 in [13])**.** *If $x, y \in D_f$, then*

$$f(y) - f(x) = \int_0^1 \langle g(x + t(y - x)), y - x \rangle dt.$$

Next, we recall the definition of a vector-valued integral.

Definition 2.4. *Let $t \mapsto v(t) \in \mathbb{R}^n$ where $t \in [a, b]$. Then u is the integral of v if*

$$\langle u, w \rangle = \int_a^b \langle v(t), w \rangle \, dt \text{ for all } w \in \mathbb{R}^n.$$

Note that this definition is in fact independent of the reference inner product.

We will use the following bound on norms of vector-values integrals.

Theorem 2.5 (Proposition 1.5.4 in [13])**.** *Let $t \mapsto v(t) \in \mathbb{R}^n$ where $t \in [a, b]$. If v is integrable, then*

$$\left\| \int_a^b v(t) dt \right\| \leq \int_a^b \|v(t)\| dt.$$

Finally, we will will require the following version of the fundamental theorem for gradients.

Theorem 2.6 (Theorem 1.5.6 in [13])**.** *If $x, y \in D_f$, then*

$$g(y) - g(x) = \int_0^1 H(x + t(y - x))(y - x) dt.$$

2.3 Inexact gradient and Newton methods

The minimizer of f is denoted by x_*.

We consider approximate gradients $d_i \approx g(x_i)$ at a given iterate x_i $(i = 0, 1, \ldots)$; to be precise we assume the following for a given $\varepsilon \geq 0$:

$$\|d_i - g(x_i)\| \leq \varepsilon \|g(x_i)\| \quad i = 0, 1, \ldots \tag{1}$$

Note that $\varepsilon = 0$ yields the gradient, i.e. $d_i = g(x_i)$.

Inexact gradient descent method

Input: $f : \mathbb{R}^n \to \mathbb{R}$, $x_0 \in \mathbb{R}^n$, $0 \leq \varepsilon < 1$.

for $i = 0, 1, \ldots$

Select any d_i that satisfies (1);

Choose a step length $\gamma > 0$

$x_{i+1} = x_i - \gamma d_i$

We will consider two ways of choosing the step length γ:

1. Exact line search: $\gamma = \text{argmin}_{\gamma \in \mathbb{R}} f(x_i - \gamma d_i)$;
2. Fixed step length: γ takes the same value at each iteration, and this value is known beforehand.

We note once more that, at iteration i, we obtain the Newton direction by using the $\langle \cdot, \cdot \rangle_{x_i}$ inner product. Similarly, by using an inner product $\langle \cdot, \cdot \rangle_B$ for some positive definite, self-adjoint operator B, we obtain the direction $-B^{-1} g(x_i)$, which is of the type used in quasi-Newton methods. Finally, the Euclidean dot product yields the familiar steepest descent direction $-\nabla f(x_i)$.

3 Classes of convex functions

In this section we review two classes of convex functions, namely smooth, strongly convex functions and self-concordant functions. We also show that, in a certain sense, the latter class may be seen as a special case of the former.

3.1 Convex functions

Recall that a differentiable function f is convex on on open convex set $D \subset \mathbb{R}^n$ if and only if

$$f(y) \geq f(x) + \langle g(x), y - x \rangle \quad \forall x, y \in D. \tag{2}$$

Also recall that a twice continuously differentiable function is convex on D if and only if $H(x) \succeq 0$ for all $x \in D$.

3.2 Smooth strongly convex functions

A differentiable function f with $D_f = \mathbb{R}^n$ is called L-smooth and μ-strongly convex if it satisfies the following two properties:

(a) L**-smoothness**: there exists some $L > 0$ such that $\frac{1}{L}\|g(u) - g(v)\| \leq \|u - v\|$ holds for all pairs u, v and corresponding gradients $g(u), g(v)$.

(b) μ**-strong convexity**: there exists some $\mu > 0$ such that the function $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex.

The class of such functions is denoted by $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$. Note that this function class is defined in terms of the reference inner product and its induced norm. In particular, it is not invariant under a change of inner product: under a change of inner product a smooth, strongly convex function remains smooth, strongly convex, but the parameters μ and L depend on the inner product.

If $D \subseteq \mathbb{R}^n$ is an open, convex set, then we denote the set of functions that satisfy properties (a) and (b) on D by $\mathcal{F}_{\mu,L}(D)$.

Lemma 3.1. *Let D be an open convex set and $f : D \to \mathbb{R}^d$ be twice continuously differentiable. The following statements are equivalent:*

(a) f is convex and L-smooth on D,

(b) $0 \preceq H(x) \preceq LI \ \ \forall x \in D$,

(c) $\langle g(y) - g(x), y - x \rangle \geq \frac{1}{L}\|g(y) - g(x)\|^2 \ \ \forall x, y \in D$.

Proof. $(a) \Rightarrow (b)$: First of all, convexity of f is equivalent to $H(x) \succeq 0$ for all $x \in D$. If f is also L-smooth of D, then $\|g(x) - g(y)\| \leq L\|x - y\|$ for all $x, y \in D$. By the definition of the Hessian one has that, for any $\epsilon > 0$, there is a $\delta > 0$ such that

$$\|g(y) - g(x) - H(x)(y - x)\| \leq \epsilon\|y - x\|$$

if $\|y - x\| \leq \delta$. This implies

$$\begin{aligned} \|H(x)(y - x)\| &\leq \epsilon\|y - x\| + \|g(y) - g(x)\| \quad \text{(triangle inequality)} \\ &\leq \epsilon\|y - x\| + L\|y - x\| \quad (L\text{-smoothness}). \end{aligned}$$

Thus we have

$$\frac{\|H(x)(y - x)\|}{\|y - x\|} \leq \epsilon + L,$$

which in turn implies

$$\|H(x)\| = \max_{\|v\|=1} \|H(x)v\| \leq L,$$

since $\epsilon \geq 0$ was arbitrary. Finally we use the fact that, since $H(x)$ is self-adjoint and positive semidefinite, $\|H(x)\|$ equals the largest eigenvalue of $H(x)$, so $\|H(x)\| \leq L$ is the same as $H(x) \preceq LI$.

$(b) \Rightarrow (c)$: If $\|H(x)\| \leq L$, then $H(x) - \frac{1}{L}H^2(x) \succeq 0$ for all $x \in D$, and Theorem 2.6 implies

$$\begin{aligned}
\langle g(y) - g(x), y - x\rangle &= \left\langle y - x, \int_0^1 H(x + t(y-x))(y-x)dt \right\rangle \\
&= \int_0^1 \langle y - x, H(x + t(y-x))(y-x)\rangle dt \\
&\geq \int_0^1 \langle y - x, \frac{1}{L}H^2(x + t(y-x))(y-x)\rangle dt \\
&= \frac{1}{L}\int_0^1 \|H(x + t(y-x))(y-x)\|^2 dt \\
&\geq \frac{1}{L}\left(\int_0^1 \|H(x + t(y-x))(y-x)\| dt\right)^2 \quad \text{(Jensen inequality)} \\
&\geq \frac{1}{L}\left\|\int_0^1 H(x + t(y-x))(y-x)dt\right\|^2 \quad \text{(Theorem 2.5)} \\
&= \frac{1}{L}\|g(y) - g(x)\|^2 \quad \text{(Theorem 2.6).}
\end{aligned}$$

$(c) \Rightarrow (a)$: Condition (c), together with the Cauchy-Schwartz inequality, immediately imply L-smoothness. To show convexity, note that, by Theorem 2.3,

$$\begin{aligned}
f(y) - f(x) - \langle g(x), y - x\rangle &= \int_0^1 \frac{1}{t}\langle g(x + t(y-x)) - g(x), t(y-x)\rangle dt \\
&\geq \int_0^1 \frac{1}{tL}\|g(x + t(y-x)) - g(x)\|^2 dt \geq 0,
\end{aligned}$$

where the first inequality is from condition (c). Thus we obtain the convexity inequality (2). □

The last Lemma allows us to derive the following necessary and sufficient conditions for membership of $\mathcal{F}_{\mu,L}(D)$. The condition (d) below will be used extensively in the proofs that follow.

Theorem 3.2. *Let D be an open convex set and $f : D \to \mathbb{R}^d$ be twice continuously differentiable. The following statements are equivalent:*

(a) f is μ-strongly convex and L-smooth on D, i.e. $f \in \mathcal{F}_{\mu,L}(D)$,

(b) $\mu I \preceq H(x) \preceq LI \; \forall x \in D$,

(c) $f(x) - \frac{\mu}{2}\|x\|^2$ is convex and $(L-\mu)$-smooth on D,

(d) for all $x, y \in D$ we have

$$\langle g(x) - g(y), x - y\rangle \geq \frac{1}{1 - \frac{\mu}{L}}\left(\frac{1}{L}\|g(y) - g(x)\|^2 + \mu\|x - y\|^2 - 2\frac{\mu}{L}\langle g(x) - g(y), x - y\rangle\right). \tag{3}$$

Proof. The equivalences $(a) \Leftrightarrow (b) \Leftrightarrow (c)$ follow directly from Lemma 3.1 and the relevant definitions.

$(c) \Leftrightarrow (d)$: Requiring $h(x) = f(x) - \frac{\mu}{2}\|x\|^2$ to be convex and $(L-\mu)$-smooth on D can equivalently be formulated as requiring h to satisfy

$$\langle g_h(x) - g_h(y), x - y\rangle \geq \frac{1}{L - \mu}\|g_h(y) - g_h(x)\|^2$$

for all $x, y \in D$, where g_h is the gradient of h. Equivalently:

$$(L-\mu)\left[\langle g_f(x) - g_f(y), x - y\rangle - \mu\|x - y\|^2\right] \geq \|g_f(y) - g_f(x)\|^2 + \mu^2\|x - y\|^2 - 2\mu\langle g_f(y) - g_f(x), y - x\rangle,$$

which is exactly condition (d) in the statement of the theorem. □

We note once more that the spectrum of the Hessian is not invariant under change in inner product, thus the values μ and L are intrinsically linked to the reference inner product. If $D = \mathbb{R}^n$, a stronger condition than condition (d) in the last theorem holds, namely

$$f(y) - f(x) - \langle g(x), y - x\rangle \geq \frac{1}{2(1-\frac{\mu}{L})}\left(\frac{1}{L}\|g(y) - g(x)\|^2 + \mu\|x - y\|^2 - 2\frac{\mu}{L}\langle g(x) - g(y), x - y\rangle\right). \quad (4)$$

3.3 Self-concordance

Self-concordant functions are special convex functions introduced by Nesterov and Nemirovski [12], that play a certral role in the analysis of interior point algorithms. We will use the (slightly) more general definition by Renegar [13].

Definition 3.3 (Self-concordant functional)**.** *Let $B_x(x, 1)$ be the open unit ball centered at x for the $\|\cdot\|_x$ norm. Then f is called self-concordant if:*

1. *For all $x \in D_f$ one has $B_x(x, 1) \subseteq D_f$;*
2. *For all $y \in B_x(x, 1)$ one has*

$$1 - \|y - x\|_x \leq \frac{\|v\|_y}{\|v\|_x} \leq \frac{1}{1 - \|y - x\|_x} \text{ for all } v \neq 0.$$

A subclass of self-concordant functions, that play a key role in interior point analysis, are the so-called self-concordant barriers.

Definition 3.4 (Self-concordant barrier)**.** *A self-concordant function f is called a ϑ-self-concordant barrier if there is a finite value $\vartheta \geq 1$ given by*

$$\vartheta := \sup_{x \in D_f} \|g_x(x)\|_x^2.$$

A self-concordant barrier tends to ∞ as x approaches the boundary of D_f, hence the name.

An equivalent characterization of self-concordance is as follows.

Theorem 3.5 (Theorem 2.2.1 in [13])**.** *Assume f such that for all $x \in D_f$ one has $B_x(x, 1) \subseteq D_f$.*
Then f is self-concordant if, and only if, for all $x \in D_f$ and $y \in B_x(x, 1)$:

$$\|H_x(y)\|_x, \|H_x(y)^{-1}\|_x \leq \frac{1}{(1 - \|y - x\|_x)^2}$$

or, equivalently

$$\|I - H_x(y)\|_x, \|I - H_x(y)^{-1}\|_x \leq \frac{1}{(1 - \|y - x\|_x)^2} - 1.$$

This alternative characterization allows us to establish the following link between self-concordant and L-smooth, μ-strictly convex functions.

Theorem 3.6. *Assume $f : D_f \to \mathbb{R}$ is self-concordant, $x \in D_f$, and $\delta < 1$. If*

$$D = \{y \mid \|x - y\|_x < \delta\} = B_x(x, \delta),$$

then $f \in F_{\mu,L}(D)$ with

$$\mu = (1 - \delta)^2,\ L = \frac{1}{(1 - \delta)^2}.$$

Proof. By Theorem 3.5, the spectrum of $H_x(y)$ is contained in the interval $\left[(1 - \|y - x\|_x)^2, \frac{1}{(1-\|y-x\|_x)^2}\right]$, which in turn is contained in the interval $\left[(1 - \delta)^2, \frac{1}{(1-\delta)^2}\right]$ for all $y \in B_x(x, \delta)$.

Theorem 3.2 now yields the required result. □

4 Performance estimation problems

Performance estimation problems, as introduced by Drori and Teboulle [6], are semidefinite programming (SDP) problems that bound the worst-case performance of certain iterative optimization algorithms. Essentially, the goal is to find the objective function from a given function class, that exhibits the worst-case behavior for a given iterative algorithm.

In what follows we list the SDP performance estimation problems that we will use.

The performance estimation problems have variables that correspond to (unknown) iterates x_0 and x_1, the minimizer x_*, as well as the gradients and function values at these points, namely g_i ($i \in \{*, 0, 1\}$)) correspond to $g(x_i)$ ($i \in \{*, 0, 1\}$)), and f_i ($i \in \{*, 0, 1\}$)) correspond to $f(x_i)$ ($i \in \{*, 0, 1\}$)). We may assume $x_* = g_* = 0$ and $f_* = 0$ without loss of generality.

The objective is to find the maximum (worst-case) value of either $f_1 - f_*$, $\|g_1\|$, or $\|x_1 - x_*\|$. Note again that the norm may be any induced norm on $\mathbb{R}^n$.

4.1 Performance estimation with exact line search

- Parameters: $L \geq \mu > 0$, $R > 0$;
- Variables: $\{(x_i, g_i, f_i)\}_{i \in S}$ $(S = \{*, 0, 1\})$.

Worst-case function value

$$\left.\begin{array}{ll} \max & f_1 - f_* \\ \text{s.t.} & f_i - f_j - \langle g_j, x_i - x_j \rangle \geq \frac{1}{2(1-\mu/L)} \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - 2\frac{\mu}{L} \langle g_j - g_i, x_j - x_i \rangle \right) \quad \forall i, j \in S \\ & g_* = 0 \\ & \langle x_1 - x_0, g_1 \rangle = 0 \\ & \langle g_0, g_1 \rangle \leq \varepsilon \|g_0\| \|g_1\| \\ & f_0 - f_* \leq R \end{array}\right\} \tag{5}$$

The first constraint corresponds to (4), and models the necessary condition for $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$. The second constraint corresponds to the fact that the gradient is zero at a minimizer. The third constraint is the well-known property of exact line search, while the fourth constraint is satisfied if the approximate gradient condition (1) holds. Finally, the fifth constraint ensures that the problem is bounded.

Note that the resulting problem may be written as an SDP problem, with 4×4 matrix variable given by the Gram matrix of the vectors x_0, x_1, g_0, g_1 with respect to the reference inner product. In particular the fourth contraint may be written as the linear matrix inequality:

$$\begin{pmatrix} \varepsilon \|g_0\|^2 & \langle g_0, g_1 \rangle \\ \langle g_0, g_1 \rangle & \varepsilon \|g_1\|^2 \end{pmatrix} \succeq 0. \tag{6}$$

Also note that the optimal value of the resulting SDP problem is independent of the inner product. The SDP problem (5) was first studied in [8].

Worst-case gradient norm

The second variant of performance estimation is to find the worst case convergence of the gradient norm.

$$\left.\begin{array}{ll} \max & \|g_1\|^2 \\ \text{s.t.} & \langle g_i - g_j, x_i - x_j \rangle \geq \frac{1}{1-\frac{\mu}{L}} \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - 2\frac{\mu}{L} \langle g_i - g_j, x_i - x_j \rangle \right) \quad \forall i, j \in S \\ & g_* = 0 \\ & \langle x_1 - x_0, g_1 \rangle = 0 \\ & \langle g_0, g_1 \rangle \leq \varepsilon \|g_0\| \|g_1\| \\ & \|g_0\|^2 \leq R \end{array}\right\} \tag{7}$$

Worst-case distance to optimality

The third variant of performance estimation is to find the worst case convergence of the distance to optimality.

$$\left.\begin{array}{ll} \max & \|x_1 - x_*\|^2 \\ \text{s.t.} & \langle g_i - g_j, x_i - x_j\rangle \geq \frac{1}{1-\frac{\mu}{L}}\left(\frac{1}{L}\|g_i - g_j\|^2 + \mu\|x_i - x_j\|^2 - 2\frac{\mu}{L}\langle g_i - g_j, x_i - x_j\rangle\right) \quad \forall i,j \in S \\ & g_* = 0 \\ & \langle x_1 - x_0, g_1\rangle = 0 \\ & \langle g_0, g_1\rangle \leq \varepsilon\|g_0\|\|g_1\| \\ & \|x_0\|^2 \leq R \end{array}\right\} \tag{8}$$

In what follows we will give upper bounds on the optimal values of these performance estimation SDP problems.

4.2 PEP with fixed step sizes

For fixed step sizes, the performance estimation problems (5), (7), and (8) change as follows:

- for given step size $\gamma > 0$, the condition $x_1 = x_0 - \gamma d$ is used to eliminate x_1, where d is the approximate gradient at x_0.
- The condition $\|d - g_0\|^2 \leq \varepsilon^2\|g_0\|^2$ is added, that corresponds to (1), and $\langle g_0, g_1\rangle \leq \varepsilon\|g_0\|\|g_1\|$ is omitted.

5 Error bounds from performance estimation

The optimal values of the performance estimation problems in the last section give bounds on the worst-case convergence rate of the gradient method.

5.1 PEP with exact line search

Theorem 5.1. *Consider the inexact gradient method with exact line search applied to some $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$. If $\kappa := \frac{\mu}{L}$ and $\varepsilon \in \left[0, \frac{2\sqrt{\kappa}}{1+\kappa}\right]$, one has*

$$\begin{aligned} f(x_1) - f(x_*) &\leq \left(\frac{1-\kappa+\varepsilon(1-\kappa)}{1+\kappa+\varepsilon(1-\kappa)}\right)^2 (f(x_0) - f(x_*)), \\ \|g(x_1)\| &\leq \left(\varepsilon + \sqrt{1-\varepsilon^2}\frac{1-\kappa}{2\sqrt{\kappa}}\right)\|g(x_0)\|, \\ \|x_1 - x_*\| &\leq \left(\varepsilon + \sqrt{1-\varepsilon^2}\frac{1-\kappa}{2\sqrt{\kappa}}\right)\|x_0 - x_*\|. \end{aligned}$$

Proof. The three inequalities in the statement of the theorem follow from corresponding upper bounds on the optimal values of the SDP performance estimation problems (5), (7), and (8), respectively.

The first of the SDP performance estimation problems, namely (5), is exactly the same as the one in [8] (see (8) there). The first inequality therefore follows from Theorem 1.2 in [8].

It remains to demonstrate suitable upper bounds on the SDP performance estimation problems (7), and (8). This is done by aggregating the constraints of these respective SDP problems by using suitable (Lagrange) multipliers.[1]

[1]The multipliers used in the proof were obtained by solving the SDP problems (7) and (8) numerically for different values of μ, L and R, and subsequently guessing the correct analytical expressions of the multipliers by looking at the optimal solution of the dual SDP problem. The correctness of these expressions for the multipliers is verified in the proof. Thus the performance estimation problems were only used to find the proof of Theorem 5.1, and play no role in the proof itself.

To this end, consider the following constraints from (7) and their associated multipliers:

$$\begin{aligned}
&\langle g_0 - g_1, x_0 - x_1\rangle \geq \frac{1}{1-\frac{\mu}{L}}\left(\frac{1}{L}\|g_0-g_1\|^2 + \mu\|x_1-x_0\|^2 - 2\frac{\mu}{L}\langle g_0-g_1, x_0-x_1\rangle\right) && : (L-\mu)\lambda,\\
&\langle g_1, x_1 - x_0\rangle \leq 0 && : L+\mu,\\
&\begin{pmatrix}\epsilon||g_0||^2 & \langle g_0, g_1\rangle\\ \langle g_0, g_1\rangle & \epsilon||g_1||^2\end{pmatrix} \succeq 0 && : S,
\end{aligned}$$

with $S = \begin{pmatrix} s_{11} & s_{12}\\ s_{12} & s_{22}\end{pmatrix}$, and:

$$\begin{aligned}
\lambda &= \frac{2\varepsilon\sqrt{\kappa}}{\sqrt{1-\varepsilon^2}\,(1-\kappa)} + 1,\\
s_{11} &= \frac{3\varepsilon}{2} - \frac{\varepsilon\left(\kappa+\frac{1}{\kappa}\right)}{4} + \frac{1-\kappa}{2\sqrt{\kappa(1-\varepsilon^2)}} - \frac{\varepsilon^2(1-\kappa)}{\sqrt{\kappa(1-\varepsilon^2)}},\\
s_{22} &= \frac{2\sqrt{\kappa(1-\varepsilon^2)} - \varepsilon(1-\kappa)}{(1-\varepsilon^2)(1-\kappa) + 2\varepsilon\sqrt{\kappa(1-\varepsilon^2)}},\\
s_{21} &= \frac{\varepsilon\,(1-\kappa)}{2\sqrt{\kappa(1-\varepsilon^2)}} - 1.
\end{aligned}$$

Assuming the corresponding multipliers are of appropriate signs (see discussion below), the proof consists in reformulating the following weighted sum of the previous inequalities (the validity of this inequality follows from the signs of the multipliers):

$$\begin{aligned}
0 \geq &(L-\mu)\lambda\left[\frac{1}{1-\frac{\mu}{L}}\left(\frac{1}{L}\|g_0-g_1\|^2 + \mu\|x_1-x_0\|^2 - 2\frac{\mu}{L}\langle g_0-g_1, x_0-x_1\rangle\right) - \langle g_0-g_1, x_0-x_1\rangle\right]\\
&+ (L+\mu)\left[\langle g_1, x_1-x_0\rangle\right] - \text{Trace}\left(\begin{pmatrix} s_{11} & s_{12}\\ s_{12} & s_{22}\end{pmatrix}\begin{pmatrix}\epsilon||g_0||^2 & \langle g_0, g_1\rangle\\ \langle g_0, g_1\rangle & \epsilon||g_1||^2\end{pmatrix}\right).
\end{aligned} \tag{9}$$

To this end, we first show that all multipliers are nonnegative; that is, $(L-\mu)\lambda \geq 0$, $L+\mu \geq 0$ and $S \succeq 0$. The nonnegativity of the first two is clear from their expressions (sum of nonnegative terms). Concerning S, let us note that $s_{22} \geq 0 \Leftrightarrow \varepsilon \in \left[\frac{-2\sqrt{\kappa}}{1+\kappa}, \frac{2\sqrt{\kappa}}{1+\kappa}\right]$. When $\varepsilon < \frac{2\sqrt{\kappa}}{1+\kappa}$, s_{22} ensures that there exists a positive eigenvalue for S, since $s_{22} > 0$. In order to prove that both eigenvalues of S are nonnegative, one may verify:

$$\det S = s_{11}s_{22} - s_{21}^2 = 0.$$

Therefore, one eigenvalue of S is positive and the other one is zero when $\varepsilon < \frac{2\sqrt{\kappa}}{1+\kappa}$, and in the simpler case $\varepsilon = \frac{2\sqrt{\kappa}}{1+\kappa}$, we have $S = 0$, and hence the inequality (9) is valid.

Reformulating the valid inequality (9) yields:

$$\begin{aligned}
\|g_1\|^2 \leq &\left(\varepsilon + \sqrt{1-\varepsilon^2}\frac{1-\kappa}{2\sqrt{\kappa}}\right)^2\|g_0\|^2\\
&- \kappa\frac{2\varepsilon\sqrt{\kappa} + (1-\kappa)\sqrt{1-\varepsilon^2}}{(1-\kappa)\sqrt{1-\varepsilon^2}}\left\|\frac{\varepsilon(1+\kappa)}{\sqrt{\kappa}\left(\sqrt{1-\varepsilon^2}(1-\kappa) + 2\varepsilon\sqrt{\kappa}\right)}g_1 - \frac{1+\kappa}{2\kappa}g_0 + L(x_0-x_1)\right\|^2,\\
\leq &\left(\varepsilon + \sqrt{1-\varepsilon^2}\frac{1-\kappa}{2\sqrt{\kappa}}\right)^2\|g_0\|^2,
\end{aligned}$$

where the last inequality follows from the sign of the coefficient:

$$\kappa\frac{2\varepsilon\sqrt{\kappa} + (1-\kappa)\sqrt{1-\varepsilon^2}}{(1-\kappa)\sqrt{1-\varepsilon^2}} \geq 0.$$

Next, we prove the exact same guarantee as for the gradient norm, but in the case of distance to optimality $\|x_1 - x_*\|^2$.

Let us consider the following constraints from (8) with associated multipliers:

$$\begin{aligned}
&\frac{1}{1-\frac{\mu}{L}}\left(\frac{1}{L}\|g_0\|^2 + \mu\|x_0 - x_*\|^2 - 2\frac{\mu}{L}\langle g_0, x_0 - x_*\rangle\right) + \langle g_0, x_* - x_0\rangle \leq 0 && : \lambda_0,\\
&\frac{1}{1-\frac{\mu}{L}}\left(\frac{1}{L}\|g_1\|^2 + \mu\|x_1 - x_*\|^2 - 2\frac{\mu}{L}\langle g_1, x_1 - x_*\rangle\right) + \langle g_1, x_* - x_1\rangle \leq 0 && : \lambda_1,\\
&\langle g_1, x_1 - x_0\rangle \leq 0 && : \lambda_2,\\
&\begin{pmatrix} \epsilon||g_0||^2 & \langle g_0, g_1\rangle \\ g_0^\mathsf{T} g_1 & \epsilon||g_1||^2 \end{pmatrix} \succeq 0 && : S,
\end{aligned}$$

with $S = \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix}$, and:

$$\begin{aligned}
\lambda_0 &= \frac{1-\kappa}{\mu}\left[1 - 2\varepsilon^2 + \frac{\varepsilon\sqrt{1-\varepsilon^2}}{2\sqrt{\kappa}(1-\kappa)}(-1-\kappa^2+6\kappa)\right],\\
\lambda_1 &= \frac{1}{\mu} - \frac{1}{L},\\
\lambda_2 &= \frac{1}{\mu} + \frac{1}{L},\\
L\mu s_{11} &= \frac{3\varepsilon}{2} - \frac{\varepsilon\left(\kappa + \frac{1}{\kappa}\right)}{4} + \frac{1-\kappa}{2\sqrt{\kappa(1-\varepsilon^2)}} - \frac{\varepsilon^2(1-\kappa)}{\sqrt{\kappa(1-\varepsilon^2)}},\\
L\mu s_{22} &= \frac{2\sqrt{\kappa(1-\varepsilon^2)} - \varepsilon(1-\kappa)}{(1-\varepsilon^2)(1-\kappa) + 2\varepsilon\sqrt{\kappa(1-\varepsilon^2)}},\\
L\mu s_{21} &= \frac{\varepsilon(1-\kappa)}{2\sqrt{\kappa(1-\varepsilon^2)}} - 1.
\end{aligned}$$

As in the case of the gradient norm, we proceed by reformulating the weighted sum of the constraints. For doing that, we first check nonnegativity of the weights $\lambda_0, \lambda_1, \lambda_2 \geq 0$ and $S \succeq 0$.

As in the previous case, $s_{22} \geq 0 \Leftrightarrow \varepsilon \in \left[\frac{-2\sqrt{\kappa}}{1+\kappa}, \frac{2\sqrt{\kappa}}{1+\kappa}\right]$. We therefore only need to check the sign of λ_0 in order to have the desired results (the $S \succeq 0$ requirement is the same as for the convergence in gradient norm, and the others are easily verified). Concerning λ_0, we have

$$\lambda_0 \geq 0 \Leftrightarrow \frac{\kappa-1}{\kappa+1} \leq \varepsilon \leq \frac{2\sqrt{\kappa}}{\kappa+1},$$

with $\frac{\kappa-1}{\kappa+1} \leq 0$, and hence $\lambda_0 \geq 0$ in the region of interest.

Aggregating the constraints with the corresponding multipliers yields:

$$\begin{aligned}
\|x_1 - x_*\|^2 \leq{}& \left(\varepsilon + \sqrt{1-\varepsilon^2}\frac{1-\kappa}{2\sqrt{\kappa}}\right)^2 \|x_0 - x_*\|^2\\
&- \frac{2\varepsilon\sqrt{(1-\varepsilon^2)\kappa + (1-\varepsilon^2)(1-\kappa)}}{\kappa(1-\kappa)} \times\\
&\left\|\left(1 - \frac{\varepsilon(1-\kappa)}{2\sqrt{(1-\varepsilon^2)\kappa}}\right)\frac{g_0}{L} - \frac{1+\kappa}{2}(x_0 - x_*) + \frac{1-\kappa}{2\varepsilon\sqrt{(1-\varepsilon^2)\kappa} + (1-\varepsilon^2)(1-\kappa)}\frac{g_1}{L}\right\|^2,\\
\leq{}& \left(\varepsilon + \sqrt{1-\varepsilon^2}\frac{1-\kappa}{2\sqrt{\kappa}}\right)^2 \|x_0 - x_*\|^2,
\end{aligned}$$

where the last inequality follows from the sign of the coefficient

$$\frac{2\varepsilon\sqrt{(1-\varepsilon^2)\,\kappa} + \left(1-\varepsilon^2\right)(1-\kappa)}{\kappa(1-\kappa)} \geq 0.$$

This completes the proof. □

One has the following variation on Theorem 5.1 that deals with the case where $f \in \mathcal{F}_{\mu,L}(D)$ for some open convex set $D \subset \mathbb{R}^n$.

Theorem 5.2. *Consider the inexact gradient method with exact line search applied to some $f \in \mathcal{F}_{\mu,L}(D)$ where $D \subset \mathbb{R}^n$ is open and convex, from a starting point $x_0 \in D$. Assume that $\{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\} \subset D$. If $\kappa := \frac{\mu}{L}$ and $\varepsilon \in \left[0, \frac{2\sqrt{\kappa}}{1+\kappa}\right]$, one has*

$$\begin{aligned} \|g(x_1)\| &\leq \left(\varepsilon + \sqrt{1-\varepsilon^2}\frac{1-\kappa}{2\sqrt{\kappa}}\right)\|g(x_0)\|, \\ \|x_1 - x_*\| &\leq \left(\varepsilon + \sqrt{1-\varepsilon^2}\frac{1-\kappa}{2\sqrt{\kappa}}\right)\|x_0 - x_*\|. \end{aligned}$$

Proof. The proof follows from the proof of Theorem 5.1 after the following observations:

1. The proof of the last two inequalities in Theorem 5.1 only relies on the inequality (3), which holds for any open, convex $D \subset \mathbb{R}^n$, i.e. not only for $D = \mathbb{R}^n$, by Theorem 3.2.
2. By the assumption on the level set of f, exact line search yields a point $x_1 \in D$, as required.

□

Note that the first inequality in Theorem 5.1 (convergence in function value) does not extend readily to all convex D, since its proof requires the inequality (4).

5.2 PEP with fixed step sizes

We now state a result that is similar to Theorem 5.1, but deals with fixed step lengths in stead of exact line search.

Theorem 5.3. *Consider the inexact gradient method with fixed step length γ applied to some $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$. If $\varepsilon \leq \frac{2\mu}{L+\mu}$, and $\gamma \in \left[0, \frac{2\mu-\varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)}\right]$, one has*

$$\begin{aligned} f(x_1) - f(x_*) &\leq (1-(1-\varepsilon)\mu\gamma)^2(f(x_0) - f(x_*)), \\ \|g(x_1)\| &\leq (1-(1-\varepsilon)\mu\gamma)\|g_0\|, \\ \|x_1 - x_*\| &\leq (1-(1-\varepsilon)\mu\gamma)\|x_0 - x_*\|. \end{aligned}$$

Proof. The proof is similar to that of Theorem 5.1, and is sketched in the appendix. □

Note that, if $\gamma = \frac{2\mu-\varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)}$, the factor $(1-(1-\varepsilon)\mu\gamma)$ that appears in the inequalities in Theorem 5.3 reduces to

$$1-(1-\varepsilon)\mu\gamma = \frac{1-\kappa}{1+\kappa} + \varepsilon,$$

where $\kappa = \mu/L$ as before.

Next, we again consider a variant constrained to an open, convex set $D \subset \mathbb{R}^n$.

Theorem 5.4. *Assume $f \in \mathcal{F}_{\mu,L}(D)$ for some open convex set D. Let $x_0 \in D$ so that $B(x_0, 2\|x_0 - x_*\|) \subset D$. If $x_1 = x_0 - \gamma d$, with $\|d - g(x_0)\| \leq \varepsilon\|g(x_0)\|$, $\varepsilon \leq \frac{2\kappa}{1+\kappa}$, and*

$$\gamma = \frac{2\mu - \varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)},$$

then

$$\begin{aligned} \|g_{x_0}(x_1)\|_{x_0} &\leq \left(\frac{1-\kappa}{1+\kappa} + \varepsilon\right)\|g_{x_0}(x_0)\|_{x_0}, \\ \|x_1 - x_*\|_{x_0} &\leq \left(\frac{1-\kappa}{1+\kappa} + \varepsilon\right)\|x_0 - x_*\|_{x_0}, \end{aligned}$$

where $\kappa = \mu/L$.

Proof. Note that the result follows from the proof of Theorem 5.3, provided that $x_1 \in D$. In other words, we need to show that the condition $x_1 \in D$ is redundant. This follows from:

$$\begin{aligned} \|x_1 - x_0\|_{x_0} &\leq \|x_1 - x_*\|_{x_0} + \|x_* - x_0\|_{x_0} \quad \text{(triangle inequality)} \\ &\leq \left(\frac{1-\kappa}{1+\kappa} + \varepsilon\right)\|x_0 - x_*\|_{x_0} + \|x_* - x_0\|_{x_0} \quad \text{(by Theorem 5.3)} \\ &\leq 2\|x_* - x_0\|_{x_0} \quad \text{(by } \varepsilon \leq \tfrac{2\kappa}{1+\kappa}\text{)}, \end{aligned}$$

which implies $x_1 \in D$ due to the assumption $B(x_0, 2\|x_0 - x_*\|) \subset D$. □

6 Implications for Newton's method for self-concordant f

Theorem 5.4 has interesting implications when minimizing a self-concordant function f with minimizer x_* by Newton's method. The implications become clear when fixing a point $x_0 \in D_f$, and using the inner product $\langle \cdot, \cdot \rangle_{x_0}$. Then the gradient at x_0 becomes $g_{x_0}(x_0) = H_{x_0}^{-1}(x_0)(g(x_0))$, which is minus the Newton direction at x_0. We will consider approximate Newton directions in the sense of (1), i.e. directions $-d$ that satisfy $\|d - g_{x_0}(x_0)\|_{x_0} \leq \varepsilon\|g_{x_0}(x_0)\|_{x_0}$, where $\varepsilon > 0$ is given. We only state results for the fixed step-length case, for later use. Similar results may be derived when using exact line search.

Corollary 6.1. *Assume f is self-concordant with minimizer x_*. Let $0 < \delta < 1$ be given and $x_0 \in D_f$ so that $\|x_0 - x_*\|_{x_0} \leq \frac{1}{2}\delta$. If If $x_1 = x_0 - \gamma d$, where $\|d - g_{x_0}(x_0)\|_{x_0} \leq \varepsilon\|g_{x_0}(x_0)\|_{x_0}$ with $\varepsilon \leq \frac{2(1-\delta)^4}{1+(1-\delta)^4}$, and*

$$\gamma = \frac{2(1-\delta)^4 - \varepsilon(1+(1-\delta)^4)}{(1-\varepsilon)(1-\delta)^2((1-\delta)^4+1)},$$

then

$$\begin{aligned} \|g_{x_0}(x_1)\|_{x_0} &\leq \left(\frac{1-\kappa_\delta}{1+\kappa_\delta} + \varepsilon\right)\|g_{x_0}(x_0)\|_{x_0}, \\ \|x_1 - x_*\|_{x_0} &\leq \left(\frac{1-\kappa_\delta}{1+\kappa_\delta} + \varepsilon\right)\|x_0 - x_*\|_{x_0}, \end{aligned}$$

where $\kappa_\delta = (1-\delta)^4$.

Proof. By Theorem 3.6, if we fix the inner product $\langle \cdot, \cdot \rangle_{x_0}$, then $f \in \mathcal{F}_{\mu,L}(B_{x_0}(x_0, \delta))$ with

$$\mu = (1-\delta)^2, \; L = \frac{1}{(1-\delta)^2}. \tag{10}$$

As a consequence $\kappa_\delta := \kappa = \mu/L = (1-\delta)^4$. (We use the notation $\kappa = \kappa_\delta$ to emphasize that κ depends on δ (only).) The required result now follows from Theorem 5.4. □

We note that, for $\varepsilon = 0$, the inequalities imply convergence whenever $\kappa_\delta > (6 - \sqrt{32})/2$, which is satisfied if $\delta \leq 0.3564$.

A final, but important observation is that the results in Corollary 6.1 remain valid if we use the $\langle \cdot, \cdot \rangle_{x_*}$ inner product, as opposed to $\langle \cdot, \cdot \rangle_{x_0}$. This implies that we (approximately) use the direction $-g_{x_*}(x_0) = -H^{-1}(x_*)g(x_0)$. Such a direction may seem to be of no practical use, since x_* is not known, but in the next section we will analyze an interior point method that uses precisely such search directions.

For easy reference, we therefore state the worst-case convergence result when using the $\langle \cdot, \cdot \rangle_{x_*}$ inner product.

Corollary 6.2. *Assume f is self-concordant with minimizer x_*. Let $0 < \delta < 1$ be given and $x_0 \in D_f$ so that $\|x_0 - x_*\|_{x_*} \leq \frac{1}{2}\delta$. If $x_1 = x_0 - \gamma d$, where $\|d - g_{x_*}(x_0)\|_{x_*} \leq \varepsilon \|g_{x_*}(x_0)\|_{x_*}$ with $\varepsilon \leq \frac{2(1-\delta)^4}{1+(1-\delta)^4}$, and step length*

$$\gamma = \frac{2(1-\delta)^4 - \varepsilon(1 + (1-\delta)^4)}{(1-\varepsilon)(1-\delta)^2((1-\delta)^4 + 1)},$$

then

$$\begin{aligned} \|g_{x_0}(x_1)\|_{x_*} &\leq \left(\frac{1-\kappa_\delta}{1+\kappa_\delta} + \varepsilon \right) \|g_{x_0}(x_0)\|_{x_*}, \\ \|x_1 - x_*\|_{x_*} &\leq \left(\frac{1-\kappa_\delta}{1+\kappa_\delta} + \varepsilon \right) \|x_0 - x_*\|_{x_*}, \end{aligned}$$

where $\kappa_\delta = (1-\delta)^4$.

7 Analysis of the method of Abernathy-Hazan

Given a convex body $\mathcal{K} \subset \mathbb{R}^n$ and a vector $\hat{\theta} \in \mathbb{R}^n$, the Abernathy and Hazan [1] describe an interior point method to solve the convex optimization problem

$$\min_{x \in \mathcal{K}} \hat{\theta}^{\mathsf{T}} x, \tag{11}$$

if one only has access to a membership oracle for $\mathcal{K}$. This method has generated much recent interest, since it is formally equivalent to a simulated annealing algorithm, and may be implemented by only sampling from $\mathcal{K}$. The interior point method in question used the so-called entropic (self-concordant) barrier function, introduced by Bubeck and Eldan [4], and we first review the necessary background.

7.1 Background on the entropic barrier method

The following discussion is condensed from [1].

The method is best described by considering the Boltzman probability distribution on $\mathcal{K}$:

$$P_\theta(x) := \exp(-\theta^\top x - A(\theta)) \quad \text{where} \quad A(\theta) := \ln \int_K \exp(-\theta^\top x') dx',$$

where $\theta = \eta\hat{\theta}$ for some fixed parameter $\eta > 0$. We write $X \sim P_\theta$ if the random variable X takes values in $\mathcal{K}$ according to the Boltzman probability distribution on $\mathcal{K}$ with density P_θ.

The convex function $A(\cdot)$ is known as the *log partition function*, and has derivatives:

$$\begin{aligned}\nabla A(\theta) &= -\mathbb{E}_{X\sim P_\theta}[X] \\ \nabla^2 A(\theta) &= \mathbb{E}_{X\sim P_\theta}[(X-\mathbb{E}_{X\sim P_\theta}[X])(X-\mathbb{E}_{X\sim P_\theta}[X])^\top].\end{aligned}$$

The Fenchel conjugate of $A(\theta)$ is

$$A^*(x) := \sup_{\theta\in\mathbb{R}^n} \theta^\top x - A(\theta).$$

The domain of $A^*(\cdot)$ is precisely the space of gradients of $A(\cdot)$, and this is the set $\mathrm{int}(-\mathcal{K})$.

The following key result shows that A^* provides a self-concordant barrier for the set $\mathcal{K}$.

Theorem 7.1 (Bubeck-Eldan [4])**.** *The function $x \mapsto A^*(-x)$ is a ϑ-self-concordant barrier function on $\mathcal{K}$ with $\vartheta \le n(1+o(1))$.*

The function $x \mapsto A^*(-x)$ is denoted by $A^*_-(\cdot)$ and called the entropic barrier for $\mathcal{K}$.

At every step of the associated interior point method, one wishes to minimize (approximately) a self-concordant function of the form

$$f(x) = \eta\hat{\theta}^\top x + A^*_-(x). \tag{12}$$

Subsequently the value of η is increased, and the process is repeated.

In keeping with our earlier notation, we denote the minimizer of f on $\mathcal{K}$ by x_*. Thus x_* is the point on the central path corresponding to the parameter η. We also assume a current iterate $x_0 \in \mathrm{int}(\mathcal{K})$ is available so that $\|x_* - x_0\|_{x_*} = \frac{1}{2}\delta < \frac{1}{2}$.

Abernathy and Hazan [1, Appendix C] propose to use the following direction to minimize f:

$$-d = -\nabla^2 f(x_*)^{-1}\nabla f(x_0). \tag{13}$$

The underlying idea is that $\nabla^2 f(x_*)^{-1}$ may be approximated to any given accuracy through sampling, based on the following result.

Lemma 7.2 ([4])**.** *One has*

$$\nabla^2 f(x_*)^{-1} = \nabla^2 A(\theta) = \mathbb{E}_{X\sim P_\theta}[(X-\mathbb{E}_{X\sim P_\theta}[X])(X-\mathbb{E}_{X\sim P_\theta}[X])^\top],$$

where $\theta = \eta\hat{\theta}$.

The proof follows immediately from the relationship between the hessians of a convex function and its conjugate, as given in [5].

Thus we may approximate $\nabla^2 f(x_*)^{-1}$ by an empirical covariance matrix as follows. If $X_i \sim P_\theta$ $(i = 1,\dots,N)$ are i.i.d., then we define the associated estimator of the covariance matrix of the X_i's as

$$\hat{\Sigma} := \frac{1}{N}\sum_{i=1}^N (X_i - \bar{X})(X_i-\bar{X})^\top \quad \text{where } \bar{X} = \frac{1}{N}\sum_{i=1}^N X_i. \tag{14}$$

The estimator $\hat{\Sigma}$ is known as the empirical covariance matrix, and it may be observed by sampling $X \sim P_\theta$. This may be done efficiently: for example, Lovász and Vempala [10] showed that one may sample (approximately) from log-concave distributions on compact bodies in polynomial time, by using the Markov-chain Monte-Carlo sampling method called hit-and-run, introduced by Smith [14].

The following concentration result (i.e. error bound) is known for the empirical covariance matrix.

Theorem 7.3 (cf. Theorems 4.1 and 4.2 in [3])**.** *Assume $\epsilon \in (0,1)$ and $X_i \sim P_\theta$ $(i = 1,\dots,N)$ are i.i.d. and $\hat{\Sigma}$ is the empirical covariance matrix in* (14)*. Then there exists a $C(\epsilon) > 0$ (depending only on ϵ), so that for $N \ge C(\epsilon)n$, the following holds with overwhelming probability:*

$$(1-\epsilon)y^\top\hat{\Sigma}y \le y^\top\Sigma y \le (1+\epsilon)y^\top\hat{\Sigma}y \qquad \forall y\in\mathbb{R}^n \tag{15}$$

$$(1-\epsilon)y^\top\hat{\Sigma}^{-1}y \le y^\top\Sigma^{-1}y \le (1+\epsilon)y^\top\hat{\Sigma}^{-1}y \qquad \forall y\in\mathbb{R}^n \tag{16}$$

where

$$\Sigma = \mathbb{E}_{X\sim P_\theta}[(X - \mathbb{E}_{X\sim P_\theta}[X])(X - \mathbb{E}_{X\sim P_\theta}[X])^\top \tag{17}$$

is the covariance matrix.

The phrase overwhelming probability in the theorem refers to the fact that the probability goes to 1 as n goes to infinity.

The exact details of hit-and-run sampling are outside the scope of this paper. For simplicity, we will therefore assume, in what follows, the availability of an approximate covariance matrix $\hat{\Sigma}$ that satisfies (15) and (16).

7.2 Analysis of the approximate direction

We can now show that an approximation of the search direction of Habernathy-Hazan (13) satisfies our 'approximate negative gradient' condition (1).

Theorem 7.4. *Let $\epsilon > 0$ be given, the covariance matrix Σ as in* (17)*, and a symmetric matrix $\hat{\Sigma}$ that approximates Σ as in* (15) *and* (16)*. Further, let f be as in* (12) *with minimizer x_* on a given convex body $\mathcal{K}$. Then the direction $-d = -\hat{\Sigma}\nabla f(x_0)$ at $x_0 \in \mathcal{K}$ satisfies*

$$\|\nabla^2 f(x_*)^{-1}\nabla f(x_0) - d\|_{x_*} \leq \sqrt{\frac{2\epsilon}{1-\epsilon}}\|\nabla^2 f(x_*)^{-1}\nabla f(x_0)\|_{x_*}.$$

In other words, one has $\|g_{x_*}(x_0) - d\|_{x_*} \leq \varepsilon\|g_{x_*}(x_0)\|_{x_*}$ where $\varepsilon = \sqrt{\frac{2\epsilon}{1-\epsilon}}$, i.e. condition (1) *holds for the inner product $\langle\cdot,\cdot\rangle_{x_*}$, when the reference inner product $\langle\cdot,\cdot\rangle$ is the Euclidean dot product.*

Proof. We fix the reference inner product $\langle\cdot,\cdot\rangle$ as the Euclidean dot product, so that $H(x_*) = \nabla^2 f(x_*) = \Sigma^{-1}$ and $g(x_0) = \nabla f(x_0)$. One has

$$\begin{aligned}
\|H^{-1}(x_*)g(x_0) - d\|_{x_*}^2 &= \langle H^{-1}(x_*)g(x_0) - \hat{\Sigma}g(x_0), H^{-1}(x_*)g(x_0) - \hat{\Sigma}g(x_0)\rangle_{x_*} \\
&= \langle H^{-1}(x_*)g(x_0) - \hat{\Sigma}g(x_0), g(x_0) - H(x_*)\hat{\Sigma}g(x_0)]\rangle \\
&= g(x_0)^\top H^{-1}(x_*)g(x_0) - 2g(x_0)^\top\hat{\Sigma}g(x_0) + [\hat{\Sigma}g(x_0)]^\top H(x_*)[\hat{\Sigma}g(x_0)] \\
&\leq (1+\epsilon)g(x_0)^\top\hat{\Sigma}g(x_0) - 2g(x_0)^\top\hat{\Sigma}g(x_0) + (1+\epsilon)[\hat{\Sigma}g(x_0)]^\top\hat{\Sigma}^{-1}[\hat{\Sigma}g(x_0)] \\
&= 2\epsilon\cdot g(x_0)^\top\hat{\Sigma}g(x_0),
\end{aligned}$$

where the inequality is from (15) and (16). Finally, using (15) once more, one obtains

$$\begin{aligned}
\|H^{-1}(x_*)g(x_0) - d\|_{x_*}^2 &\leq \frac{2\epsilon}{1-\epsilon}g(x_0)^\top H^{-1}(x_*)g(x_0) \\
&= \frac{2\epsilon}{1-\epsilon}\|g_{x_*}(x_0)\|_{x_*}^2,
\end{aligned}$$

as required. □

We need to consider another variant of the search direction in (13), since $\nabla f(x_0)$ will not be available exactly in general. Indeed, one can only obtain $\nabla f(x_0)$ approximately via the relation

$$\nabla f(x_0) = \eta\hat{\theta} + \nabla A_-^*(x_0) = \eta\hat{\theta} + \arg\max_{\theta\in\mathbb{R}^n}\left[-\theta^\top x_0 - A(\theta)\right],$$

where the last equality follows from the relationship between first derivatives of conjugate functions.

Thus $\nabla f(x_0)$ may be approximated by solving an unconstrained concave maximization problem in θ approximately, and, for this purpose, one may use the derivatives of $A(\theta)$ as given above. In particular, we will assume that we have available a $\tilde{g}(x_0) \approx \nabla f(x_0)$ in the sense that

$$\|\tilde{g}_{x_*}(x_0) - g_{x_*}(x_0)\|_{x_*} \leq \epsilon'\|g_{x_*}(x_0)\|_{x_*}, \tag{18}$$

where $\tilde{g}_{x_*}(x_0) := \Sigma\tilde{g}(x_0)$, $g_{x_*}(x_0) = \Sigma\nabla f(x_0)$ as before, and $\epsilon' > 0$ is given.

Thus we will consider the search direction

$$-\tilde{d} := -\hat{\Sigma}\tilde{g}(x_0) \approx -\Sigma\nabla f(x_0). \tag{19}$$

Corollary 7.5. *Under the assumptions of Theorem 7.4, define for a given $\epsilon' > 0$, the direction $-\tilde{d}$ at $x_0 \in D_f$ as in* (19)*, where $\tilde{g}(x_0) \approx \nabla f(x_0)$ satisfies* (18)*. Then one has*

$$\|\tilde{d} - g_{x_*}(x_0)\|_{x_*} \leq \left(\epsilon' \cdot \sqrt{\frac{1+\epsilon}{1-\epsilon}} + \sqrt{\frac{2\epsilon}{1-\epsilon}}\right) \|g_{x_*}(x_0)\|_{x_*}.$$

In other words, one has $\|g_{x_*}(x_0) - \tilde{d}\|_{x_*} \leq \varepsilon\|g_{x_*}(x_0)\|_{x_*}$ where $\varepsilon = \epsilon' \cdot \sqrt{\frac{1+\epsilon}{1-\epsilon}} + \sqrt{\frac{2\epsilon}{1-\epsilon}}$, i.e. condition (1) *holds for the inner product $\langle\cdot,\cdot\rangle_{x_*}$, when the reference inner product $\langle\cdot,\cdot\rangle$ is the Euclidean dot product.*

Proof. Recall, the notation $d = \hat{\Sigma}\nabla f(x_0)$ from Theorem 7.4, and note that, by definition,

$$\begin{aligned}
\|\tilde{d} - d\|_{x_*}^2 &= \|\hat{\Sigma}\left(\tilde{g}(x_0) - g(x_0)\right)\|_{x_*}^2 \\
&= \langle\hat{\Sigma}(\tilde{g}(x_0) - g(x_0), \Sigma^{-1}\hat{\Sigma}(\tilde{g}(x_0) - g(x_0)\rangle \\
&\leq (1+\epsilon)(\tilde{g}(x_0) - g(x_0))^\top\hat{\Sigma}(\tilde{g}(x_0) - g(x_0)) \qquad \text{(by (16))} \\
&\leq \left(\frac{1+\epsilon}{1-\epsilon}\right)\|\tilde{g}_{x_*}(x_0) - g_{x_*}(x_0)\|_{x_*}^2 \qquad \text{(by (15))} \\
&\leq (\epsilon')^2 \cdot \frac{1+\epsilon}{1-\epsilon}\|g_{x_*}(x_0)\|_{x_*}^2 \qquad \text{(by (18)).}
\end{aligned}$$

To complete the proof now only requires the triangle inequality,

$$\|g_{x_*}(x_0) - \tilde{d}\|_{x_*} \leq \|g_{x_*}(x_0) - d\|_{x_*} + \|d - \tilde{d}\|_{x_*},$$

as well as the inequality from Theorem 7.4. □

7.3 Complexity of a short-step interior point method

We now sketch how to bound the worst-case iteration complexity of a short-step interior point method using the entropic barrier, that is similar to the algorithm outlined in [1, Algorithm 3 and Appendix C]. In particular, is uses the search direction (19), described above.

The key observation is that one may analyse the complexity of interior point methods by only analysing the progress during one iteration; see e.g. [13, §2.4]. Thus our analysis of the previous section may be applied readily.

In what follows we need to modify our previous notation slightly to account for the changing value of the parameter η in the interior point algorithm. In particular, we will denote the point on the central path corresponding to a given $\eta > 0$, by $x(\eta)$ (as opposed to x_*). In other words,

$$x(\eta) = \arg\min_{x\in\mathcal{K}} \eta\hat{\theta}^\top x + A_-^*(x).$$

We may now state the short-step, interior point method that we will analyse as Algorithm 1.

Data: Tolerances: $\epsilon, \epsilon', \bar{\epsilon} > 0$; Proximity to central path parameter: $\delta \in (0,1)$; Entropic barrier parameter: $1 \leq \vartheta \leq n + o(1)$; Objective vector $\hat{\theta} \in \mathbb{R}^n$; an $x_0 \in \mathcal{K}$ and $\eta_0 > 0$ such that $\|x_0 - x(\eta_0)\|_{x(\eta_0)} \leq \frac{1}{2}\delta$.
Result: $\bar{\epsilon}$-optimal solution to $\min_{x \in \mathcal{K}} \hat{\theta}^\top x$
$\varepsilon = \epsilon' \cdot \sqrt{\frac{1+\epsilon}{1-\epsilon}} + \sqrt{\frac{2\epsilon}{1-\epsilon}}$;
Fixed step length: $\gamma = \frac{2(1-\delta)^4 - \varepsilon(1+(1-\delta)^4)}{(1-\varepsilon)(1-\delta)^2((1-\delta)^4+1)}$;
Iteration: $k = 0$;
while $\frac{\vartheta}{\eta_k} > \bar{\epsilon}$ **do**
  Compute $\hat{\Sigma}$ that satisfies (15) and (16);
  Compute $\tilde{g}(x_k)$ that satisfies (18);
  $\tilde{d} = \hat{\Sigma}\tilde{g}(x_k)$;
  $x_{k+1} = x_k - \gamma\tilde{d}$;
  $\eta_{k+1} = \left(1 + \frac{1}{16\sqrt{\vartheta}}\right)\eta_k$;
  $k \leftarrow k + 1$;
end

Algorithm 1: Short-step interior point method using the entropic barrier

We will show the following worst-case iteration complexity result.

Theorem 7.6. *Consider Algorithm 1 with the following input parameter settings: $\epsilon > 0$, and $\epsilon' > 0$ any values such that $\varepsilon = \epsilon' \cdot \sqrt{\frac{1+\epsilon}{1-\epsilon}} + \sqrt{\frac{2\epsilon}{1-\epsilon}} \leq \frac{1}{32}$, and $\delta = \frac{1}{4}$. If the algorithm is initialized with an $x_0 \in \mathcal{K}$ and $\eta_0 > 0$ such that $\|(x_0) - x(\eta_0)\|_{x(\eta_0)} \leq \frac{1}{2}\delta$, then it terminates after at most*

$$k = \left\lceil 20\sqrt{\vartheta} \ln\left(\frac{\vartheta}{\eta_0 \bar{\epsilon}}\right)\right\rceil$$

iterations. The result is an $x_k \in \mathcal{K}$ such that

$$\hat{\theta}^\top x_k - \min_{x \in \mathcal{K}} \hat{\theta}^\top x \leq \bar{\epsilon}.$$

Proof. The proof follows the usual lines of analysis of short-step interior point methods; in particular we will repeatedly refer to Renegar [13, §2.4]. We only need to show that, at the start of each iteration k, one has

$$\|x_k - x(\eta_k)\|_{x(\eta_k)} \leq \frac{1}{2}\delta.$$

Since, on the central path one has $\hat{\theta}^\top x(\eta) - \min_{x \in \mathcal{K}} \hat{\theta}^\top x \leq \vartheta/\eta$, the required result will then follow in the usual way (following the proof of relation (2.18) in [13, p. 47]).

Without loss of generality we therefore only consider the first iteration, with a given $x_0 \in \mathcal{K}$ and $\eta_0 > 0$ such that $\|x_0 - x(\eta_0)\|_{x(\eta_0)} \leq \frac{1}{2}\delta$, and proceed to show that $\|x_1 - x(\eta_1)\|_{x(\eta_1)} \leq \frac{1}{2}\delta$.

First, we bound the difference between the successive 'target' points on the central path, namely $x(\eta_0)$ and $x(\eta_1)$, where $\eta_1 = \left(1 + \frac{k}{\sqrt{\vartheta}}\right)\eta_0$ with $k = 1/16$. By the same argument as in [13, p. 46], one obtains:

$$\begin{aligned} \|x(\eta_1) - x(\eta_0)\|_{x(\eta_0)} &\leq k + \frac{3k^2}{(1-k)^3} \\ &\leq 0.0767 \text{ for } k = 1/16. \end{aligned}$$

Moreover, by Corollary 6.2,

$$\begin{aligned} \|x_1 - x(\eta_0)\|_{x(\eta_0)} &\leq \left(\frac{1-(1-\delta)^4}{1+(1-\delta)^4} + \varepsilon\right)\|x_0 - x(\eta_0)\|_{x(\eta_0)} \\ &\leq 0.1596 \cdot \frac{1}{2}\delta \leq 0.02. \end{aligned}$$

Using the triangle inequality,

$$\begin{aligned}\|x_1 - x(\eta_1)\|_{x(\eta_0)} &\leq \|x_1 - x(\eta_0)\|_{x(\eta_0)} + \|x(\eta_1) - x(\eta_0)\|_{x(\eta_0)} \\ &\leq 0.02 + 0.0767 = 0.0967.\end{aligned}$$

Finally, by the definition of self-concordance, one has

$$\|x_1 - x(\eta_1)\|_{x(\eta_1)} \leq \frac{\|x_1 - x(\eta_1)\|_{x(\eta_0)}}{1 - \|x(\eta_0) - x(\eta_1)\|_{x(\eta_0)}} \leq \frac{0.0967}{1 - 0.0767} \leq 0.1047 < \frac{1}{2}\delta,$$

as required. □

8 Concluding remarks

An unresolved question in our analysis is to understand the class of functions where the following inequality holds

$$f(y) - f(x) - \langle g(x), y - x\rangle \geq \frac{1}{L}\|g(y) - g(x)\|^2$$

for all x, y in a given open convex set D, where $L > 0$ is fixed. In particular, does it hold for all $f \in \mathcal{F}_{0,L}(D)$?

A recent, related result by Azagra and Mudarra [2], shows that, if f satisfies this inequality on a given subset of $\mathbb{R}^n$, then f has an extension to some $\tilde{f} \in \mathcal{F}_{0,L}(\mathbb{R}^n)$, in the sense that the values and gradients of f and $\tilde{f}$ coincide on the given subset. This implies that if a function satisfies inequality (4) on some subset of $\mathbb{R}^n$, then it has an extension to $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$. It is not clear though, if any $f \in \mathcal{F}_{\mu,L}(D)$, for given open convex D, has an extension to $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$.

References

[1] J. Abernethy and E. Hazan. Faster Convex Optimization: Simulated Annealing with an Efficient Universal Barrier. arXiv 1507.02528, July 2015.

[2] D. Azagra, and C. Mudarra. An Extension Theorem for convex functions of class $C^{1,1}$ on Hilbert spaces. *Journal of Mathematical Analysis and Applications*, 446.2, 1167–1182, 2017.

[3] R. Adamczak, A.E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the AMS*, 23(2), 535-561, 2010.

[4] S. Bubeck and R. Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. *arXiv:1412.1587*, 2014.

[5] J.P. Crouzeix. A relationship between the second derivatives of a convex function and of its conjugate. *Mathematical Programming*, 13 364-365, 1977.

[6] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

[7] A. T. Kalai and S. Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2), 253–266 (2006)

[8] E. de Klerk, F. Glineur and A.B. Taylor. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *arXiv*, 1606.09365, 2016.

[9] J. Li, M.S. Andersen, and L. Vandenberghe. Inexact proximal Newton methods for self-concordant functions. Mathematical Methods of Operations Research, to appear. DOI 10.1007/s00186-016-0566-9

[10] L. Lovasz and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307-358, 2007.

[11] Yu. Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., 2004.

[12] Yu. Nesterov and A.S. Nemirovski, *Interior point polynomial algorithms in convex programming*. SIAM, 1994.

[13] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*, SIAM, 2001.

[14] R. Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions, *Operations Research* 32(6), 1296-1308, 1984.

[15] A.B. Taylor, J.M. Hendrickx, and F. Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *arXiv*, 1512.07516v2, 2016.

Appendix: proofs

Proof of Theorem 5.3

Convergence of gradient norm

As in the proof of Theorem 5.1, consider the following inequalities along with their associated multipliers:

$$\begin{aligned}
&\langle g_0 - g_1, x_0 - x_1\rangle \geq \frac{1}{1-\frac{\mu}{L}}\left(\frac{1}{L}\|g_0-g_1\|^2 + \mu\|x_1-x_0\|^2 - 2\frac{\mu}{L}\langle g_0-g_1, x_0-x_1\rangle\right) && : \lambda_0,\\
&\|d-g_0\|^2 - \varepsilon^2\|g_0\|^2 \leq 0 && : \lambda_1.
\end{aligned}$$

In the following developments, we will also use the form of the algorithm:

$$x_1 = x_0 - \gamma d,$$

and the notation $\rho_\varepsilon(\gamma) := 1-(1-\varepsilon)\mu\gamma$. Recall that we want to prove that the rate $\rho_\varepsilon^2(\gamma)$ is valid on the interval

$$\gamma \in \left[0, \frac{2\mu - \varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)}\right],$$

when $\frac{2\mu-\varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)} \geq 0 \Leftrightarrow \varepsilon \leq \frac{2\mu}{L+\mu}$ (that is, we only consider $\gamma \geq 0$). We use the following values for the multipliers:

$$\begin{aligned}
\lambda_0 &= \frac{2}{\gamma(1-\varepsilon)}\rho_\varepsilon(\gamma),\\
\lambda_1 &= \frac{\gamma\mu}{\varepsilon}\rho_\varepsilon(\gamma).
\end{aligned}$$

In that case, one can write the weighted sum of the previous constraints in the following form:

$$\begin{aligned}
\rho_\varepsilon^2(\gamma)\|g_0\|^2 \geq \|g_1\|^2 &+ \frac{2-(1-\varepsilon)\gamma(L+\mu)}{(1-\varepsilon)\gamma(L-\mu)}\left\|\frac{\gamma(L+\mu)((\varepsilon-1)\gamma\mu+1)}{(\varepsilon-1)\gamma(L+\mu)+2}d - \frac{2((\varepsilon-1)\gamma\mu+1)}{(1-\varepsilon)\gamma(L+\mu)+2}g_0 + g_1\right\|^2\\
&+ \frac{(1-\varepsilon)\gamma(1-(1-\varepsilon)\gamma\mu)\Big(-(1-\varepsilon)\gamma\mu(L+\mu) - \varepsilon(L+\mu) + 2\mu\Big)}{\varepsilon(2-(1-\varepsilon)\gamma(L+\mu))}\left\|\frac{1}{\varepsilon-1}d + g_0\right\|^2.
\end{aligned}$$

Therefore, the guarantee

$$\rho_\varepsilon^2(\gamma)\|g_0\|^2 \geq \|g_1\|^2$$

is valid as long as both the Lagrange multipliers, and the coefficients of the norms in the previous expression are nonnegative. That is, under the following conditions:

- the Lagrange multipliers are nonnegative as long as $\rho_\varepsilon(\gamma) \geq 0$, that is, when

$$\gamma \leq \frac{1}{(1-\varepsilon)\mu},$$

which is valid for all values of γ in the interval of interest (see below).

- The coefficients of the norms are also nonnegative, since

$$\begin{aligned} 2-(1-\varepsilon)\gamma(L+\mu) \geq 0 &\Leftrightarrow \gamma \leq \frac{2}{(1-\varepsilon)(L+\mu)}, \\ (1-(1-\varepsilon)\gamma\mu) \geq 0 &\Leftrightarrow \gamma \leq \frac{1}{(1-\varepsilon)\mu}, \\ \Big(-(1-\varepsilon)\gamma\mu(L+\mu)-\varepsilon(L+\mu)+2\mu\Big) \geq 0 &\Leftrightarrow \gamma \leq \frac{2\mu-\varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)}, \end{aligned}$$

which are all valid on the interval of interest for γ, as:

$$\frac{2\mu-\varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)} = \frac{2}{(1-\varepsilon)(L+\mu)} - \frac{\varepsilon}{(1-\varepsilon)\mu} \leq \frac{2}{(1-\varepsilon)(L+\mu)} \leq \frac{1}{(1-\varepsilon)\mu}.$$

Convergence of distance to optimality

Consider the following inequalities and the associated multipliers:

$$\begin{aligned} &\frac{1}{1-\frac{\mu}{L}}\left(\frac{1}{L}\|g_0\|^2+\mu\|x_0-x_*\|^2-2\frac{\mu}{L}\langle g_0, x_0-x_*\rangle\right)+\langle g_0, x_*-x_0\rangle \leq 0 && : \lambda_0, \\ &\|d-g_0\|^2-\varepsilon^2\|g_0\|^2 \leq 0 && : \lambda_1. \end{aligned}$$

As in the case of the gradient norm, we use the notation $\rho_\varepsilon(\gamma) := 1-(1-\varepsilon)\mu\gamma$. Let us recall that we want to prove that the rate $\rho_\varepsilon^2(\gamma)$ is valid on the interval

$$\gamma \in \left[0, \frac{2\mu-\varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)}\right],$$

when $\frac{2\mu-\varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)} \geq 0 \Leftrightarrow \varepsilon \leq \frac{2\mu}{L+\mu}$ (we only consider $\gamma \geq 0$). We now use the following values for the multipliers:

$$\begin{aligned} \lambda_0 &= 2\gamma(1-\varepsilon)\rho_\varepsilon(\gamma), \\ \lambda_1 &= \frac{\gamma}{\mu\varepsilon}\rho_\varepsilon(\gamma). \end{aligned}$$

In that case, one can write the weighted sum of the previous constraints in the following form:

$$\begin{aligned} &(1-\gamma\mu(1-\varepsilon))^2\|x_0-x_*\| \geq \|x_1-x_*\| \\ &+\gamma\mu^2(1-\varepsilon)\frac{2-\gamma(1-\varepsilon)(L+\mu)}{L-\mu}\left\|\frac{L-\mu}{(1-\varepsilon)\mu^2(2-\gamma(1-\varepsilon)(L+\mu))}d-\frac{(L+\mu)(1-\gamma\mu(1-\varepsilon))}{\mu^2(2-\gamma(1-\varepsilon)(L+\mu))}g_0+x_0-x_*\right\|^2 \\ &+\gamma\frac{(1-\gamma\mu(1-\varepsilon))(2\mu-\varepsilon(L+\mu)-\gamma\mu(1-\varepsilon)(L+\mu))}{\varepsilon\mu^2(1-\varepsilon)(2-\gamma(1-\varepsilon)(L+\mu))}\|d-(1-\varepsilon)g_0\|^2 \end{aligned}$$

Hence, all coefficients and multipliers are positive as long as

$$\begin{aligned} &2\mu-\varepsilon(L+\mu)-\gamma\mu(1-\varepsilon)(L+\mu) \geq 0 \Leftrightarrow \gamma \leq \frac{\mu-\varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)}, \\ &1-(1-\varepsilon)\mu\gamma \geq 0 \Leftrightarrow \gamma \leq \frac{1}{(1-\varepsilon)\mu}, \\ &2-\gamma(1-\varepsilon)(L+\mu) \geq 0 \Leftrightarrow \gamma \leq \frac{2}{(1-\varepsilon)(L+\mu)}. \end{aligned}$$

We refer to previous discussions for the details leading to the conclusion:

$$(1-\gamma\mu(1-\varepsilon))^2\|x_0-x_*\| \geq \|x_1-x_*\|.$$

Convergence of function values

As in the previous section, we use the notation $\rho_\varepsilon(\gamma) := 1-(1-\varepsilon)\mu\gamma$, and consider the case

$$\gamma \in \left[0, \frac{2\mu-\varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)}\right].$$

For proving the desired convergence rate in terms of function values, we consider the following set of inequalities (and associated multipliers):

$$\begin{aligned}
&f_0 - f_1 - \langle g_1, x_0 - x_1\rangle \\
&\qquad \geq \frac{1}{2(1-\mu/L)}\left(\frac{1}{L}\|g_0-g_1\|^2 + \mu\|x_0-x_1\|^2 - 2\frac{\mu}{L}\langle g_1-g_0, x_1-x_0\rangle\right) && : \lambda_{01} = \rho_\varepsilon(\gamma),\\
&f_* - f_0 - \langle g_0, x_* - x_0\rangle \\
&\qquad \geq \frac{1}{2(1-\mu/L)}\left(\frac{1}{L}\|g_*-g_0\|^2 + \mu\|x_*-x_0\|^2 - 2\frac{\mu}{L}\langle g_0-g_*, x_0-x_*\rangle\right) && : \lambda_{*0} = \rho_\varepsilon(\gamma)(1-\rho_\varepsilon(\gamma)),\\
&f_* - f_1 - \langle g_1, x_* - x_1\rangle \\
&\qquad \geq \frac{1}{2(1-\mu/L)}\left(\frac{1}{L}\|g_*-g_1\|^2 + \mu\|x_*-x_1\|^2 - 2\frac{\mu}{L}\langle g_1-g_*, x_1-x_*\rangle\right) && : \lambda_{*1} = 1-\rho_\varepsilon(\gamma),\\
&\|d-g_0\|^2 - \varepsilon^2\|g_0\|^2 \leq 0 && : \lambda_2 = \frac{\gamma}{2\varepsilon}\rho_\varepsilon(\gamma).
\end{aligned}$$

Note that $\rho_\varepsilon(\gamma) \leq 1$ and that the multipliers are nonnegative in the cases of interest. We can write the weighted sum of the previous constraints in the following form :

$$\begin{aligned}
&\rho_\varepsilon^2(\gamma)(f(x_0)-f(x_*))\\
&\geq f(x_1)-f(x_*)\\
&+\frac{\gamma\rho_\varepsilon(\gamma)(L(-2\varepsilon\gamma\mu+\rho_\varepsilon(\gamma)-1)+\mu(\rho_\varepsilon(\gamma)+1))}{2\varepsilon(L(\rho_\varepsilon(\gamma)-1)+\mu(\rho_\varepsilon(\gamma)+1))}\left\|d+\frac{g_0((\varepsilon+1)L(\rho_\varepsilon(\gamma)-1)-(\varepsilon-1)\mu(\rho_\varepsilon(\gamma)+1))}{L(2\varepsilon\gamma\mu-\rho_\varepsilon(\gamma)+1)-\mu(\rho_\varepsilon(\gamma)+1)}\right\|^2\\
&+\frac{L\mu\left(1-\rho_\varepsilon^2(\gamma)\right)}{2(L-\mu)}\left\|-\frac{d\gamma}{\rho_\varepsilon(\gamma)+1}-\frac{g_0\rho_\varepsilon(\gamma)}{\mu\rho_\varepsilon(\gamma)+\mu}-\frac{g_1}{\mu\rho_\varepsilon(\gamma)+\mu}+x_0-x_*\right\|^2\\
&+\frac{\rho_\varepsilon(\gamma)(L+\mu)-(L-\mu)}{2\mu(\rho_\varepsilon(\gamma)+1)(L-\mu)}\left\|\frac{2d\gamma L\mu\rho_\varepsilon(\gamma)}{L(\rho_\varepsilon(\gamma)-1)+\mu(\rho_\varepsilon(\gamma)+1)}+\frac{g_0\rho_\varepsilon(\gamma)(L(\rho_\varepsilon(\gamma)-1)-\mu(\rho_\varepsilon(\gamma)+1))}{L(\rho_\varepsilon(\gamma)-1)+\mu(\rho_\varepsilon(\gamma)+1)}+g_1\right\|^2\\
&-\frac{\rho_\varepsilon(\gamma)((\varepsilon+1)\gamma L-\rho_\varepsilon(\gamma)-1)((\varepsilon-1)\gamma\mu-\rho_\varepsilon(\gamma)+1)}{L(2\varepsilon\gamma\mu-\rho_\varepsilon(\gamma)+1)-\mu(\rho_\varepsilon(\gamma)+1)}\|g_0\|^2\\
&\geq f(x_1)-f(x_*).
\end{aligned}$$

In order to prove the last inequality, we have to show that the coefficients of the norms of the decomposition are nonnegative.

- Term 1: substituting $\rho_\varepsilon(\gamma)$ by its expression, nonnegativity of the coefficient follows from

$$\begin{aligned}
\gamma &\leq \frac{1}{(1-\varepsilon)\mu}\\
\gamma &\leq \frac{2-\varepsilon(L-\mu)\left(L\mu(1-\varepsilon^2)\right)^{-1/2}}{(L+\mu)}\\
\gamma &\leq \frac{2}{(1-\varepsilon)(L+\mu)}
\end{aligned}$$

which hold, as $\gamma \leq \frac{2-\varepsilon(L-\mu)\left(L\mu(1-\varepsilon^2)\right)^{-1/2}}{(L+\mu)} \leq \frac{2}{(1-\varepsilon)(L+\mu)} \leq \frac{1}{(1-\varepsilon)\mu}$ on the interval of interest for γ.

- Term 2: always nonnegative as $0 \leq \rho_\varepsilon(\gamma) \leq 1$ on the interval of interest for γ.
- Term 3: substituting $\rho_\varepsilon(\gamma)$ by its expression, one can easily verify that the coefficient is positive when

$$\gamma \leq \frac{2}{(1-\varepsilon)(L+\mu)},$$

  which is true on the interval of interest for γ.
- Term 4: cancels out by substituting $\rho_\varepsilon(\gamma)$ by its expression.