HAL
archives-ouvertes.fr

Worst-Case Convergence Analysis of Inexact Gradient and Newton Methods Through Semidefinite Programming Performance Estimation

Etienne de Klerk, François Glineur, Adrien Taylor

▶ **To cite this version:**

Etienne de Klerk, François Glineur, Adrien Taylor. Worst-Case Convergence Analysis of Inexact Gradient and Newton Methods Through Semidefinite Programming Performance Estimation. SIAM Journal on Optimization, Society for Industrial and Applied Mathematics, 2020, 30 (3), pp.2053-2082. 10.1137/19M1281368 . hal-02956367

HAL Id: hal-02956367
https://hal.inria.fr/hal-02956367

Submitted on 7 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1709.05191v2 [math.OC] 22 Jun 2020

WORST-CASE CONVERGENCE ANALYSIS OF INEXACT GRADIENT AND NEWTON METHODS THROUGH SEMIDEFINITE PROGRAMMING PERFORMANCE ESTIMATION*

ETIENNE DE KLERK†, FRANÇOIS GLINEUR‡, AND ADRIEN B. TAYLOR§

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, including the case of inexact search directions. The analysis uses semidefinite programming performance estimation, as pioneered by Drori and 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*, 11(7), 1185-1199, 2017]. To illustrate the applicability of the tools, we demonstrate a novel complexity analysis of short step interior point methods using inexact search directions. As an example in this framework, we sketch how to give a rigorous worst-case complexity analysis of a recent interior point method by Abernethy and Hazan [*PMLR*, 48:2520–2528, 2016].

Key words. performance estimation problems, gradient method, inexact search direction, semidefinite programming, interior point methods

AMS subject classifications. 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[1] and relies on semidefinite programming (SDP) performance estimation problems, as introduced by Drori and Teboulle [15]. As a result, we develop a set of tools that may be used to design or analyse a wide range of interior point (and other) algorithms. Our analysis is in fact an extension of the worst-case analysis of the gradient method in [22], combined with the fact that Newton's method may be viewed as a gradient method with respect to a suitable local (intrinsic) inner product. As a result, we obtain worst-case convergence results for a single iteration of a wide range of methods, including Newton's method and the steepest descent method.

Related work. Our work is similar in spirit to recent analysis by Li et al [24] of inexact proximal Newton methods for self-concordant functions, but our approach and results are different. Their approach is oriented toward inexactness from the difficulty of computing the proximal Newton step, whereas ours is oriented toward the difficulty of computing the Hessian. Also, the authors of [24] do not use SDP performance estimation.

Since the seminal work by Drori and Teboulle [15], several authors have extended the SDP performance estimation framework. The authors of [35] introduced tightness guarantees for smooth (strongly) convex optimization, and for larger classes of problems in [36] (where a list of sufficient conditions for applying the methodology is provided). It was also used to deal with nonsmooth problems [11, 36], monotone inclusions and variational inequalities [16, 17, 21, 31], and even to study fixed-point iterations of non-expansive operators [25]. Fixed-step gradient descent was among the first algorithms to be studied with this methodology in different settings: for (possibly composite) smooth (possibly strongly) convex optimization [14, 15, 35, 36], and its line-search version was studied using the same methodology in [22]. The performance estimation framework was also used for obtaining new methods with optimized worst-case performance guarantees in different settings [11, 13, 19, 20]. In particular, such new methods were obtained by optimization of their algorithmic parameters in [11, 19, 20], and by analogy with conjugate-gradient type methods (doing greedy span-searches) in [13]. Performance estimation is also related to the line of work on *integral quadratic constraints* started by Lessard, Recht, and Packard [23], which also allowed designing optimized methods, see [8, 38]. The approach in [23] may be seen as a (relaxed) version of an SDP performance estimation problem where Lyapunov functions are used to certify error bounds [37].

Outline and contributions of this paper. The contribution of this work is two-fold:

1. We extend the SDP performance estimation framework to include smooth, strongly convex functions that are

* **Funding:** François Glineur is supported by the Belgian Interuniversity Attraction Poles, and by the ARC grant 13/18-054 (Communauté française de Belgique). Adrien Taylor acknowledges support from the European Research Council (grant SEQUOIA 724063).

† Tilburg University, The Netherlands, E.deKlerk@uvt.nl.

‡ UCL / CORE and ICTEAM, Louvain-la-Neuve, Belgium, Francois.Glineur@uclouvain.be.

§ INRIA, Département d'informatique de l'ENS, Ecole normale supérieure, CNRS, PSL Research University, Paris, France, Adrien.Taylor@inria.fr.

[1] Our analysis is computer-assisted in the following sense: computation was used to identify proofs, that could subsequently be verified in a standard mathematically rigorous way. Hence our results do not rely on the outcome of computations.

not defined over the entire $\mathbb{R}^n$. By considering arbitrary inner products, we unify the analysis for gradient descent-type methods and Newton's method, thus extending the results in [22] in a significant way. In particular, we are able to give a new error analysis of inexact Newton methods for self-concordant functions. Thus we provide the first extension of SDP performance estimation to second order methods.

2. As an application of the tools we develop, we give an analysis of inexact short step interior point methods. As an example of how our analysis may be used, we sketch 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 recent interest, since it links simulated annealing and interior point methods. However, no detailed analysis of the method is provided in [1], and we supply some crucial details that are missing there.

In Section 2, we review some basics on gradient descent methods, coordinate-free calculus, and convex functions. Thereafter, in Section 3, we describe inequalities for various classes of convex functions that will be used in performance estimation problems. A novel aspect here is that we allow the convex domain to be arbitrary in the study of smooth, strongly convex functions. Section 4 introduces the SDP performance estimation problems for various classes of convex functions and domains, and the error bounds from the analytical solutions of the performance estimation are given in Section 5. In Section 6 we show that our general framework for smooth, strongly convex functions includes self-concordant functions, and we are thus able to study inexact Newton(-type) methods for self-concordant functions. We then switch to an application of the tools we have developed, namely the complexity analysis of short step interior point methods that use inexact search directions (Section 7). As an example of inexact interior point methods, we consider a recent algorithm by Abernethy and Hazan [1], that uses inexact Newton-type directions for the so-called entropic (self-concordant) barrier; see Section 7.1. We stress that this example only serves to illustrate our results, and we do not give the full details of the analysis, which is beyond the scope of our paper. The additional analysis on how sampling may be used in the method by Abernethy and Hazan [1] to compute the inexact Newton-type directions is given in the separate work [4].

2. Preliminaries. Throughout f denotes a differentiable convex function, whose domain is denoted by $D_f \subset \mathbb{R}^n$, whereas D is used to denote open sets that may not correspond to the domain of f. We will indicate additional assumptions on f as needed. We will mostly use the notation from the book by Renegar [30], for easy reference.

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

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

If f is twice differentiable, the second derivative (or Hessian) of f at x is defined as the (unique) linear operator $H(x)$ that satisfies

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

Note that $g(x)$, and therefore also $H(x)$, depend on the reference inner product. If $\langle \cdot, \cdot \rangle$ is the Euclidean dot product then $g(x) = \nabla f(x) = \left[\frac{\partial f(x)}{\partial x_i} \right]_{i=1,\ldots,n}$, and the Hessian, when written as a matrix with respect to the standard basis, takes the familiar form $[H(x)]_{ij} =: [\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ $(i, j \in \{1, \ldots, n\})$.

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 to $B^{-1}g(x)$, and the Hessian at x changes to $B^{-1}H(x)$.

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 point 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* $\langle \cdot, \cdot \rangle$. The induced norm for the intrinsic inner product is denoted by: $\|u\|_x = \sqrt{\langle u, u \rangle_x}$. For the intrinsic inner product, 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 [30], and all proofs may be found there.

THEOREM 2.2 (Theorem 1.5.1 in [30]). *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 2.3. *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-valued integrals.

THEOREM 2.4 (Proposition 1.5.4 in [30]). *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 require the following version of the fundamental theorem for gradients.

THEOREM 2.5 (Theorem 1.5.6 in [30]). *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. 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 $\varepsilon \in [0, 1)$:

$$\|d_i - g(x_i)\| \leq \varepsilon \|g(x_i)\| \quad i = 0, 1, \ldots \tag{2.1}$$

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

Algorithm 2.1 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 (2.1)
Choose a step size $\gamma > 0$
Update $x_{i+1} = x_i - \gamma d_i$

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

1. Exact line search: $\gamma = \text{argmin}_{\gamma \in \mathbb{R}} f(x_i - \gamma d_i)$;
2. Fixed step size: 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 (if $\varepsilon = 0$). More generally, 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 the Newton direction when $B = H(x_i)$, a quasi-Newton direction if $B \approx H(x_i)$, and the familiar steepest descent direction $-\nabla f(x_i)$ if $B = I$.

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, by extending some known results on the first class.

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

$$f(y) \geq f(x) + \langle g(x), y - x \rangle \quad \forall x, y \in D. \tag{3.1}$$

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. For our purposes, it is important not to fix the inner product a priori.

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 $\mathcal{F}_{\mu,L}(D)$. The following lemma refines and extends well-known results from the literature on smooth convex functions. The novelty here is that we allow for arbitrary inner products, and any open, convex domain D. Similar results, are given, for example in the textbook by Nesterov [27, Chapter 2], and we only prove the results not covered there. We will use the Löwner partial order notation: for symmetric matrices (or self-adjoint linear operators) A and B, '$A \preceq B$' means that $B - A$ is positive semidefinite. We will denote the identity matrix (or operator) by I.

LEMMA 3.1. *Let D be an open convex set and $f : D \to \mathbb{R}$ 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. $(a) \Rightarrow (b)$: The proof of this implication is similar to that of [27, Theorems 2.1.5 and 2.1.6] (see relation (2.1.16) there), and is therefore omitted here.

$(b) \Rightarrow (c)$: If $\|H(x)\| \leq L$, then $H(x) - \frac{1}{L}H^2(x) \succeq 0$ for all $x \in D$, and Theorem 2.5 implies

$$\begin{aligned}
\langle g(y) - g(x), y - x \rangle &= \left\langle y - x, \int_0^1 H(x + t(y - x))(y - x)dt \right\rangle \\
&= \int_0^1 \langle y - x, H(x + t(y - x))(y - x) \rangle dt \\
&\geq \int_0^1 \langle y - x, \frac{1}{L}H^2(x + t(y - x))(y - x) \rangle dt \\
&= \frac{1}{L}\int_0^1 \|H(x + t(y - x))(y - x)\|^2 dt \\
&\geq \frac{1}{L}\left(\int_0^1 \|H(x + t(y - x))(y - x)\| dt\right)^2 \quad \text{(Jensen inequality)} \\
&\geq \frac{1}{L}\left\|\int_0^1 H(x + t(y - x))(y - x)dt\right\|^2 \quad \text{(Theorem 2.4)} \\
&= \frac{1}{L}\|g(y) - g(x)\|^2 \quad \text{(Theorem 2.5)}.
\end{aligned}$$

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

convexity, note that, by Theorem 2.2,

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

where the first inequality is from condition (c). Thus we obtain the convexity inequality (3.1). ∎

An interesting question is to understand the class of functions where the following inequality holds

$$f(y) - f(x) - \langle g(x), y - x\rangle \geq \frac{1}{2L}\|g(y) - g(x)\|^2 \tag{3.2}$$

for all x, y in a given open convex set D, where $L > 0$ is fixed. (Note that (3.2) implies condition (c) in the lemma, by adding (3.2) to itself after interchanging x and y.)

Indeed, for performance estimation problems, one attempts to find a function from a specific class that corresponds to the worst-case input for a given iterative algorithm. It is therefore of both practical and theoretical interest to understand the function class that satisfies (3.2).

The inequality (3.2) is known to hold for $D = \mathbb{R}^n$ if, and only if, $f \in \mathcal{F}_{0,L}(\mathbb{R}^n)$, by results of Taylor, Glineur and Hendrickx [35] (see also Azagra and Mudarra [2]). In an earlier version of this paper, we asked if this is true for more general open convex sets $D \subset \mathbb{R}^n$, but this turns out not to be the case: Drori [12] recently constructed an example of a bivariate function with open domain, say D, such that $f \in \mathcal{F}_{0,L}(D)$ with $L = 1$, but where (3.2) does not hold for all $x, y \in D$. For this reason, item (c) in Lemma 3.1, cannot be changed to the stronger condition (3.2).

Lemma 3.1 allows us to derive the following necessary and sufficient conditions for membership of $\mathcal{F}_{\mu,L}(D)$. The condition (d) below is new, and will be used extensively in the proofs that follow.

THEOREM 3.2. *Let D be an open convex set and $f : D \to \mathbb{R}$ 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

$$\langle g(x) - g(y), x - y\rangle \geq \frac{1}{1 - \frac{\mu}{L}}\left(\frac{1}{L}\|g(x) - g(y)\|^2 + \mu\|x - y\|^2 - 2\frac{\mu}{L}\langle g(x) - g(y), x - y\rangle\right). \tag{3.3}$$

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

$$f(y) - f(x) - \langle g(x), y - x\rangle \geq \frac{1}{2(1 - \frac{\mu}{L})}\left(\frac{1}{L}\|g(y) - g(x)\|^2 + \mu\|x - y\|^2 - 2\frac{\mu}{L}\langle g(x) - g(y), x - y\rangle\right). \tag{3.4}$$

For the reasons discussed after Lemma 3.1, the inequality (3.4) does not hold in general if $D \neq \mathbb{R}^n$, due to the results by Drori [12].

3.3. Self-concordance. Self-concordant functions are special convex functions introduced by Nesterov and Nemirovski [28], that play a central role in the analysis of interior point algorithms. We will use the (slightly) more general definition by Renegar [30].

DEFINITION 3.3 (Self-concordant functional). *Let $f : D_f \to \mathbb{R}$ be such that $H(x) \succ 0$ for all $x \in D_f$, and let $B_x(x,1)$ denote the open unit ball centered at $x \in D_f$ 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*

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

An equivalent characterization of self-concordance, due to Renegar [30], is as follows.

THEOREM 3.4 (Theorem 2.2.1 in [30]). *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 $y \in B_x(x,1)$:*

$$\max\left\{\|H_x(y)\|_x, \|H_x(y)^{-1}\|_x\right\} \leq \frac{1}{(1-\|y-x\|_x)^2}. \tag{3.5}$$

This alternative characterization allows us to establish a new link between self-concordant and L-smooth, μ-strongly convex functions.

COROLLARY 3.5. *Assume $f : D_f \to \mathbb{R}$ is self-concordant, $x \in D_f$, and $\delta < 1$. Then $f \in F_{\mu,L}(D)$ with respect to the inner product $\langle\cdot,\cdot\rangle_x$, if $D = \{y \mid \|y-x\|_x < \delta\} = B_x(x,\delta)$, $\mu = (1-\delta)^2$, and $L = \frac{1}{(1-\delta)^2}$.*

Conversely, assume $f : D_f \to \mathbb{R}$ is twice continuously differentiable, and $H(x) \succ 0$ for all $x \in D_f$. If, for all $x \in D_f$ and $\delta \in (0,1)$, it holds that

1. *$D := B_x(x,\delta) \subset D_f$,*
2. *$f \in F_{\mu,L}(D)$ with respect to the inner product $\langle\cdot,\cdot\rangle_x$, and $\mu = (1-\delta)^2$ and $L = \frac{1}{(1-\delta)^2}$,*

then f is self-concordant.

Proof. For the first implication, observe that for a self-concordant f, Theorem 3.4 implies 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 $\left[(1-\delta)^2, \frac{1}{(1-\delta)^2}\right]$ for all $y \in B_x(x,\delta)$. Theorem 3.2 now yields the required result.

To prove the converse, assume $y \in B_x(x,1)$ and set $\delta = \|x-y\|_x$, $D = B_x(x,\delta)$, $\mu = (1-\delta)^2$, and $L = \frac{1}{(1-\delta)^2}$.

Since $f \in F_{\mu,L}(D)$ by assumption, it holds that

$$\mu I \preceq H_x(y) \preceq LI,$$

which is the same as the condition (3.5) that guarantees self-concordance. ∎

4. Performance estimation problems. Performance estimation problems, as introduced by Drori and Teboulle [15], 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 employ to study worst-case bounds for one iteration of gradient methods, applied to a smooth, strongly convex f that admits a minimizer, denoted by x_*. These 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 correspond to $g(x_i)$ ($i \in \{*,0,1\}$)), and f_i 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 identify the worst-case after one iteration is performed, namely to find the maximum value of either $f_1 - f_*$, $\|g_1\|$, or $\|x_1 - x_*\|$, given an upper bound on the initial value of one of the quantities $f_0 - f_*$, $\|g_0\|$, or $\|x_0 - x_*\|$ (this upper bound will be denoted R below).

Note again that the norm may be any induced norm on $\mathbb{R}^n$.

Note that we only consider one iteration, i.e. the transition from x_0 to x_1, whereas SDP performance estimation is typically used to study several iterations; see e.g. [15, 35]. For our purposes, it suffices to consider one iteration — we will elaborate on this later on. For the time being, we consider functions where the domain is all of $\mathbb{R}^n$, and will move to restricted domains later on.

Performance estimation with exact line search.

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

Worst-case function value.

(4.1)
$$\left.\begin{array}{ll} \max & f_1 - f_* \\ \text{s.t.} & f_i - f_j - \langle g_j, x_i - x_j \rangle \geq \frac{1}{2(1-\mu/L)} \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - 2\frac{\mu}{L} \langle g_j - g_i, x_j - x_i \rangle \right) \ \forall i, j \in S \\ & g_* = 0 \\ & \langle x_1 - x_0, g_1 \rangle = 0 \\ & \langle g_0, g_1 \rangle \leq \varepsilon \|g_0\| \|g_1\| \\ & f_0 - f_* \leq R \end{array}\right\}$$

The first constraint corresponds to (3.4), 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 (2.1) 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:

(4.2)
$$\begin{pmatrix} \varepsilon \|g_0\|^2 & \langle g_0, g_1 \rangle \\ \langle g_0, g_1 \rangle & \varepsilon \|g_1\|^2 \end{pmatrix} \succeq 0.$$

Also note that the optimal value of the resulting SDP problem is independent of the inner product. The SDP problem (4.1) was first studied in [22].

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

(4.3)
$$\left.\begin{array}{ll} \max & \|g_1\|^2 \\ \text{s.t.} & \langle g_i - g_j, x_i - x_j \rangle \geq \frac{1}{1-\frac{\mu}{L}} \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - 2\frac{\mu}{L} \langle g_i - g_j, x_i - x_j \rangle \right) \quad \forall i, j \in S \\ & g_* = 0 \\ & \langle x_1 - x_0, g_1 \rangle = 0 \\ & \langle g_0, g_1 \rangle \leq \varepsilon \|g_0\| \|g_1\| \\ & \|g_0\|^2 \leq R \end{array}\right\}$$

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

(4.4)
$$\left.\begin{array}{ll} \max & \|x_1 - x_*\|^2 \\ \text{s.t.} & \langle g_i - g_j, x_i - x_j \rangle \geq \frac{1}{1-\frac{\mu}{L}} \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - 2\frac{\mu}{L} \langle g_i - g_j, x_i - x_j \rangle \right) \quad \forall i, j \in S \\ & g_* = 0 \\ & \langle x_1 - x_0, g_1 \rangle = 0 \\ & \langle g_0, g_1 \rangle \leq \varepsilon \|g_0\| \|g_1\| \\ & \|x_0 - x_\star\|^2 \leq R \end{array}\right\}$$

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

PEP with fixed step sizes. For fixed step sizes, the performance estimation problems (4.1), (4.3), and (4.4) change as follows:

1. 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.
2. The vector d is viewed as a variable in the performance estimation problem.
3. The exact line search condition, $\langle x_1 - x_0, g_1 \rangle = 0$, is omitted.
4. The condition $\|d - g_0\|^2 \leq \varepsilon^2 \|g_0\|^2$ is added, that corresponds to (2.1), 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 rates of the gradient method for different performance measures. In this section, we provide bounds that were obtained using the solutions to the previously presented performance estimation problems. As the bounds we present below are not always the *tightest* ones, each result is provided with comments regarding tightness.

To the best of our knowledge, there are no results dealing with the exact same settings in the literature. Among others, one can find detailed analyses of first-order method under deterministic uncertainty models in [9, 10, 33]. Related results involving relative inaccuracy (as in this work) can be found in older works by Polyak [29] where the focus is on smooth convex minimization.

To give some meaningful comparison, we will compare our results to standard ones for the simpler case $\varepsilon = 0$.

5.1. PEP with exact line search. The first result concerns error bounds for the inexact gradient method with exact line search when no restriction is applied on the domain. This problem, originally due to Cauchy, was studied in detail in [22] using SDP performance estimation. Here we will generalize the main result from [22] to include arbitrary inner products. The Cauchy problem remains of enduring interest; see e.g. the recent work by Bolte and Pauwels [5, §5.3].

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 $\kappa := \frac{\mu}{L}$ (the inverse condition number: $\kappa \in (0,1]$) 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 (4.1), (4.3), and (4.4), respectively.

The first of the SDP performance estimation problems, namely (4.1), is exactly the same as the one in [22] (see (8) there). The first inequality therefore follows from Theorem 1.2 in [22].

It remains to demonstrate suitable upper bounds on the SDP performance estimation problems (4.3), and (4.4). This is done by aggregating the constraints of these respective SDP problems by using suitable (Lagrange) multipliers.[2]

To this end, consider the following constraints from (4.3) and their associated multipliers:

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

[2]The multipliers used in the proof were obtained by solving the SDP problems (4.3) and (4.4) numerically for different values of μ and L (one may assume w.l.o.g. that $R = 1$), and subsequently guessing the correct analytical expressions of the multipliers by looking at the optimal solution of the dual SDP problem. For example it quickly became clear that the second multiplier was $(L + \mu)$, and that the matrix S always had rank one. The correctness of these expressions for the multipliers is verified in the proof. Thus numerical computations were only used to find the proof of Theorem 5.1, i.e. to guess the correct analytical expressions for the multipliers, and play no role in the proof itself.

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

$$\begin{aligned}
\lambda &= \frac{2\varepsilon\sqrt{\kappa}}{\sqrt{1-\varepsilon^2}\,(1-\kappa)} + 1, \\
s_{11} &= \frac{3\varepsilon}{2} - \frac{\varepsilon\left(\kappa + \frac{1}{\kappa}\right)}{4} + \frac{1-\kappa}{2\sqrt{\kappa(1-\varepsilon^2)}} - \frac{\varepsilon^2(1-\kappa)}{\sqrt{\kappa(1-\varepsilon^2)}}, \\
s_{22} &= \frac{2\sqrt{\kappa(1-\varepsilon^2)} - \varepsilon(1-\kappa)}{(1-\varepsilon^2)(1-\kappa) + 2\varepsilon\sqrt{\kappa(1-\varepsilon^2)}}, \\
s_{12} &= \frac{\varepsilon\,(1-\kappa)}{2\sqrt{\kappa(1-\varepsilon^2)}} - 1.
\end{aligned}$$

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):

$$\begin{aligned}
0 \geq &(L-\mu)\lambda \left[\frac{1}{1-\frac{\mu}{L}} \left(\frac{1}{L}\|g_0 - g_1\|^2 + \mu\|x_1 - x_0\|^2 - 2\frac{\mu}{L}\langle g_0 - g_1, x_0 - x_1\rangle \right) - \langle g_0 - g_1, x_0 - x_1\rangle \right] \\
&+ (L+\mu)\left[\langle g_1, x_1 - x_0\rangle\right] - \mathrm{Trace}\left(\begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} \begin{pmatrix} \epsilon||g_0||^2 & \langle g_0, g_1\rangle \\ \langle g_0, g_1\rangle & \epsilon||g_1||^2 \end{pmatrix} \right).
\end{aligned} \tag{5.1}$$

We first show that multipliers are nonnegative (resp. positive semidefinite) where required; that is, $(L-\mu)\lambda \geq 0$ and $S \succeq 0$. The nonnegativity of the first term is clear from $0 < \mu \leq L$ and $\lambda \geq 0$. 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:

$$\det S = s_{11}s_{22} - s_{12}^2 = 0.$$

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 (5.1) is valid.

Reformulating the valid inequality (5.1) 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 (4.4) 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:

$$\begin{aligned}
\lambda_0&=\frac{1-\kappa}{\mu}\left[1-2\varepsilon^2+\frac{\varepsilon\sqrt{1-\varepsilon^2}}{2\sqrt{\kappa}(1-\kappa)}(-1-\kappa^2+6\kappa)\right],\\
\lambda_1&=\frac{1}{\mu}-\frac{1}{L},\\
\lambda_2&=\frac{1}{\mu}+\frac{1}{L},\\
L\mu s_{11}&=\frac{3\varepsilon}{2}-\frac{\varepsilon\left(\kappa+\frac{1}{\kappa}\right)}{4}+\frac{1-\kappa}{2\sqrt{\kappa(1-\varepsilon^2)}}-\frac{\varepsilon^2(1-\kappa)}{\sqrt{\kappa(1-\varepsilon^2)}},\\
L\mu s_{22}&=\frac{2\sqrt{\kappa(1-\varepsilon^2)}-\varepsilon(1-\kappa)}{(1-\varepsilon^2)(1-\kappa)+2\varepsilon\sqrt{\kappa(1-\varepsilon^2)}},\\
L\mu s_{12}&=\frac{\varepsilon(1-\kappa)}{2\sqrt{\kappa(1-\varepsilon^2)}}-1.
\end{aligned}$$

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 $S\succeq 0$.

Similarly to 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:

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

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. ∎

Theorem 5.1 provides both tight and non-tight results, as follows:

1. the result in function values cannot be improved, by [22, Example 5.2];
2. likewise, the result in gradient norm cannot be improved; we give an example proving this in Appendix A.
3. The result in distance to optimality is not tight.

Our bounds on the rates satisfy

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

where the left hand term is the optimal value of (4.1) and the right hand term is the optimal value of (4.3).

We may compare our results to known (classical) results with $\varepsilon = 0$. In that case, we have that the best possible rate for function value is given by (see [22, Theorem 5.2]),

$$f(x_1) - f(x_*) \leq \left(\frac{1-\kappa}{1+\kappa}\right)^2 (f(x_0) - f(x_*)).$$

By smoothness and strong convexity, we derive in a standard way for the gradient norm (the exact same reasoning holds for the distance to optimum)

$$||g(x_1)|| \leq \frac{1}{\sqrt{\kappa}}\left(\frac{1-\kappa}{1+\kappa}\right)||g(x_0)||,$$

whereas Theorem 5.1 provides the strictly better guarantee for $\kappa \in (0,1)$, namely:

$$||g(x_1)|| \leq \frac{1}{2\sqrt{\kappa}}(1-\kappa)\,||g(x_0)||.$$

The above rates are valid when performing one iteration. Better rates can be guaranteed if more than one iteration is performed, which can be done in the same framework. However, we do not pursue our investigations in that direction, as the subsequent analysis of Newtons method only requires the best possible one-iteration inequalities, as provided by the above new improved bounds.

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 twice continuously differentiable $f \in \mathcal{F}_{\mu,L}(D)$ 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 $\kappa := \frac{\mu}{L} \in (0,1]$ and $\varepsilon \in \left[0, \frac{2\sqrt{\kappa}}{1+\kappa}\right]$, one has*

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

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 (3.3), 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. ∎

Concerning tightness and comparisons with known results, the same remarks as for Theorem 5.1 apply here. Although the setting is nonstandard for first-order methods, comparisons made for the case $D = \mathbb{R}^n$ are still valid as the worst-case bounds for gradient and distance are the same (i.e., results of Theorem 5.2 already improve upon the literature/folklore knowledge in the simpler setting $D = \mathbb{R}^n$).

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 (3.4).

5.2. PEP with fixed step sizes. We now state a result that is similar to Theorem 5.1, but deals with fixed step sizes instead of exact line search.

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

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

Proof. The proof is similar to that of Theorem 5.1, and is sketched in Appendix B. ∎

In Theorem 5.3, all results are provably tight. Indeed, one can easily verify that those bounds are achieved with equality (for all three cases: function, distance and gradient) on the one-dimensional minimization problem $\min_{x\in\mathbb{R}} f(x)$ with the quadratic function $f(x) = \frac{\mu}{2}x^2$ and the search direction $-g(x_0)(1-\varepsilon) = -\mu(1-\varepsilon)x_0$ (that satisfies the relative accuracy criterion (2.1)). Note that, in the case $\varepsilon = 0$, one can therefore also recover all standard convergence guarantees that are tight on quadratics (see e.g., [34] and the references therein).

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, and f twice continuously differentiable. Let $x_0 \in D$ so that $B(x_0, 2\|x_0 - x_*\|) \subset D$. If $x_1 = x_0 - \gamma d$, with $\|d - g(x_0)\| \leq \varepsilon\|g(x_0)\|$, $\varepsilon \in \left[0, \frac{2\kappa}{1+\kappa}\right]$, and*

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

then

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

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 a consequence of the hypotheses. This follows from:

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

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

The same remarks as for Theorem 5.3 apply here: the results are tight on quadratics, as the worst-case bounds match those in the case $D = \mathbb{R}^d$.

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 the opposite of the Newton step at x_0. We will consider approximate Newton directions in the sense of (2.1), 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 error bounds can be obtained using Theorem 5.2 for inexact Newton methods with exact line search, that are used, e.g. in long step interior point methods with inexact search directions; see, e.g. [30, §2.5.3].

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 $x_1 = x_0 - \gamma d$, where $\|d - g_{x_0}(x_0)\|_{x_0} \leq \varepsilon \|g_{x_0}(x_0)\|_{x_0}$ with $\varepsilon \in \left[0, \frac{2(1-\delta)^4}{1+(1-\delta)^4}\right]$, and*

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

then

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

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

Proof. By Corollary 3.5, 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

$$\mu = (1-\delta)^2, \; L = \frac{1}{(1-\delta)^2}. \tag{6.1}$$

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. ∎

In view of our earlier remarks on tightness of the bounds in Theorem 5.4, it is important to note that the bounds in Corollary 6.1 are not tight in general. The reason is that we only used the fact that, for a given $x_0 \in D_f$ and $\delta \in (0,1)$, one has $f \in \mathcal{F}_{\mu,L}(B_{x_0}(x_0, \delta))$ for the values of μ and L as given in (6.1). This is weaker than requiring self-concordance of f, as the following example shows.

EXAMPLE 6.2. *Consider the univariate $f(x) = \frac{1}{12}x^4$ with $D_f = (0, \infty)$. At $x_0 = 1$, one has $H(x_0) = 1$. If we set $\delta = \frac{1}{2}$, (6.1) yields $\mu = \frac{1}{4}$ and $L = 4$, and we have $B_{x_0}(x_0, \delta) = \left(\frac{1}{2}, \frac{3}{2}\right)$. Since $H_{x_0}(y) = y^2$ for all $y \in \mathbb{R}$, one has $\mu < H_{x_0}(y) < L$ if $y \in B_{x_0}(x_0, \delta)$, and therefore $f \in \mathcal{F}_{\mu,L}(B_{x_0}(x_0, \delta))$. On the other hand, f is not self-concordant on its domain, since it does not satisfy the condition $|f'''(x)| \leq 2f''(x)^{3/2}$ if $x \in (0,1)$.*

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 6.3. *Assume f is self-concordant with minimizer x_*. Let $\delta \in (0,1)$ 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 $\varepsilon \in \left[0, \frac{2(1-\delta)^4}{1+(1-\delta)^4}\right]$, and step size*

$$\gamma = \frac{2(1-\delta)^4 - \varepsilon(1+(1-\delta)^4)}{(1-\varepsilon)(1-\delta)^2((1-\delta)^4+1)}, \tag{6.2}$$

then

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

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

We may compare Corollaries 6.1 and 6.3 to the following results that may be obtained from standard interior point analysis.

THEOREM 6.4 (Based on Theorems 1.6.2 and 2.2.3 in [30]). *Let f be a self-concordant function with minimizer x_*, and let $x_0 \in B_{x_0}(x_*, 1)$. Define $x_1 = x_0 - \gamma[H(x_0)^{-1}g(x_0) + e(x_0)]$ for some $\gamma \in (0,1)$, where $e(x_0)$ denotes an error in the Newton direction at the point x_0. If $\|e(x_0)\|_{x_0} \leq \varepsilon \|H(x_0)^{-1}g(x_0)\|_{x_0}$, then*

$$\|x_1 - x_*\|_{x_0} \leq \frac{(1-\gamma+\gamma^2\varepsilon)\|x_0 - x_*\|_{x_0} + \gamma\|x_0 - x_*\|_{x_0}^2}{\gamma(1 - \|x_0 - x_*\|_{x_0})}. \tag{6.3}$$

Similarly, if we define instead $x_1 = x_0 - \gamma[H(x_*)^{-1}g(x_0) + e(x_0)]$, i.e. replace $H(x_0)$ by $H(x_*)$ in the definition of x_1, then

$$\|x_1 - x_*\|_{x_*} \leq \frac{(1-\gamma+\gamma^2\varepsilon)\|x_0 - x_*\|_{x_*} + \gamma\|x_0 - x_*\|_{x_*}^2}{\gamma(1 - \|x_0 - x_*\|_{x_*})}, \tag{6.4}$$

under the assumption $x_0 \in B_{x_*}(x_*, 1)$.

Note that the only difference between the inequalities (6.3) and (6.4) is the choice of local norm.

To compare Theorem 6.4 to Corollaries 6.1 and 6.3, we present a plot of the respective upper bounds in Figure 1 for for different values of ε. The value of the step size γ is as in (6.2) with $\delta = 2\|x_0 - x_*\|_{x_*}$.

A few remarks on Figure 1:

1. Although the figure only compares inequality (6.4) in Theorem 6.4 to the bound in Corollary 6.3, the exact same plots remain valid when comparing the Newton direction bounds, namely inequality (6.3) in Theorem 6.4 to the bound in Corollary 6.1. The only difference is the scaling on the axes, since one should then switch from the $\|\cdot\|_{x_*}$ norm to the $\|\cdot\|_{x_0}$ norm. In this case the value γ is still given by (6.2), but with $\delta = 2\|x_0 - x_*\|_{x_0}$.
2. It is clear that our new bounds in Corollary 6.3 (and Corollary 6.1) improve on the known bounds in most cases. Even when $\varepsilon = 0$, we still improve if $\|x_0 - x_*\|_{x_*}$ is sufficiently large. As ε grows, our bounds clearly improve on those in Theorem 6.4.
3. In the figure, our new error bound remains bounded as the initial distance $\|x_0 - x_*\|_{x_*}$ approaches 1, but this is not the case for the bound from Theorem 6.4. Thus our new results capture a desirable feature of the convergence near the boundary of the Dikin ellipsoid.

7. Complexity of a short step interior point method using inexact search directions. We now sketch a proof of how to bound the worst-case iteration complexity of a short step interior point method using inexact search directions.

Given a convex body $\mathcal{K} \subset \mathbb{R}^n$ and a vector $\hat{\theta} \in \mathbb{R}^n$, we consider the convex optimization problem

$$\min_{x \in \mathcal{K}} \hat{\theta}^\mathsf{T} x. \tag{7.1}$$

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

DEFINITION 7.1 (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*

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

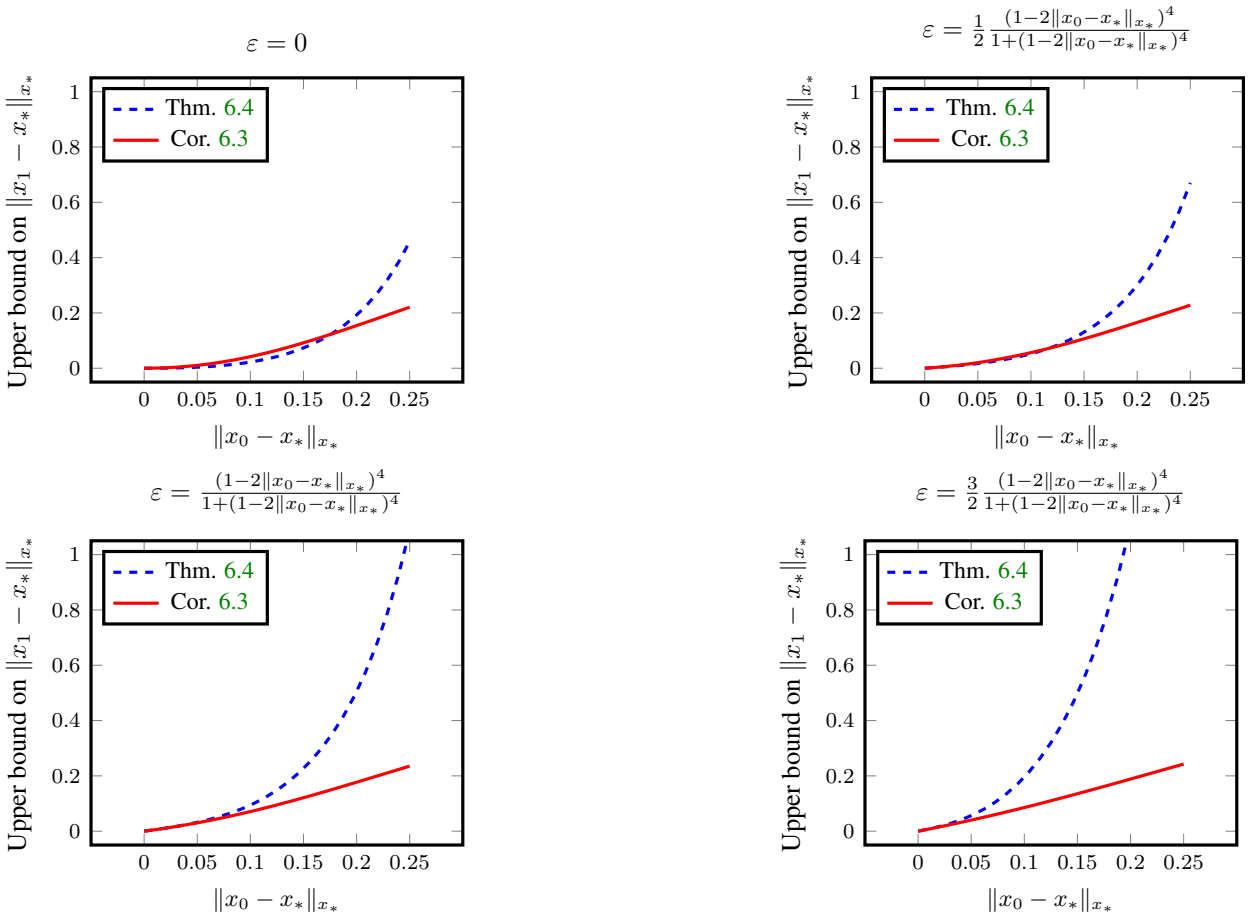

FIG. 1. *Upper bounds on* $\|x_1 - x_*\|_{x_*}$ *from Theorem 6.4 and Corollary 6.3.*

We will assume that we know a self-concordant barrier function with domain given by the interior of $\mathcal{K}$, say $f_{\mathcal{K}}$.

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. [30, §2.4]. Thus our analysis of the previous section may be applied readily. At each interior point iteration, one approximately minimizes a self-concordant function of the form

$$f(x) = \eta\hat{\theta}^\top x + f_{\mathcal{K}}(x), \tag{7.2}$$

where $\eta > 0$ is a fixed parameter. We denote its unique minimizer by $x(\eta)$, and call it the point on the *central path* corresponding to η. Subsequently, in the next interior point iteration, the value of η is increased, and the process is repeated.

We may now state two variants (A and B) of a short step, interior point method using inexact search directions (see Algorithm 7.1). Variant A corresponds to the short step interior point method analysed by Renegar [30, §2.4.2], but allows for inexact Newton directions. Variant B captures the framework of the interior point method of Abernethy and Hazan [1, Appendix D, supplementary material], to be discussed in Section 7.1.

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

THEOREM 7.2. *Consider Algorithm 7.1 with the following input parameter settings:* $\bar{\epsilon} > 0$, $0 \leq \varepsilon \leq \frac{1}{6}$, *and* $\delta = \frac{1}{4}$. *Both variants of the algorithm then terminate after at most*

$$N = \left\lceil 40\sqrt{\vartheta}\ln\left(\frac{6\vartheta}{5\eta_0\bar{\epsilon}}\right)\right\rceil$$

Algorithm 7.1 Short step interior point method using inexact directions (variants A and B)

Tolerances: $\varepsilon > 0$ (for search direction error), $\bar{\epsilon} > 0$ (for stopping criterion)
Proximity to central path parameter: $\delta \in (0, 1)$
Barrier parameter for $f_{\mathcal{K}}$: $\vartheta \geq 1$
Objective vector $\hat{\theta} \in \mathbb{R}^n$
Given an $x_0 \in \mathcal{K}$ and $\eta_0 > 0$ such that $\|x_0 - x(\eta_0)\|_{x(\eta_0)} \leq \frac{1}{2}\delta$ (variant A) or $\|x_0 - x(\eta_0)\|_{x_0} \leq \frac{1}{2}\delta$ (variant B).
Set the step size $\gamma = \frac{2(1-\delta)^4 - \varepsilon(1+(1-\delta)^4)}{(1-\varepsilon)(1-\delta)^2((1-\delta)^4+1)}$
Iteration: $k = 0$
while $\frac{\vartheta}{\eta_k} > \frac{5}{6}\bar{\epsilon}$ **do**
  compute d that satisfies $\|d - g_{x_k}(x_k)\|_{x_k} \leq \varepsilon\|g_{x_k}(x_k)\|_{x_k}$ (variant A) or $\|d - g_{x(\eta)}(x_k)\|_{x(\eta)} \leq \varepsilon\|g_{x(\eta)}(x_k)\|_{x(\eta)}$ (variant B)
  $x_{k+1} = x_k - \gamma d$
  $\eta_{k+1} = \left(1 + \frac{1}{32\sqrt{\vartheta}}\right)\eta_k$
  $k \leftarrow k + 1$
end while
return x_k an $\bar{\epsilon}$-optimal solution to $\min_{x\in\mathcal{K}} \hat{\theta}^\top x$

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

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

Proof. The proof follows the usual lines of analysis of short step interior point methods; in particular we will repeatedly refer to Renegar [30, §2.4]. We only analyse variant B of Algorithm 7.2, as the analysis of variant A is similar, but simpler.

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 [30, 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 $\eta_1 = \left(1 + \frac{\alpha}{\sqrt{\vartheta}}\right)\eta_0$ with $\alpha = 1/32$. By the same argument as in [30, p. 46], one obtains:

$$\begin{aligned}\|x(\eta_1) - x(\eta_0)\|_{x(\eta_0)} &\leq \alpha + \frac{3\alpha^2}{(1-\alpha)^3} \\ &\leq 0.0345 \text{ for } \alpha = 1/32.\end{aligned}$$

Moreover, by Corollary 6.3,

$$\begin{aligned}\|x_1 - x(\eta_0)\|_{x(\eta_0)} &\leq \left(\frac{1-(1-\delta)^4}{1+(1-\delta)^4} + \varepsilon\right)\|x_0 - x(\eta_0)\|_{x(\eta_0)} \\ &\leq 0.6860 \cdot \frac{1}{2}\delta \leq 0.0857.\end{aligned}$$

Using the triangle inequality,

$$\begin{aligned}\|x_1 - x(\eta_1)\|_{x(\eta_0)} &\leq \|x_1 - x(\eta_0)\|_{x(\eta_0)} + \|x(\eta_1) - x(\eta_0)\|_{x(\eta_0)} \\ &\leq 0.0857 + 0.0345 = 0.1202.\end{aligned}$$

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

$$\|x_1 - x(\eta_1)\|_{x(\eta_1)} \leq \frac{\|x_1 - x(\eta_1)\|_{x(\eta_0)}}{1 - \|x(\eta_0) - x(\eta_1)\|_{x(\eta_0)}} \leq \frac{0.1202}{1 - 0.0345} \leq 0.1245 < \frac{1}{2}\delta,$$

as required. □

It is insightful to note that, in the proof of Theorem 7.2, it would not suffice to use the classical bound from Theorem 6.4. Indeed, we used $\|x_1 - x(\eta_0)\|_{x(\eta_0)} \leq 0.0857$ in the proof, obtained from our new bound in Corollary 6.3. If we had used Theorem 6.4 instead, we would only obtain $\|x_1 - x(\eta_0)\|_{x(\eta_0)} \leq 0.1042$ (by using $\gamma = 0.67$), which would be too weak to complete the argument. Of course, one could prove a variation on Theorem 7.2 by using Theorem 6.4 and smaller values of δ and ε. Having said that, it is clear that our analysis adds in a meaningful way to the classical interior point analysis, removing the need to use weaker parameter values.

7.1. Analysis of the method of Abernathy-Hazan. Abernathy and Hazan [1] describe an interior point method to solve the convex optimization problem (7.1) if one only has access to a membership oracle for $\mathcal{K}$ (see Abernethy and Hazan [1, Appendix D, supplementary material]). As mentioned earlier, it falls within the framework of variant B of Algorithm 7.1 above.

This method has generated recent interest, since it is closely related to a simulated annealing algorithm, and may be implemented by only sampling from $\mathcal{K}$. Polynomial-time complexity of certain simulated annealing methods for convex optimization was first shown by Kalai and Vempala [18], and the link with interior point methods casts light on their result.

The interior point method in question used the so-called entropic (self-concordant) barrier function, introduced by Bubeck and Eldan [6], and we first review the necessary background.

7.1.1. 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 $\text{int}(-\mathcal{K})$.

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

THEOREM 7.3 (Bubeck-Eldan [6]). *The function $x \mapsto A^*(-x)$ is a ϑ-self-concordant barrier function on $\mathcal{K}$ with $\vartheta \leq n(1 + o(1))$.*

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 (7.2), where we now have the barrier function $f_K(x) = A^*_-(x)$.

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

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

$$-d = -\nabla^2 f(x_*)^{-1}\nabla f(x_0). \tag{7.3}$$

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 7.4 ([6]). *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 [7].

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

$$\hat{\Sigma} := \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^\top \quad \text{where } \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{7.4}$$

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 [26] 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 [32].

The following concentration result (i.e. error bound) is known for the empirical covariance matrix. We state this result to motivate our framework of analysis only — we will not use it.

THEOREM 7.5 (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.,*

$$\Sigma = \mathbb{E}_{X\sim P_\theta}\left[(X - \mathbb{E}_{X\sim P_\theta}[X])(X - \mathbb{E}_{X\sim P_\theta}[X])^\top\right] \tag{7.5}$$

is the covariance matrix, and $\hat{\Sigma}$ *is the empirical covariance matrix in* (7.4)*. Then there exist absolute constants* $c > 0$ *and* $C > 0$*, such that, for* $N \geq C\frac{\|\Sigma\|^2}{\epsilon^2}\log^2\left(\frac{2\|\Sigma\|^2}{\epsilon^2}\right) n$*, the following holds with probability at least* $1 - \exp(-c\sqrt{n})$*:*

$$(1-\epsilon)y^\top\hat{\Sigma}y \leq y^\top\Sigma y \leq (1+\epsilon)y^\top\hat{\Sigma}y \qquad \forall y \in \mathbb{R}^n \tag{7.6}$$

$$(1-\epsilon)y^\top\hat{\Sigma}^{-1}y \leq y^\top\Sigma^{-1}y \leq (1+\epsilon)y^\top\hat{\Sigma}^{-1}y \qquad \forall y \in \mathbb{R}^n. \tag{7.7}$$

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 (7.6) and (7.7). In other words, Theorem 7.5 only serves to motivate our assumption, we will not use it in our analysis. We give a full analysis of the sampling process in the separate work [4], where we show how to find a sample covariance matrix $\hat{\Sigma}$ that satisfies (7.6) and (7.7).

7.1.2. Analysis of the approximate direction in the Abernethy-Hazan algorithm. We can now show that an approximation of the search direction of Abernethy-Hazan (7.3) satisfies our 'approximate negative gradient' condition (2.1).

THEOREM 7.6. *Let* $\epsilon > 0$ *be given, the covariance matrix* Σ *as in* (7.5)*, and a symmetric matrix* $\hat{\Sigma}$ *that approximates* Σ *as in* (7.6) *and* (7.7)*. Further, let* f *be as in* (7.2) *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* (2.1) *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 (7.6) and (7.7). Finally, using (7.6) 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 (7.3), 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

$$\|\tilde{g}_{x_*}(x_0) - g_{x_*}(x_0)\|_{x_*} \leq \epsilon' \|g_{x_*}(x_0)\|_{x_*}, \tag{7.8}$$

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. To motivate our assumption, we note that the function $\theta \mapsto -\theta^\top x_0 - A(\theta)$ is self-concordant [6], and we may therefore use Corollary 6.1 to bound the complexity of approximating its maximizer. As with the Hessian approximation, we again omit the (lengthy) details; see Section 8 for a further discussion of our assumptions.

Thus we will consider the search direction

$$-\tilde{d} := -\hat{\Sigma}\tilde{g}(x_0) \approx -\Sigma\nabla f(x_0). \tag{7.9}$$

COROLLARY 7.7. *Under the assumptions of Theorem 7.6, define for a given $\epsilon' > 0$, the direction $-\tilde{d}$ at $x_0 \in D_f$ as in (7.9), where $\tilde{g}(x_0) \approx \nabla f(x_0)$ satisfies (7.8). 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 (2.1) 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 7.6, and note that, by definition,

$$\begin{aligned}
\|\tilde{d} - d\|_{x_*}^2 &= \|\hat{\Sigma}\left(\tilde{g}(x_0) - g(x_0)\right)\|_{x_*}^2 \\
&= \langle \hat{\Sigma}(\tilde{g}(x_0) - g(x_0), \Sigma^{-1}\hat{\Sigma}(\tilde{g}(x_0) - g(x_0)\rangle \\
&\leq (1+\epsilon)(\tilde{g}(x_0) - g(x_0))^\top \hat{\Sigma}(\tilde{g}(x_0) - g(x_0)) \qquad \text{(by (7.7))} \\
&\leq \left(\frac{1+\epsilon}{1-\epsilon}\right) \|\tilde{g}_{x_*}(x_0) - g_{x_*}(x_0)\|_{x_*}^2 \qquad \text{(by (7.6))} \\
&\leq (\epsilon')^2 \left(\frac{1+\epsilon}{1-\epsilon}\right) \|g_{x_*}(x_0)\|_{x_*}^2 \qquad \text{(by (7.8))}.
\end{aligned}$$

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 7.6. ∎

8. Concluding remarks. In this paper we have extended the SDP performance estimation analysis to second order methods, and demonstrated an example of how to use the resulting error bounds in the complexity analysis of inexact interior point methods. Our analysis of the interior point method of Abernethy and Hazan [1] gives an outline of a new proof of the polynomial complexity of the method. Having said that, we have assumed in this paper that the gradient and Hessian of the entropic barrier function may be computed within a fixed relative accuracy through sampling. It is possible to complete the analysis of the sampling process in the spirit of the work of Kalai and Vempala [18] and Lovász and Vempala [26]. This requires detailed analysis of the hit-and-run sampling method for log-concave distributions on convex bodies, combined with the self-concordance property of the entropic barrier function. Since the resulting analysis is lengthy, and of a very different type than that presented here, we present it in a separate work, namely [4].

Acknowledgement. Etienne de Klerk would like to thank Riley Badenbroek for pointing out the result in Theorem 6.4, and providing its proof, and for also pointing out a mistake in the proof of Theorem 7.2 in an earlier version of this paper.

References.

[1] J. Abernethy and E. Hazan. Faster Convex Optimization: Simulated Annealing with an Efficient Universal Barrier. Proceedings of The 33rd International Conference on Machine Learning, PMLR 48:2520–2528, 2016. http://proceedings.mlr.press/v48/abernethy16.html

[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] R. Badenbroek and E. de Klerk. Complexity analysis of a sampling-based interior point method for convex optimization. arXiv:1811.07677 [math.OC], 2018.

[5] J. Bolte and E. Pauwels Curiosities and counterexamples in smooth convex optimization. arXiv:2001.07999 [math.OC], 2018.

[6] S. Bubeck and R. Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. In: *Conference on Learning Theory*, 279–279, 2015.

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

[8] S. Cyrus, B. Hu, B. Van Scoy, and L. Lessard. A Robust Accelerated Optimization Algorithm for Strongly Convex Functions. *Proceedings of the 2018 Annual American Control Conference (ACC)*, pp. 1376–1381, 2018.

[9] A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3), 1171–1183, 2008.

[10] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2), 37–75, 2014.

[11] Y. Drori and M. Teboulle. An optimal variant of Kelley's cutting-plane method. *Mathematical Programming*, 160(1-2):321–351, 2016.

[12] Y. Drori. On the Properties of Convex Functions over Open Sets. arXiv:1812.02419 [math.OC], 2018.

[13] Y. Drori and A.B. Taylor. Efficient first-order methods for convex minimization: a constructive approach. *Mathematical Programming*, available online: https://doi.org/10.1007/s10107-019-01410-2

[14] Y. Drori. *Contributions to the Complexity Analysis of Optimization Algorithms*. PhD thesis, Tel-Aviv University, 2014.

[15] 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.

[16] G. Gu and J. Yang. Optimal nonergodic sublinear convergence rate of proximal point algorithm for maximal monotone inclusion problems. *arXiv:1904.05495 [Math.OC]*, 2019.

[17] G. Gu and J. Yang. On the optimal ergodic sublinear convergence rate of the relaxed proximal point algorithm for variational inequalities. *arXiv:1905.06030 [Math.OC]*, 2019.

[18] A. T. Kalai and S. Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2), 253–266, 2006.
[19] D. Kim and J.F. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1-2), 81–107, 2016.
[20] D. Kim and J. A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *arXiv:1803.06600 [Math.OC]*, 2018.
[21] D. Kim. Accelerated proximal point method for maximally monotone operators. *arXiv:1905.05149 [Math.OC]*, 2019.
[22] E. de Klerk, F. Glineur, and A.B. Taylor. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11(7), 1185–1199, 2017.
[23] L. Lessard, B. Recht, and A. Packard. Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints. *SIAM Journal on Optimization*, 26(1), 57–95, 2016.
[24] J. Li, M.S. Andersen, and L. Vandenberghe. Inexact proximal Newton methods for self-concordant functions. *Mathematical Methods of Operations Research*, 85, 19–41, 2017.
[25] F. Lieder. On the convergence rate of the Halpern-iteration. Technical report, 2017.
[26] L. Lovasz and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.
[27] Yu. Nesterov. *Lectures on convex optimization*, 2nd ed. Springer Optimization and Its Applications 137, Springer Nature Switzerland, 2018.
[28] Yu. Nesterov and A.S. Nemirovski, *Interior point polynomial algorithms in convex programming*. SIAM, 1994.
[29] B.T. Polyak. Convergence of methods of feasible directions in extremal problems. USSR Computational Mathematics and Mathematical Physics, 11(4), 53–70, 1971.
[30] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*, SIAM, 2001.
[31] E. K. Ryu, A. B. Taylor, C. Bergeling, and P. Giselsson. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *arXiv:1812.00146 [Math.OC]*, 2018.
[32] R. Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions, *Operations Research* 32(6), 1296–1308, 1984.
[33] M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, 1458–1466, 2011.
[34] A.B. Taylor, J.M. Hendrickx, and F. Glineur. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *Journal of Optimization Theory and Applications*, 178(2), 455–476, 2018.
[35] A.B. Taylor, J.M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017.
[36] A.B. Taylor, J.M. Hendrickx, and F. Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3), 1283–1313, 2017.
[37] A. Taylor, B. Van Scoy, and L. Lessard Lyapunov functions for first-order methods: Tight automated convergence guarantees. In Proceedings of the 35th International Conference on Machine Learning (ICML), pages 4897–4906, 2018.
[38] B. Van Scoy, R. A. Freeman, and K. M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2018.

Appendix A. Worst-case example for Theorem 5.1. Here, show that the bound

$$\|g(x_1)\| \leq \left(\varepsilon + \sqrt{1-\varepsilon^2}\tfrac{(1-\kappa)}{2\sqrt{\kappa}}\right)\|g(x_0)\|$$

from Theorem 5.1 cannot be improved on the class of smooth strongly convex functions, by giving an example of a function $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ where the bound holds with equality.

To this end, consider the two triplets $(x_0, g_0, f_0), (x_1, g_1, f_1) \in \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}$:

$$x_0 = \begin{bmatrix}0\\0\end{bmatrix}, \quad g_0 = \begin{bmatrix}1\\0\end{bmatrix}, \quad f_0 = 0,$$

and

$$x_1 = \begin{bmatrix} -\frac{\left(1-\epsilon^2\right)(1+\kappa)}{2\mu} \\ \frac{\epsilon\sqrt{1-\epsilon^2}(1+\kappa)}{2\mu} \end{bmatrix}, \quad g_1 = \begin{bmatrix} \frac{\sqrt{1-\epsilon^2}\epsilon(1-\kappa)}{2\sqrt{\kappa}} + \epsilon^2 \\ \frac{\left(1-\epsilon^2\right)(1-\kappa)}{2\sqrt{\kappa}} + \epsilon\sqrt{1-\epsilon^2} \end{bmatrix}, \quad f_1 = -\frac{\left(1-\epsilon^2\right)(1+\kappa)}{4\mu},$$

along with the inexact search direction $d \in \mathbb{R}^2$

$$d = \begin{bmatrix} 1-\epsilon^2 \\ -\epsilon\sqrt{1-\epsilon^2} \end{bmatrix},$$

where $\kappa := \frac{\mu}{L}$ is the (inverse) condition ratio, and $\epsilon \geq 0$. The following facts hold:

1. the pair of triplets satisfies the conditions

$$f_0 - f_1 + g_0^\mathsf{T} x_1 - x_0 + \tfrac{1}{2(1-\mu/L)}\left(\tfrac{1}{L}\|g_0 - g_1\|^2 + \mu\|x_0 - x_1\|^2 - 2\tfrac{\mu}{L}g_0 - g_1^\mathsf{T} x_0 - x_1\right) = 0,$$
$$f_1 - f_0 + g_1^\mathsf{T} x_0 - x_1 + \tfrac{1}{2(1-\mu/L)}\left(\tfrac{1}{L}\|g_0 - g_1\|^2 + \mu\|x_0 - x_1\|^2 - 2\tfrac{\mu}{L}g_0 - g_1^\mathsf{T} x_0 - x_1\right) = 0,$$

2. $x_1 = x_0 - \gamma d$ with $\gamma = \frac{1+\kappa}{2\mu}$,
3. $d^\mathsf{T} g_1 = 0$,
4. $\|g_0 - d\| = \varepsilon\|g_0\|$,
5. $\frac{\|g_1\|}{\|g_0\|} = \left(\varepsilon + \sqrt{1-\varepsilon^2}\frac{(1-\kappa)}{2\sqrt{\kappa}}\right)$.

Therefore, by [35, Theorem 4], there exists a function $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ satisfying

$$f(x_0) = f_0,\ f(x_1) = f_1,\ g(x_0) = g_0,\ g(x_1) = g_1.$$

For this function, one iteration of gradient descent with exact line search starting at x_0 achieves exactly

$$\|g(x_1)\| = \left(\varepsilon + \sqrt{1-\varepsilon^2}\frac{(1-\kappa)}{2\sqrt{\kappa}}\right)\|g(x_0)\|,$$

yielding the desired result.

Appendix B. 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:

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

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:

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

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

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

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

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

  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:

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

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:

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

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

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

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.

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

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

   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.
2. Term 2: always nonnegative as $0 \leq \rho_\varepsilon(\gamma) \leq 1$ on the interval of interest for γ.
3. 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.
4. Term 4: cancels out by substituting $\rho_\varepsilon(\gamma)$ by its expression.