

2016/27

# On the Worst-case Complexity of the Gradient Method with Exact Line Search for Smooth Strongly Convex Functions

ETIENNE DE KLERK, FRANÇOIS GLINEUR AND ADRIEN B. TAYLOR



50 YEARS OF  
**CORE**  
DISCUSSION PAPERS

## **CORE**

Voie du Roman Pays 34, L1.03.01

Tel (32 10) 47 43 04

Fax (32 10) 47 43 01

Email: [immaq-library@uclouvain.be](mailto:immaq-library@uclouvain.be)

<http://www.uclouvain.be/en-44508.html>

# On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions

Etienne de Klerk\*

François Glineur<sup>†</sup>

Adrien B. Taylor<sup>†</sup>

June 30, 2016

## Abstract

We consider the gradient (or steepest) descent method with exact line search applied to a strongly convex function with Lipschitz continuous gradient. We establish the exact worst-case rate of convergence of this scheme, and show that this worst-case behavior is exhibited by a certain convex quadratic function. We also extend the result to a noisy variant of gradient descent method, where exact line-search is performed in a search direction that differs from negative gradient by at most a prescribed relative tolerance.

The proof is computer-assisted, and relies on the resolution of semidefinite programming performance estimation problems as introduced in the paper [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].

**Keywords:** gradient method, steepest descent, semidefinite programming, performance estimation problem,

**AMS classification:** 90C25, 90C22, 90C20.

## 1 Introduction

The gradient (or steepest) descent method for unconstrained method was devised by Augustin-Louis Cauchy (1789–1857) in the 19th century, and remains one of the most iconic algorithms for unconstrained optimization. Indeed, it is usually the first algorithm that is taught during introductory courses on nonlinear optimization. It is therefore somewhat surprising that the worst-case convergence rate of the method is not yet precisely understood for smooth strongly convex functions.

In this paper, we settle the worst-case convergence rate question of the gradient descent method with exact line search for strongly convex, continuously differentiable functions  $f$  with Lipschitz continuous gradient. Formally we consider the following function class.

**Definition 1.1.** A continuously differentiable function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is called  $L$ -smooth,  $\mu$ -strongly convex with parameters  $L > 0$  and  $\mu > 0$  if

1.  $\mathbf{x} \mapsto f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$  is a convex function on  $\mathbb{R}^n$ , where the norm is the Euclidean norm;
2.  $\|\nabla f(\mathbf{x} + \Delta\mathbf{x}) - \nabla f(\mathbf{x})\| \leq L\|\Delta\mathbf{x}\|$  holds for all  $\mathbf{x} \in \mathbb{R}^n$  and  $\Delta\mathbf{x} \in \mathbb{R}^n$ .

The class of  $L$ -smooth,  $\mu$ -strongly convex functions will be denoted by  $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$ .

Note that, if  $f$  is twice continuously differentiable, then  $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$  is equivalent to

$$LI \succeq \nabla^2 f(\mathbf{x}) \succeq \mu I \quad \forall \mathbf{x} \in \mathbb{R}^n$$

---

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

<sup>†</sup>UCL/CORE and ICTEAM, Louvain-la-Neuve, Belgium, Francois.Glineur@uclouvain.be, Adrien.Taylor@uclouvain.be. A.B. Taylor is a F.R.I.A. fellow (F.R.S.-FNRS). The UCL/CORE authors are supported by the Belgian Interuniversity Attraction Poles, and by the ARC grant 13/18-054 (Communauté française de Belgique).

where the notation  $A \succeq B$  for symmetric matrices  $A$  and  $B$  means the matrix  $A - B$  is positive semidefinite, and  $I$  is the identity matrix. Equivalently, the eigenvalues of the Hessian  $\nabla^2 f(\mathbf{x})$  lie in the interval  $[\mu, L]$  for all  $\mathbf{x}$ .

The gradient method with exact line search may be described as follows.

**Gradient descent method with exact line search**

**Input:**  $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ ,  $\mathbf{x}_0 \in \mathbb{R}^n$ .

**for**  $i = 0, 1, \dots$

$\gamma = \operatorname{argmin}_{\gamma \in \mathbb{R}} f(\mathbf{x}_i - \gamma \nabla f(\mathbf{x}_i))$

$\mathbf{x}_{i+1} = \mathbf{x}_i - \gamma \nabla f(\mathbf{x}_i)$

Our main result may now be stated concisely.

**Theorem 1.2.** *Let  $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ ,  $\mathbf{x}_*$  a global minimizer of  $f$  on  $\mathbb{R}^n$ , and  $f_* = f(\mathbf{x}_*)$ . Each iteration of the gradient method with exact line search satisfies*

$$f(\mathbf{x}_{i+1}) - f_* \leq \left( \frac{L - \mu}{L + \mu} \right)^2 (f(\mathbf{x}_i) - f_*) \quad i = 0, 1, \dots \quad (1)$$

Note that the result in Theorem 1.2 is known for the case of quadratic functions in  $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$ , that is for functions of the form

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top Q \mathbf{x} + \mathbf{c}^\top \mathbf{x}$$

where  $\mathbf{c} \in \mathbb{R}^n$ , and the eigenvalues of the  $n \times n$  symmetric positive definite matrix  $Q$  lie in the interval  $[\mu, L]$ ; see e.g. [1, §1.2], [9, pp. 60–62], or [3, pp. 235–238]. Moreover, the bound (12) is known to be tight for the following example.

**Example 1.3.** *Consider the following quadratic function from [1, Example on p. 69]:*

$$f(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^n \lambda_i x_i^2$$

where

$$0 < \mu = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n = L,$$

and the starting point

$$\mathbf{x}_0 = (1/\mu, 0, \dots, 0, 1/L)^\top.$$

One may readily show that

$$\mathbf{x}_1 = \left( \frac{L - \mu}{L + \mu} \right) (1/\mu, 0, \dots, 0, -1/L)^\top,$$

and, for all  $i = 0, 1, \dots$

$$\mathbf{x}_{2i} = \left( \frac{L - \mu}{L + \mu} \right)^{2i} \mathbf{x}_0, \quad \mathbf{x}_{2i+1} = \left( \frac{L - \mu}{L + \mu} \right)^{2i} \mathbf{x}_1.$$

Since  $f_* = 0$ , it is straightforward to verify that equality

$$f(\mathbf{x}_{i+1}) - f_* = \left( \frac{L - \mu}{L + \mu} \right)^2 (f(\mathbf{x}_i) - f_*) \quad i = 0, 1, \dots,$$

holds as required. □

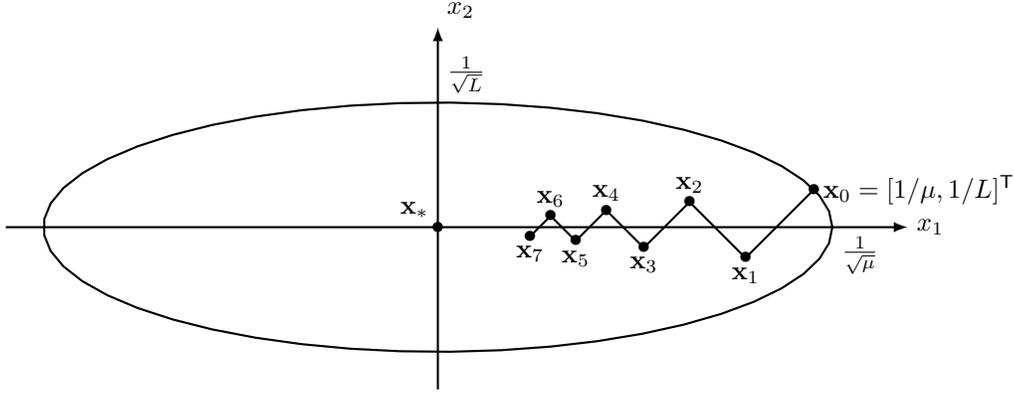


Figure 1: Illustration of the construction in Example 1.3 for the case  $n = 2$ .

The construction in Example 1.3 is illustrated in Figure 1.

As an immediate consequence of Theorem 1.2 and Example 1.3, one has the following tight bound on the number of steps needed to obtain  $\epsilon$ -relative accuracy on the objective function for a given  $\epsilon > 0$ .

**Corollary 1.4.** *Given  $\epsilon > 0$ , the gradient method with exact line search yields a solution with relative accuracy  $\epsilon$  for any function  $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$  after at most  $N = \left\lceil \frac{1}{2} \log\left(\frac{1}{\epsilon}\right) / \log\left(\frac{L+\mu}{L-\mu}\right) \right\rceil$  iterations, i.e.*

$$\frac{f(\mathbf{x}_N) - f_*}{f(\mathbf{x}_0) - f_*} \leq \epsilon,$$

where  $\mathbf{x}_0$  is the starting point. Moreover, this iteration bound is tight for the quadratic function defined in Example 1.3.

For non-quadratic functions in  $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$ , only bounds weaker than (12) are known. For example, in [3, p. 240], the following bound is shown:

$$(f(\mathbf{x}_{i+1}) - f_*) \leq \left(1 - \frac{\mu}{L}\right) (f(\mathbf{x}_i) - f_*) \quad i = 0, 1, \dots$$

In [8, Theorem 3.4] a stronger result than Theorem 1.2 was claimed, but this was retracted in a subsequent erratum<sup>1</sup>, and only an asymptotic result is claimed in the erratum.

A result related to Theorem 1.2 is given in [5] where Armijo-rule line search is used instead of exact line search. An explicit rate in the strongly convex case is given there in Proposition 3.3.5 on page 53 (definition of the method is (3.1.2) on page 44). More general upper bounds on the convergence rates of gradient-type methods for convex functions may be found in the books [6, 7].

## 2 Background results

In this section we collect some known results on strongly convex functions and on the gradient method. We will need these results in the proof of our main result, Theorem 1.2.

<sup>1</sup>The erratum is available at: <http://users.iems.northwestern.edu/~nocedal/book/2ndprint.pdf>

## 2.1 Properties of the gradient method with exact line search

Let  $\mathbf{x}_i$  ( $i = 1, 2, \dots, N$ ) be the iterates produced by the gradient method with exact line search started at  $\mathbf{x}_0$ . Those iterates are defined by the following two conditions for  $i = 0, 1, \dots, N - 1$

$$\mathbf{x}_{i+1} - \mathbf{x}_i + \gamma \nabla f(\mathbf{x}_i) = 0, \text{ for some } \gamma \geq 0, \quad (2)$$

$$\nabla f(\mathbf{x}_{i+1})^\top (\mathbf{x}_{i+1} - \mathbf{x}_i) = 0 \quad (3)$$

where the first condition (2) states that we move in the direction of the negative gradient, and the second condition (3) expresses the exact line search condition.

A consequence of those conditions is that successive gradients are orthogonal, i.e.

$$\nabla f(\mathbf{x}_{i+1})^\top \nabla f(\mathbf{x}_i) = 0 \quad i = 0, 1, \dots, N - 1. \quad (4)$$

Instead of relying on conditions (2)–(3) that define the iterates of the gradient method with exact line search, our analysis will be based on the weaker conditions (3)–(4), which are also satisfied by other sequences of iterates.

## 2.2 Interpolation with functions in $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$

We consider the following interpolation problem over the class of functions  $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$ .

**Definition 2.1.** Consider an integer  $N \geq 1$  and given data  $\{(\mathbf{x}_i, f_i, \mathbf{g}_i)\}_{i \in \{0,1,\dots,N\}}$  where  $\mathbf{x}_i \in \mathbb{R}^n$ ,  $f_i \in \mathbb{R}$  and  $\mathbf{g}_i \in \mathbb{R}^n$ . If there exists a function  $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$  such that

$$f(\mathbf{x}_i) = f_i, \quad \nabla f(\mathbf{x}_i) = \mathbf{g}_i, \quad \forall i \in \{0, 1, \dots, N\},$$

then we say that  $\{(\mathbf{x}_i, f_i, \mathbf{g}_i)\}_{i \in \{0,1,\dots,N\}}$  is  $\mathcal{F}_{\mu,L}$ -interpolable.

A necessary and sufficient condition for  $\mathcal{F}_{\mu,L}$ -interpolability is given in the next theorem, taken from [11].

**Theorem 2.2** ([11]). A data set  $\{(\mathbf{x}_i, f_i, \mathbf{g}_i)\}_{i \in \{0,1,\dots,N\}}$  is  $\mathcal{F}_{\mu,L}$ -interpolable if and only if the following inequality

$$f_i - f_j - \mathbf{g}_j^\top (\mathbf{x}_i - \mathbf{x}_j) \geq \frac{1}{2(1 - \mu/L)} \left( \frac{1}{L} \|\mathbf{g}_i - \mathbf{g}_j\|^2 + \mu \|\mathbf{x}_i - \mathbf{x}_j\|^2 - 2 \frac{\mu}{L} (\mathbf{g}_j - \mathbf{g}_i)^\top (\mathbf{x}_j - \mathbf{x}_i) \right)$$

holds for all  $i \neq j \in \{0, 1, \dots, N\}$ .

In principle, Theorem 2.2 allows one to generate all possible valid inequalities that hold for functions in  $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$  in terms of their function values and gradients at a set of points  $\mathbf{x}_0, \dots, \mathbf{x}_N$ . This will be the essential for the proof of our main result, Theorem 1.2.

## 3 A performance estimation problem

The proof technique we will use for Theorem 1.2 is inspired by recent work on the so-called performance estimation problem, as introduced in [2] and further developed in [11]. The idea is to formulate the computation of the worst-case behavior of certain iterative methods as an explicit semidefinite programming (SDP) problem. We first recall the definition of SDP problems (in a form that is suitable to our purposes).

### 3.1 Semidefinite programs

An SDP problem is a nonlinear, convex optimization problem of the form

$$\max_{X=(x_{ij}) \in \mathbb{S}^n, X \succeq 0} \left\{ \sum_{i,j=1}^n c_{ij} x_{ij} \mid \sum_{i,j=1}^n a_{ij}^{(k)} x_{ij} \leq b_k \quad k = 1, \dots, m \right\}, \quad (5)$$

where  $\mathbb{S}^n$  is the set of symmetric matrices of size  $n$ , and matrices  $A_k = (a_{ij}^{(k)}) \in \mathbb{S}^n$  and the matrix  $C = (c_{ij}) \in \mathbb{S}^n$  are given, as well as the scalars  $b_k$  for  $k = 1, \dots, m$ .

Since every positive semidefinite matrix  $X \in \mathbb{S}^n$  is a Gram matrix, there exist vectors  $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$  such that  $x_{ij} = \mathbf{v}_i^\top \mathbf{v}_j$  for all  $i, j$ . Thus the SDP problem (5) may be equivalently rewritten as

$$\max_{\mathbf{v}_i \in \mathbb{R}^n} \left\{ \sum_{i,j=1}^n c_{ij} \mathbf{v}_i^\top \mathbf{v}_j \mid \sum_{i,j=1}^n a_{ij}^{(k)} \mathbf{v}_i^\top \mathbf{v}_j \leq b_k \quad k = 1, \dots, m \right\} \quad (6)$$

which features terms that are linear in the inner products  $\mathbf{v}_i^\top \mathbf{v}_j$  in the objective function and constraints. The associated dual SDP problem is

$$\min_{y \in \mathbb{R}^m, \mathbf{y} \geq \mathbf{0}} \left\{ \mathbf{b}^\top \mathbf{y} \mid \sum_{k=1}^m y_k A_k - C \succeq 0 \right\}. \quad (7)$$

We will later use the fact that each dual variable  $y_k$  may be viewed as a (Lagrange) multiplier of the primal constraint  $\sum_{i,j=1}^n a_{ij}^{(k)} \mathbf{v}_i^\top \mathbf{v}_j \leq b_k$ .

### 3.2 Performance estimation of the gradient method with exact line search

Consider the following SDP problem, for fixed parameters  $N \geq 1$ ,  $R > 0$ ,  $\mu > 0$  and  $L > \mu$ :

$$\text{subject to } \left. \begin{array}{ll} \max f_N - f_* & \\ \mathbf{g}_{i+1}^\top (\mathbf{x}_{i+1} - \mathbf{x}_i) & = 0 \quad i \in \{0, 1, \dots, N-1\} \\ \mathbf{g}_{i+1}^\top \mathbf{g}_i & = 0 \quad i \in \{0, 1, \dots, N-1\} \\ \{(\mathbf{x}_i, f_i, \mathbf{g}_i)\}_{i \in \{*, 0, 1, \dots, N\}} & \text{is } \mathcal{F}_{\mu, L}\text{-interpolable} \\ \mathbf{g}_* & = 0 \\ f_0 - f_* & \leq R, \end{array} \right\} \quad (8)$$

where the variables are  $\mathbf{x}_i \in \mathbb{R}^n$ ,  $f_i \in \mathbb{R}$  and  $\mathbf{g}_i \in \mathbb{R}^n$  ( $i \in \{*, 0, 1, \dots, N\}$ ).

Note that this is indeed an SDP problem of the form (6), with dual problem of the form (7), since equalities and interpolability conditions are linear in the inner products of variables  $\mathbf{x}_i$  and  $\mathbf{g}_i$ .

**Lemma 3.1.** *The optimal value of the above SDP problem (8) is an upper bound on  $f(\mathbf{x}_N) - f_*$ , where  $f$  is any function from  $\mathcal{F}_{\mu, L}(\mathbb{R}^n)$ ,  $f_*$  is its minimum and  $\mathbf{x}_N$  is the  $N$ th iterate of the gradient method with exact line search applied to  $f$  from any starting point  $\mathbf{x}_0$  that satisfies  $f(\mathbf{x}_0) - f_* \leq R$ .*

*Proof.* Fix any  $f \in \mathcal{F}_{\mu, L}(\mathbb{R}^n)$ , and let  $\mathbf{x}_0, \dots, \mathbf{x}_N$  be the iterates of the gradient method with exact line search applied to  $f$ . Now a feasible solution to the SDP problem is given by

$$\mathbf{x}_i, f_i = f(\mathbf{x}_i), \mathbf{g}_i = \nabla f(\mathbf{x}_i) \quad i \in \{*, 0, \dots, N\}.$$

The objective function value at this feasible point is  $f_N = f(\mathbf{x}_N)$ , so that the optimal value of the SDP is an upper bound on  $f(\mathbf{x}_N) - f_*$ .  $\square$

We are now ready to give a proof of our main result. We already mention that the SDP relaxation (8) is not used directly in the proof, but was used to devise the proof, in a sense that will be explained later.

## 4 Proof of Theorem 1.2

A little reflection shows that, to prove Theorem 1.2, we need only consider one iteration of the gradient method with exact line search. Thus we consider only the first iterate, given by  $\mathbf{x}_0$  and  $\mathbf{x}_1$ , as well as the minimizer  $\mathbf{x}_*$  of  $f \in \mathcal{F}_{\mu, L}$ .

Set  $f_i = f(\mathbf{x}_i)$  and  $\mathbf{g}_i = \nabla f(\mathbf{x}_i)$  for  $i \in \{*, 0, 1\}$ . The following five inequalities are now satisfied:

$$\begin{aligned}
1: \quad & f_0 \geq f_1 + \mathbf{g}_1^\top(\mathbf{x}_0 - \mathbf{x}_1) + \frac{1}{2(1 - \mu/L)} \left( \frac{1}{L} \|\mathbf{g}_0 - \mathbf{g}_1\|^2 + \mu \|\mathbf{x}_0 - \mathbf{x}_1\|^2 - 2\frac{\mu}{L} (\mathbf{g}_1 - \mathbf{g}_0)^\top (\mathbf{x}_1 - \mathbf{x}_0) \right) \\
2: \quad & f_* \geq f_0 + \mathbf{g}_0^\top(\mathbf{x}_* - \mathbf{x}_0) + \frac{1}{2(1 - \mu/L)} \left( \frac{1}{L} \|\mathbf{g}_* - \mathbf{g}_0\|^2 + \mu \|\mathbf{x}_* - \mathbf{x}_0\|^2 - 2\frac{\mu}{L} (\mathbf{g}_0 - \mathbf{g}_*)^\top (\mathbf{x}_0 - \mathbf{x}_*) \right) \\
3: \quad & f_* \geq f_1 + \mathbf{g}_1^\top(\mathbf{x}_* - \mathbf{x}_1) + \frac{1}{2(1 - \mu/L)} \left( \frac{1}{L} \|\mathbf{g}_* - \mathbf{g}_1\|^2 + \mu \|\mathbf{x}_* - \mathbf{x}_1\|^2 - 2\frac{\mu}{L} (\mathbf{g}_1 - \mathbf{g}_*)^\top (\mathbf{x}_1 - \mathbf{x}_*) \right) \\
4: \quad & -\mathbf{g}_0^\top \mathbf{g}_1 \geq 0 \\
5: \quad & \mathbf{g}_1^\top(\mathbf{x}_0 - \mathbf{x}_1) \geq 0.
\end{aligned}$$

Indeed, the first three inequalities are the  $\mathcal{F}_{\mu,L}$ -interpolability conditions, the fourth inequality is a relaxation of (4), and the fifth inequality is a relaxation of (3).

We aggregate these five inequalities by defining the following positive multipliers,

$$y_1 = \frac{L - \mu}{L + \mu}, \quad y_2 = 2\mu \frac{(L - \mu)}{(L + \mu)^2}, \quad y_3 = \frac{2\mu}{L + \mu}, \quad y_4 = \frac{2}{L + \mu}, \quad y_5 = 1, \quad (9)$$

and adding the five inequalities together after multiplying each one by the corresponding multiplier.

The result is the following inequality (as may be verified directly):

$$\begin{aligned}
f_1 - f_* \leq & \left( \frac{L - \mu}{L + \mu} \right)^2 (f_0 - f_*) - \frac{\mu L(L + 3\mu)}{2(L + \mu)^2} \left\| \mathbf{x}_0 - \frac{L + \mu}{L + 3\mu} \mathbf{x}_1 - \frac{2\mu}{L + 3\mu} \mathbf{x}_* - \frac{3L + \mu}{L^2 + 3\mu L} \mathbf{g}_0 - \frac{L + \mu}{L^2 + 3\mu L} \mathbf{g}_1 \right\|^2 \\
& - \frac{2L\mu^2}{L^2 + 2L\mu - 3\mu^2} \left\| \mathbf{x}_1 - \mathbf{x}_* - \frac{(L - \mu)^2}{2\mu L(L + \mu)} \mathbf{g}_0 - \frac{L + \mu}{2\mu L} \mathbf{g}_1 \right\|^2.
\end{aligned} \quad (10)$$

Since the last two right-hand-side terms are nonpositive, we obtain:

$$f_1 - f_* \leq \left( \frac{L - \mu}{L + \mu} \right)^2 (f_0 - f_*).$$

Since  $\mathbf{x}_0$  was arbitrary, this completes the proof of Theorem 1.2.  $\square$

#### 4.1 Remarks on the proof of Theorem 1.2.

- Although the proof of Theorem 1.2 is easy to verify, it is not apparent how the multipliers  $y_1, \dots, y_5$  in (9) were obtained. This was in fact done via preliminary computations, and subsequently guessing the values in (9), through the following steps:

1. The SDP performance estimation problem (8) with  $N = 1$  was solved numerically for various values of the parameters  $\mu$ ,  $L$  and  $R$  — actually, the values of  $L$  and  $R$  can safely be fixed to some positive constants using appropriate scaling arguments (see e.g., [11, Section 3.5] for a related argument).
2. The optimal values of the dual SDP multipliers of the constraints corresponding to the five inequalities in the proof gave the guesses for the correct values  $y_1, \dots, y_5$  as stated in in (9).
3. Finally the correctness of the guess was verified directly (by symbolic computation and by hand).

- The key inequality (10) may be rewritten in another, more symmetric way

$$(f_1 - f_*) \leq (f_0 - f_*) \left( \frac{1 - \kappa}{1 + \kappa} \right)^2 - \frac{\mu}{4} \left( \frac{\|\mathbf{s}_1\|^2}{1 + \sqrt{\kappa}} + \frac{\|\mathbf{s}_2\|^2}{1 - \sqrt{\kappa}} \right),$$

where  $\kappa = \mu/L$  is the condition number (between 0 and 1) and slack vectors  $\mathbf{s}_1$  and  $\mathbf{s}_2$  are

$$\begin{aligned}\mathbf{s}_1 &= -\frac{(1 + \sqrt{\kappa})^2}{1 + \kappa} \left( \mathbf{x}_0 - \mathbf{x}_* - \mathbf{g}_0/\sqrt{L\mu} \right) + \left( \mathbf{x}_1 - \mathbf{x}_* + \mathbf{g}_1/\sqrt{L\mu} \right) \\ \mathbf{s}_2 &= \frac{(1 - \sqrt{\kappa})^2}{1 + \kappa} \left( \mathbf{x}_0 - \mathbf{x}_* + \mathbf{g}_0/\sqrt{L\mu} \right) - \left( \mathbf{x}_1 - \mathbf{x}_* - \mathbf{g}_1/\sqrt{L\mu} \right).\end{aligned}$$

Note that the four expressions  $\mathbf{x}_i - \mathbf{x}_* \pm \mathbf{g}_i/\sqrt{L\mu}$  expressions are invariant under dilation of  $f$ , and that cases of equality in (10) simply correspond to equalities  $\mathbf{s}_1 = \mathbf{s}_2 = 0$ .

- It is interesting to note that the known proof of Theorem 1.2 for the quadratic case only requires the so-called Kantorovich inequality, that may be stated as follows.

**Theorem 4.1** (Kantorovich inequality; see e.g. Lemma 3.1 in [1]). *Let  $Q$  be a symmetric positive definite  $n \times n$  matrix with smallest and largest eigenvalues  $\mu > 0$  and  $L \geq \mu$  respectively. Then, for any unit vector  $\mathbf{x} \in \mathbb{R}^n$ , one has:*

$$(\mathbf{x}^\top Q \mathbf{x}) (\mathbf{x}^\top Q^{-1} \mathbf{x}) \leq \frac{(\mu + L)^2}{4\mu L}.$$

Thus, the inequality (10) replaces the Kantorovich inequality in the proof of Theorem 1.2 for non-quadratic  $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ .

- Finally, note that we have proven a bit more than what is stated in Theorem 1.2. Indeed, the result in Theorem 1.2 holds for any iterative method that satisfies the five inequalities used in its proof.

## 5 Extension to ‘noisy’ gradient descent with exact line search

Theorem 1.2 may be generalized to what we will call *noisy gradient descent method with exact linear search*; see e.g. [1, p.59] where it is called gradient descent method with (relative) error. Here the search direction at iteration  $i$ , say  $\mathbf{d}_i$ , satisfies

$$\| -\nabla f(\mathbf{x}_i) - \mathbf{d}_i \| \leq \varepsilon \|\nabla f(\mathbf{x}_i)\| \quad i = 0, 1, \dots, \quad (11)$$

where  $0 \leq \varepsilon < 1$  is some given *relative* tolerance on the deviation from the negative gradient. Note that the algorithm cannot be guaranteed to converge as soon as  $\varepsilon \geq 1$ , since  $\mathbf{d}_i = 0$  then becomes feasible.. We recover the normal gradient descent algorithm when  $\varepsilon = 0$ . In the case of more general values of  $\varepsilon$ , one can for example satisfy the relative error criterion when imposing a restriction of the type  $|\sin \theta| \leq \varepsilon$  on the angle  $\theta$  between search direction  $\mathbf{d}_i$  and the current gradient  $\nabla f(x_i)$ .

Using a search direction  $\mathbf{d}_i$  that satisfies (11) corresponds, for example, to an implementation of the gradient descent method where each component of  $-\nabla f(\mathbf{x}_i)$  is only calculated to a fixed number of significant digits. It is also related to the so-called *stochastic gradient descent* method that is used in training neural networks; see e.g. [4] and the references therein.

Thus we consider the following algorithm:

**Noisy gradient descent method with exact line search**

**Input:**  $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ ,  $\mathbf{x}_0 \in \mathbb{R}^n$ ,  $0 \leq \varepsilon < 1$ .

**for**  $i = 0, 1, \dots$

Compute any search direction  $\mathbf{d}_i$  that satisfies (11);

$\gamma = \operatorname{argmin}_{\gamma \in \mathbb{R}} f(\mathbf{x}_i - \gamma \mathbf{d}_i)$

$\mathbf{x}_{i+1} = \mathbf{x}_i - \gamma \mathbf{d}_i$

One may show the following generalization of Theorem 1.2.

**Theorem 5.1.** *Let  $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ ,  $\mathbf{x}_*$  a global minimizer of  $f$  on  $\mathbb{R}^n$ , and  $f_* = f(\mathbf{x}_*)$ . Given a relative tolerance  $\varepsilon$ , each iteration of the noisy gradient descent method with exact line search satisfies*

$$f(\mathbf{x}_{i+1}) - f_* \leq \left( \frac{1 - \kappa_\varepsilon}{1 + \kappa_\varepsilon} \right)^2 (f(\mathbf{x}_i) - f_*) \quad i = 0, 1, \dots \quad (12)$$

where  $\kappa_\varepsilon = \frac{\mu(1-\varepsilon)}{L(1+\varepsilon)}$ .

If  $\varepsilon = 0$  this is exactly Theorem 1.2, and the proof of Theorem 5.1 is a straightforward generalization of the proof of Theorem 1.2. The key is again to consider a wider class of iterative methods that satisfies certain inequalities. Here we use the inequalities:

$$\left. \begin{array}{l} 1: \quad f_0 \geq f_1 + \mathbf{g}_1^\top(\mathbf{x}_0 - \mathbf{x}_1) + \frac{1}{2(1-\mu/L)} \left( \frac{1}{L} \|\mathbf{g}_0 - \mathbf{g}_1\|^2 + \mu \|\mathbf{x}_0 - \mathbf{x}_1\|^2 - 2\frac{\mu}{L} (\mathbf{g}_1 - \mathbf{g}_0)^\top (\mathbf{x}_1 - \mathbf{x}_0) \right) \\ 2: \quad f_* \geq f_0 + \mathbf{g}_0^\top(\mathbf{x}_* - \mathbf{x}_0) + \frac{1}{2(1-\mu/L)} \left( \frac{1}{L} \|\mathbf{g}_* - \mathbf{g}_0\|^2 + \mu \|\mathbf{x}_* - \mathbf{x}_0\|^2 - 2\frac{\mu}{L} (\mathbf{g}_0 - \mathbf{g}_*)^\top (\mathbf{x}_0 - \mathbf{x}_*) \right) \\ 3: \quad f_* \geq f_1 + \mathbf{g}_1^\top(\mathbf{x}_* - \mathbf{x}_1) + \frac{1}{2(1-\mu/L)} \left( \frac{1}{L} \|\mathbf{g}_* - \mathbf{g}_1\|^2 + \mu \|\mathbf{x}_* - \mathbf{x}_1\|^2 - 2\frac{\mu}{L} (\mathbf{g}_1 - \mathbf{g}_*)^\top (\mathbf{x}_1 - \mathbf{x}_*) \right) \\ 4: \quad 0 \geq \mathbf{g}_1^\top(\mathbf{x}_1 - \mathbf{x}_0) \\ 5: \quad 0 \geq \mathbf{g}_0^\top \mathbf{g}_1 - \varepsilon \|\mathbf{g}_0\| \|\mathbf{g}_1\|. \end{array} \right\} \quad (13)$$

The first four inequalities are the same as before, and the fifth is satisfied by the iterates of the noisy gradient descent with exact line search. Indeed, in the first iteration one has:

$$\begin{aligned} 0 &= \mathbf{d}_0^\top \frac{\mathbf{g}_1}{\|\mathbf{g}_1\|} && \text{(exact line search)} \\ &= (\mathbf{d}_0 + \mathbf{g}_0)^\top \frac{\mathbf{g}_1}{\|\mathbf{g}_1\|} - \frac{\mathbf{g}_0^\top \mathbf{g}_1}{\|\mathbf{g}_1\|} \\ &\leq \varepsilon \|\mathbf{g}_0\| - \frac{\mathbf{g}_0^\top \mathbf{g}_1}{\|\mathbf{g}_1\|} && \text{(by Cauchy-Schwartz and (11)).} \end{aligned}$$

We rewrite the fifth inequality as the equivalent linear matrix inequality:

$$\begin{pmatrix} \varepsilon \|\mathbf{g}_0\|^2 & \mathbf{g}_0^\top \mathbf{g}_1 \\ \mathbf{g}_0^\top \mathbf{g}_1 & \varepsilon \|\mathbf{g}_1\|^2 \end{pmatrix} \succeq 0. \quad (14)$$

We first aggregate the first four inequalities in (13) by adding them together after multiplication by the respective multipliers:

$$y_1 = \rho_\varepsilon, \quad y_2 = 2\kappa_\varepsilon \frac{1 - \kappa_\varepsilon}{(1 + \kappa_\varepsilon)^2}, \quad y_3 = \frac{2\kappa_\varepsilon}{1 + \kappa_\varepsilon}, \quad y_4 = 1,$$

where  $L_\varepsilon = (1 + \varepsilon)L$ ,  $\mu_\varepsilon = (1 - \varepsilon)\mu$ ,  $\kappa_\varepsilon = \frac{\mu_\varepsilon}{L_\varepsilon}$  and  $\rho_\varepsilon = \frac{1 - \kappa_\varepsilon}{1 + \kappa_\varepsilon}$ .

Next we define a (matrix) multiplier for the linear matrix inequality (14), namely

$$\begin{pmatrix} a\rho_\varepsilon & -a \\ -a & \frac{a}{\rho_\varepsilon} \end{pmatrix} \succeq 0, \quad (15)$$

with  $a = \frac{1}{L_\varepsilon + \mu_\varepsilon}$ , and add the inner product of the left-hand-side of (14) with the multiplier matrix (15) to the aggregated constraints. It can now be checked that the resulting expression is the following (slight) generalization of (10)

$$\begin{aligned} f_1 - f_* &\leq \rho_\varepsilon^2 (f_0 - f_*) - \frac{L\mu(L_\varepsilon - \mu_\varepsilon)(L_\varepsilon + 3\mu_\varepsilon)}{2(L - \mu)(L_\varepsilon + \mu_\varepsilon)^2} \|\mathbf{x}_0 + \alpha_1 \mathbf{x}_1 - (1 + \alpha_1) \mathbf{x}_* + \alpha_2 \mathbf{g}_0 + \alpha_3 \mathbf{g}_1\|^2 \\ &\quad - \frac{2L\mu\mu_\varepsilon}{(L - \mu)(L_\varepsilon + 3\mu_\varepsilon)} \|\mathbf{x}_1 - \mathbf{x}_* + \alpha_4 \mathbf{g}_0 + \alpha_5 \mathbf{g}_1\|^2, \end{aligned}$$

with the appropriate coefficients  $\alpha_1 = -\frac{L_\varepsilon + \mu_\varepsilon}{L_\varepsilon + 3\mu_\varepsilon}$ ,  $\alpha_2 = -\frac{4L - L_\varepsilon + \mu_\varepsilon}{L(L_\varepsilon + 3\mu_\varepsilon)}$ ,  $\alpha_3 = \frac{(L_\varepsilon + \mu_\varepsilon)(-4L + 3L_\varepsilon + \mu_\varepsilon)}{L(L_\varepsilon - \mu_\varepsilon)(L_\varepsilon + 3\mu_\varepsilon)}$ ,  $\alpha_4 = -\frac{(L - \mu)(L_\varepsilon - \mu_\varepsilon)}{2L\mu(L_\varepsilon + \mu_\varepsilon)}$  and  $\alpha_5 = -\frac{L + \mu}{2L\mu}$ . This completes the proof.

## 6 Concluding remarks

The main results of this paper are the exact convergence rates of the gradient descent method with exact line search (and its noisy variant) for strongly convex functions with Lipschitz continuous gradients. The computer-assisted technique of proof is also of independent interest, and demonstrates the importance of the SDP performance estimation problems (PEPs) introduced in [2].

Indeed, to obtain our proof of Theorem 5.1, the following SDP PEP was solved numerically for various fixed values of  $R$ ,  $\mu$  and  $L$ :

$$\max f_1 - f_* \quad \text{subject to (13) and } f_0 - f_* \leq R.$$

It was observed that, for each set of values, the optimal value of the SDP corresponded exactly to the bound in Theorem 5.1 (actually, for homogeneity reasons,  $L$  and  $R$  could be fixed and only  $\mu$  needed to vary). Based on this, a rigorous proof Theorem 5.1 could be given by guessing the correct values of the dual SDP multipliers as functions of  $\mu$ ,  $L$  and  $R$ , and then verifying the guess through an explicit computation.

We believe this type of computer-assisted proof could prove useful in the analysis of more methods where exact line search is used (see for example [10] which studies conditional gradient methods).

PEPs have been used by now to study worst-case convergence rates of several first-order optimization methods [2, 10, 11]. This paper differs in an important aspect: the performance estimation problem considered actually characterizes a whole class of methods that contains the method of interest (gradient descent with exact line search) as well as many other methods. This relaxation in principle only provides an upper bound on the worst-case of gradient descent, and it is the fact that Example 1.3 matches this bound that allows us to conclude with a tight result.

The reason we could not solve the performance estimation problem for the gradient descent method itself is that equation (2), which essentially states that the step  $\mathbf{x}_{i+1} - \mathbf{x}_i$  is parallel to the gradient  $\nabla f(\mathbf{x}_i)$ , cannot be formulated as an SDP constraint. The main obstruction appears to be that requiring that two vectors are parallel is a nonconvex constraint, even when workings with their inner products<sup>2</sup>. Instead, our convex formulation enforces that those two vectors are both orthogonal to a third one, the next gradient  $\nabla f(\mathbf{x}_{i+1})$ .

## References

- [1] D.P. Bertsekas. *Nonlinear programming*. Athena scientific, 1999.
- [2] 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.
- [3] D.G. Luenberger and Y. Ye. *Linear and nonlinear programming*. Springer, 2008.
- [4] A. Neelakantan, L. Vilnis, Q.V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens. Adding gradient noise improves learning for very deep networks. *arXiv*, 1511.06807v1, 2015.
- [5] A. Nemirovski. Optimization II: Numerical methods for nonlinear continuous optimization. Lecture notes, 1999. Available from: [http://www2.isye.gatech.edu/~nemirovs/Lect\\_OptII.pdf](http://www2.isye.gatech.edu/~nemirovs/Lect_OptII.pdf).
- [6] A. Nemirovski and D.B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.
- [7] Yu. Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., 2004.
- [8] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

<sup>2</sup>One such nonconvex formulation would be  $\mathbf{g}_i^\top(\mathbf{x}_i - \mathbf{x}_{i+1}) = \|\mathbf{g}_i\| \|\mathbf{x}_i - \mathbf{x}_{i+1}\|$ .

- [9] B.T. Polyak. *Introduction to optimization*. Optimization Software, New York, 1987.
- [10] A.B. Taylor, J.M. Hendrickx, and F. Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *arXiv*, 1512.07516, 2015.
- [11] A.B. Taylor, J.M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 2016. Accepted for publication.