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

De Klerk, Etienne; Glineur, François; Taylor, Adrien B.

DOI
10.1007/s11590-016-1087-4

Publication date
2016

Document Version
Final published version

Published in
Optimization Letters

Citation (APA)
De Klerk, E., Glineur, F., & Taylor, A. B. (2016). On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, *11*(7), 1185–1199. https://doi.org/10.1007/s11590-016-1087-4

Optim Lett (2017) 11:1185–1199
DOI 10.1007/s11590-016-1087-4

ORIGINAL PAPER

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

Etienne de Klerk[1,2] · **François Glineur**[3] · **Adrien B. Taylor**[3]

Received: 29 June 2016 / Accepted: 6 October 2016 / Published online: 14 October 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 give the tight worst-case complexity bound for 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 proofs are computer-assisted, and rely on the resolutions of semidefinite programming performance estimation problems as introduced in the paper (Drori and Teboulle, Math Progr 145(1–2):451–482, 2014).

Keywords Gradient method · Steepest descent · Semidefinite programming · Performance estimation problem

A.B. Taylor is a F.R.I.A. fellow (F.R.S.-FNRS). The UCL authors are supported by the Belgian Interuniversity Attraction Poles, and by the ARC Grant 13/18-054 (Communauté française de Belgique).

✉ Etienne de Klerk
E.deKlerk@uvt.nl

François Glineur
Francois.Glineur@uclouvain.be

Adrien B. Taylor
Adrien.Taylor@uclouvain.be

1 Tilburg University, Tilburg, The Netherlands

2 Delft University of Technology, Delft, The Netherlands

3 UCL/CORE and ICTEAM, Louvain-la-Neuve, Belgium

1 Introduction

The gradient (or steepest) descent method for unconstrained method was devised by Augustin–Louis Cauchy (1789–1857) in the nineteenth 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 $f : \mathbb{R}^n \to \mathbb{R}$ is called L-smooth, μ-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, μ-strongly convex functions on $\mathbb{R}^n$ 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

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

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 matrix $\nabla^2 f(\mathbf{x})$ 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 $i = 0, 1, \ldots$

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

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

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*

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

Note that the result in Theorem 1.2, which establishes a global linear convergence rate on objective function accuracy, is known for the case of quadratic functions in $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$, that is for functions of the form

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

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, §1.3], [9, pp. 60–62], or [3, pp. 235–238]. Moreover, the bound (1) 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

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

and the starting point

$$\mathbf{x}_0 = \left(\frac{1}{\mu}, 0, \ldots, 0, \frac{1}{L}\right)^\mathsf{T}.$$

One may readily check that the gradient at $\mathbf{x}_0$ is equal to

$$\nabla f(\mathbf{x}_0) = (1, 0, \ldots, 0, 1)^\mathsf{T}$$

and that the minimum of the line-search from $\mathbf{x}_0$ in that direction is attained for step $\gamma = \frac{2}{L+\mu}$. One therefore obtains

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

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

$$\mathbf{x}_{2i} = \left(\frac{L-\mu}{L+\mu}\right)^{2i} x_0, \quad \mathbf{x}_{2i+1} = \left(\frac{L-\mu}{L+\mu}\right)^{2i} x_1.$$

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

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

holds as required. □

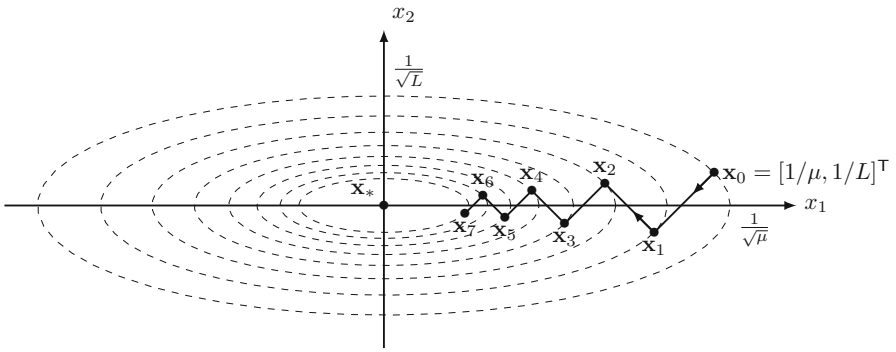

Fig. 1 Illustration of Example 1.3 for the case $n = 2$ (*small arrows* indicate direction of negative gradient)

The construction in Example 1.3 is illustrated in Fig. 1 in the case $n = 2$, where the ellipses shown are level curves of the objective function. Each step from $\mathbf{x}_i$ to $\mathbf{x}_{i+1}$ is orthogonal to the ellipse at $\mathbf{x}_i$ (since it uses the steepest descent direction) and tangent to the ellipse at $\mathbf{x}_{i+1}$ (because of the exact line-search direction), hence successive steps are orthogonal to each other.

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 ϵ-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 ϵ 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 (1) are known. For example, in [3, p. 240], the following bound is shown:

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

In [8, Theorem 3.4] a stronger result than Theorem 1.2 was claimed, but this was retracted in a subsequent erratum,[1] 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

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

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]. We mention one more particular result by Nesterov [7] that is similar to our main result in Theorem 1.2, but that uses a fixed step-length and relies on the initial distance to the solution.

Theorem 1.5 (Theorem 2.1.15 in [7]) *Given $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ and $\mathbf{x}_0 \in \mathbb{R}^n$, the gradient descent method with fixed step length $\gamma = \frac{2}{\mu+L}$ generate iterates $\mathbf{x}_i$ $(i = 0, 1, 2, \ldots)$ that satisfy*

$$f(\mathbf{x}_i) - f_* \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2i} \|\mathbf{x}_0 - \mathbf{x}_*\|^2 \quad i = 0, 1, \ldots$$

Note that this result does not imply Theorem 1.2.

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.

2.1 Properties of the gradient method with exact line search

Let $\mathbf{x}_i$ $(i = 1, 2, \ldots, N)$ 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 $i = 0, 1, \ldots, N-1$

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

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

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.

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

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 now 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 $\{(\mathbf{x}_i, f_i, \mathbf{g}_i)\}_{i\in\{0,1,\ldots,N\}}$ 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

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

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

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

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

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

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

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 $\mathbf{x}_0, \ldots, \mathbf{x}_N$. This will be 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

We will consider semidefinite programs (SDPs) of the form

$$\max_{X=(x_{ij})\in\mathbb{S}^n, X\succeq 0, \mathbf{u}\in\mathbb{R}^\ell} \left\{ \sum_{i,j=1}^n c_{ij}x_{ij} + \mathbf{c}^\mathsf{T}\mathbf{u} \;\middle|\; \sum_{i,j=1}^n a_{ij}^{(k)}x_{ij} + \mathbf{a}_k^\mathsf{T}\mathbf{u} \leq b_k \quad k = 1, \ldots, m \right\}, \tag{5}$$

where $\mathbb{S}^n$ is the set of symmetric matrices of size n, and matrices $A_k = \left(a_{ij}^{(k)}\right) \in \mathbb{S}^n$ and the matrix $C = (c_{ij}) \in \mathbb{S}^n$ are given, as well as the scalars b_k and vectors $\mathbf{a}_k \in \mathbb{R}^\ell$ $(k = 1, \ldots, m)$, and $\mathbf{c} \in \mathbb{R}^\ell$.

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

$$\max_{\mathbf{v}_i \in \mathbb{R}^n, \mathbf{u} \in \mathbb{R}^\ell} \left\{ \sum_{i,j=1}^n c_{ij} \mathbf{v}_i^\mathsf{T} \mathbf{v}_j + \mathbf{c}^\mathsf{T} \mathbf{u} \;\middle|\; \sum_{i,j=1}^n a_{ij}^{(k)} \mathbf{v}_i^\mathsf{T} \mathbf{v}_j + \mathbf{a}_k^\mathsf{T} \mathbf{u} \le b_k \quad k = 1, \ldots, m \right\} \tag{6}$$

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

$$\min_{\mathbf{y} \in \mathbb{R}^m, \mathbf{y} \ge \mathbf{0}} \left\{ \mathbf{b}^\mathsf{T} \mathbf{y} \;\middle|\; \sum_{k=1}^m y_k A_k - C \succeq 0, \; \sum_{k=1}^m y_k \mathbf{a}_k = \mathbf{c} \right\}. \tag{7}$$

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

3.2 Performance estimation of the gradient method with exact line search

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

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

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

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 Nth iterate of the gradient method with exact line search applied to f from any starting point $\mathbf{x}_0$ that satisfies $f(\mathbf{x}_0) - f_* \le R$.*

Proof Fix any $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$, and let $\mathbf{x}_0, \ldots, \mathbf{x}_N$ 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

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

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 $f(\mathbf{x}_N) - f_*$. □

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 $i \in \{*, 0, 1\}$. Note that $\mathbf{g}_* = \mathbf{0}$. The following five inequalities are now satisfied:

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

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,

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

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):

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

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. □

4.1 Remarks on the proof of Theorem 1.2

- First, 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.
- Although the proof of Theorem 1.2 is easy to verify, it is not apparent how the multipliers $y_1, \ldots, y_5$ 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 μ , 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 discussion).
  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 $y_1, \ldots, y_5$ as stated 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

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

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

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:*

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

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, we note that this proof can be modified very easily to handle the case of the fixed-step gradient method that was mentioned in Theorem 1.5. Indeed, observe that the proof aggregates the fourth and fifth inequalities with multipliers $y_4 = \frac{2}{L+\mu}$ and $y_5 = 1$, which leads to the combined inequality

$$\frac{2}{L+\mu}(-\mathbf{g}_0^\mathsf{T}\mathbf{g}_1) + \mathbf{g}_1^\mathsf{T}(\mathbf{x}_0 - \mathbf{x}_1) \geq 0 \quad \Leftrightarrow \quad \mathbf{g}_1^\mathsf{T}\left(\mathbf{x}_0 - \frac{2}{L+\mu}\mathbf{g}_0 - \mathbf{x}_1\right) \geq 0.$$

Now note that the gradient method with fixed step $\gamma = \frac{2}{L+\mu}$ satisfies this combined inequality (since the second factor in the left-hand side becomes zero), and hence the rest of the proof establishes the same rate for this method as for the gradient descent with exact line search.

Theorem 4.2 *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 fixed step length $\gamma = \frac{2}{\mu+L}$ satisfies*

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

Note that Example 1.3 also establishes that this rate is tight. Hence we have the relatively surprising fact that, when looking at the worst-case convergence rate of the objective function accuracy, performing exact line-search is not better than using a well-chosen fixed step length.

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

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

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 ε, one can for example satisfy the relative error criterion by imposing a restriction of the type $|\sin\theta| \leq \varepsilon$ on the angle θ between search direction $\mathbf{d}_i$ and the current negative gradient $-\nabla f(\mathbf{x}_i)$.

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 $i = 0, 1, \ldots$

Select any seach direction $\mathbf{d}_i$ that satisfies (11);

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

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

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 ε, each iteration of the noisy gradient descent method with exact line search satisfies*

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

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

When $\varepsilon = 0$, the rate becomes $\frac{1-\kappa}{1+\kappa} = \frac{L-\mu}{L+\mu}$, which matches 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:

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

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:

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

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

$$\begin{pmatrix} \varepsilon\|\mathbf{g}_0\|^2 & \mathbf{g}_0^\mathsf{T}\mathbf{g}_1 \\ \mathbf{g}_0^\mathsf{T}\mathbf{g}_1 & \varepsilon\|\mathbf{g}_1\|^2 \end{pmatrix} \succeq 0. \tag{14}$$

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 positive semidefinite matrix multiplier for the linear matrix inequality (14), namely

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

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

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

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. □

To conclude this section, the following example, based on the same quadratic function as Example1.3, shows that our bound (12) for the noisy gradient descent is also tight.

Example 5.2 Consider the same quadratic function as in Example 1.3:

$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i x_i^2 \quad \text{where} \quad 0 < \mu = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = L.$$

Let θ be an angle satisfying $0 \leq \theta < \frac{\pi}{2}$. Consider the noisy gradient descent method where direction $\mathbf{d}_0$ is obtained by performing a counterclockwise 2D-rotation with angle θ on the first and last coordinates of the gradient $\nabla f(\mathbf{x}_0)$. As mentioned above, this satisfies our definition with relative tolerance $\varepsilon = \sin\theta$. Define now the starting point

$$\mathbf{x}_0 = \left(\frac{1}{\mu}, 0, \ldots, 0, \frac{1}{L}\sqrt{\frac{1-\varepsilon}{1+\varepsilon}}\right)^{\mathsf{T}}.$$

Tedious but straightforward computations show that

$$\mathbf{x}_1 = \left(\frac{1-\kappa_\varepsilon}{1+\kappa_\varepsilon}\right)\left(\frac{1}{\mu}, 0, \ldots, 0, -\frac{1}{L}\sqrt{\frac{1-\varepsilon}{1+\varepsilon}}\right)^{\mathsf{T}} \quad \text{where } \kappa_\varepsilon = \frac{\mu}{L}\frac{(1-\varepsilon)}{(1+\varepsilon)}.$$

Moreover, if one chooses $\mathbf{d}_1$ by rotating the second gradient $\nabla f(\mathbf{x}_1)$ by the same angle θ in the clockwise direction, one obtains

$$\mathbf{x}_2 = \left(\frac{1-\kappa_\varepsilon}{1+\kappa_\varepsilon}\right)^2\left(\frac{1}{\mu}, 0, \ldots, 0, \frac{1}{L}\sqrt{\frac{1-\varepsilon}{1+\varepsilon}}\right)^{\mathsf{T}} = \left(\frac{1-\kappa_\varepsilon}{1+\kappa_\varepsilon}\right)^2 \mathbf{x}_0.$$

A similar reasoning for the next iterates, alternating counterclockwise and clockwise rotations, shows that

$$\mathbf{x}_{2i} = \left(\frac{1-\kappa_\varepsilon}{1+\kappa_\varepsilon}\right)^{2i} x_0, \quad \mathbf{x}_{2i+1} = \left(\frac{1-\kappa_\varepsilon}{1+\kappa_\varepsilon}\right)^{2i} x_1 \ \text{ for all } i = 0, 1, \ldots$$

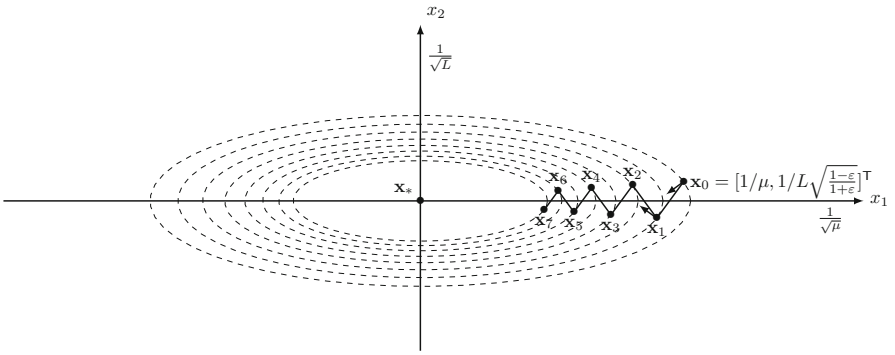

Fig. 2 Illustration Example 5.2 for $n = 2$ and $\varepsilon = 0.3$ (*small arrows* indicate direction of negative gradient)

and hence we have that equality

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

holds as announced. Figure 2 displays a few iterates, and can be compared to Fig. 1. □

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, μ and L:

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

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 μ 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 μ, 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 Eq. (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 a convex constraint in the SDP formulation. The main obstruction appears to be that requiring that two vectors are parallel is a nonconvex constraint, even when working with their inner products.[2] Instead, our convex formulation enforces that those two vectors are both orthogonal to a third one, the next gradient $\nabla f(\mathbf{x}_{i+1})$.

Acknowledgements The authors would like to thank Simon Lacoste–Julien for bringing Theorem 2.1.15 in [7] to their attention, and an anonymous referee for valuable suggestions that include the last remark in Sect. 4.1.

References

1. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Massachusetts (1999)
2. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: a novel approach. Math. Progr. **145**(1–2), 451–482 (2014)
3. Luenberger, D.G., Ye, Y.: Linear and Nonlinear Programming. Springer, Berlin (2008)
4. Neelakantan, A., Vilnis, L., Le, Q.V., Sutskever, I., Kaiser, L., Kurach, K., Martens, J.: Adding gradient noise improves learning for very deep networks (2015). arXiv:1511.06807v1
5. Nemirovski, A.: Optimization II: numerical methods for nonlinear continuous optimization. In: Lecture Notes (1999). http://www2.isye.gatech.edu/~nemirovs/Lect_OptII.pdf
6. Nemirovski, A., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)
7. Nesterov, Yu.: Introductory lectures on convex optimization: a basic course. In: Applied Optimization. Kluwer Academic Publ., Boston (2004)
8. Nocedal, J., Wright, S.: Numerical Optimization. Springer Science & Business Media, Berlin (2006)
9. Polyak, B.T.: Introduction to Optimization. Optimization Software, New York (1987)
10. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Exact worst-case performance of first-order methods for composite convex optimization (2015). arXiv:1512.07516
11. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Math. Progr. (2016). doi:10.1007/s10107-016-1009-3

[2] One such nonconvex formulation would be $\mathbf{g}_i^\mathsf{T}(\mathbf{x}_i - \mathbf{x}_{i+1}) = \|\mathbf{g}_i\|\|\mathbf{x}_i - \mathbf{x}_{i+1}\|$.