The exact worst-case convergence rate of the gradient method with fixed step lengths for L-smooth functions

Abbaszadehpeivasti, Hadi; de Klerk, Etienne; Zamani, Moslem

The exact worst-case convergence rate of the gradient method with fixed step lengths for L-smooth functions

Hadi Abbaszadehpeivasti · Etienne de Klerk · Moslem Zamani[*]

Abstract In this paper, we study the convergence rate of the gradient (or steepest descent) method with fixed step lengths for finding a stationary point of an L-smooth function. We establish a new convergence rate, and show that the bound may be exact in some cases, in particular when all step lengths lie in the interval $(0, 1/L]$. In addition, we derive an optimal step length with respect to the new bound.

Keywords L-smooth optimization · Gradient method · Performance estimation problem · Semidefinite programming

1 Introduction

We consider the non-convex unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1}$$

[*] Corresponding author.

H. Abbaszadehpeivasti
Tilburg University, Department of Econometrics and Operations Research, Tilburg, The Netherlands
E-mail: h.abbaszadehpeivasti@tilburguniversity.edu

E. De Klerk
Tilburg University, Department of Econometrics and Operations Research, Tilburg, The Netherlands
E-mail: e.deklerk@tilburguniversity.edu

M. Zamani
Tilburg University, Department of Econometrics and Operations Research, Tilburg, The Netherlands
E-mail: m.zamani_1@tilburguniversity.edu

where $f : \mathbb{R}^n \to \mathbb{R}$ is bounded from below, and let a real number $f^\star$ denote a lower bound of problem (1). In addition, we assume throughout the paper that f has an L-Lipschitz gradient, that is

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \quad \forall x, y \in \mathbb{R}^n, \tag{2}$$

for some (known) Lipschitz constant $L > 0$. Following the notation used by Nesterov [10], we let $C_L^{1,1}(\mathbb{R}^n)$ denote functions with L-Lipschitz gradient.

Problem (1) arises naturally in many applications including machine learning, signal and image processing, to name but a few [1,9]. One of the historic solution methods for problem (1) is the gradient method, proposed by Cauchy in 1847 [4].

The gradient method with fixed step lengths may be described as follows.

Algorithm 1 Gradient method with fixed step lengths

Set N and $\{t_k\}_{k=1}^N$ (step lengths) and pick $x^1 \in \mathbb{R}^n$.
For $k = 1, 2, \ldots, N$ perform the following step:
  1. $x^{k+1} = x^k - t_k \nabla f(x^k)$

Nesterov [10, page 28] gives the following convergence rate (to a stationary point) for Algorithm 1 when $t_k \in (0, \frac{2}{L})$, $k \in \{1, \ldots, N\}$:

$$\min_{1 \le k \le N+1} \|\nabla f(x^k)\| \le \left(\frac{f(x^1) - f^\star}{\sum_{k=1}^N t_k (1 - \frac{1}{2} L t_k)} \right)^{\frac{1}{2}}.$$

In the special case $t_k = \frac{1}{L}$, $k \in \{1, \ldots, N\}$, the last bound becomes

$$\min_{1 \le k \le N+1} \|\nabla f(x^k)\| \le \left(\frac{2L(f(x^1) - f^\star)}{N} \right)^{\frac{1}{2}}.$$

Recently, semidefinite programming performance estimation have been employed as a tool for the worst-case analysis of first-order methods [5,6,8,12,14]. In this method, the worst-case convergence is cast as a quadratic program with quadratic constraints and the problem is then solved by semi-definite programming methods. By employing the performance estimation method, Taylor [11, page 190], without giving a proof, states the following convergence rate:

$$\min_{1 \le k \le N+1} \|\nabla f(x^k)\| \le \left(\frac{4L(f(x^1) - f^\star)}{3N} \right)^{\frac{1}{2}}, \tag{3}$$

for $t_k = \frac{1}{L}$, $k \in \{1, \ldots, N\}$. Drori [7, Corollary 1 in Appendix] considers the case that all step lengths are smaller than $\frac{1}{L}$, and proves the following convergence rate

$$\min_{1 \le k \le N+1} \|\nabla f(x^k)\| \le \left(\frac{4(f(x^1) - f^\star)}{\sum_{k=1}^N t_k (4 - L t_k)} \right)^{\frac{1}{2}}. \tag{4}$$

It can be observed that when the step lengths are the same for each iteration and tend to $\frac{1}{L}$, the bound (4) reduces to Taylor's convergence rate.

In this paper, we investigate the convergence rate of Algorithm 1 further. By using the performance estimation method, we provide a converge rate, which is tighter than all aforementioned bounds. For example, as a part of our main result in Theorem 2, we improve on (4) by showing, for any choice of $t_k \in (0, 2/L)$ ($k \in \{1, \ldots, N\}$), that

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(x^k) \right\| \leq \left(\frac{4\Delta}{\sum_{k=1}^N \min(-L^2 t_k^3 + 4t_k, -Lt_k^2 + 4t_k) + \frac{2}{L}} \right)^{\frac{1}{2}}. \quad (5)$$

As a consequence, we also prove and improve on (3) by showing, in the special case where all $t_k = 1/L$ ($k \in \{1, \ldots, N\}$), that

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(x^k) \right\| \leq \left(\frac{4L(f(x^1) - f^\star)}{3N+2} \right)^{\frac{1}{2}}.$$

In addition, we construct an L-smooth function that attains the given bound in Theorem 2 for certain step lengths. We also propose an optimal step length that minimizes the right-hand-side of the bound (5), namely $t_k = \frac{\sqrt{4/3}}{L}$ for all $k \in \{1, \ldots, N\}$.

Outline

The paper is organized as follows. We describe the performance estimation technique in Section 2. In Section 3, we study the convergence rate by using performance estimation. Finally, we conclude the paper with a conjecture.

Notation

The n-dimensional Euclidean space is denoted by $\mathbb{R}^n$. We use $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ to denote the Euclidean inner product and norm, respectively. For a matrix A, A_{ij} denotes its (i, j)-th entry, and A^T represents the transpose of A. The notation $A \succeq 0$ means the matrix A is symmetric positive semi-definite.

2 Performance estimation

Computation of the worst-case convergence rate for a given iterative method and a given class of functions is an infinite-dimensional optimization problem. In their seminal paper [8], Drori and Teboulle take advantage of this idea, called performance estimation, and introduce some relaxation method to deal with this infinite-dimensional optimization problem. Performance estimation has been used extensively for the analysis of first-order methods [5,6,8,12,14].

Similar to problem (P) in [8], the worst-case convergence rate of Algorithm 1 may be formulated as the following abstract optimization problem,

$$\max \left(\min_{1 \le k \le N+1} \left\| \nabla f(x^k) \right\| \right)$$

$$\text{s.t. } f(x^1) - f^\star \le \Delta$$

$$\quad x^{N+1}, x^N, \ldots, x^2 \text{ are generated by Algorithm 1 w.r.t. } f, x^1 \quad (6)$$

$$\quad f(x) \ge f^\star \ \forall x \in \mathbb{R}^n$$

$$\quad f \in C_L^{1,1}(\mathbb{R}^n)$$

$$\quad x^1 \in \mathbb{R}^n,$$

where $\Delta \ge 0$ denote the difference between the given lower bound, $f^\star$, and the value of f at the starting point. In problem (6), f and x^1 are decision variables. This is an infinite-dimensional optimization problem with infinite number of constraints, and consequently intractable in general. In what follows, we provide a semidefinite programming relaxation for the problem.

Definition 1 [12, Definition 3.8.] Let $L \ge 0$. A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called L-smooth, if it satisfies the following condition,

$$\left| f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right| \le \tfrac{L}{2} \left\| y - x \right\|^2 \quad \forall x, y \in \mathbb{R}^n. \tag{7}$$

The following proposition states a well-known characterization of L-smooth functions that follows, e.g., from [10, Lemma 1.2.3], [10, Theorem 2.1.5] and [12, Lemma 3.9].

Proposition 1 *Let $L \ge 0$. $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth, if and only if it has an L-Lipschitz gradient.*

The following well-known result is a fundamental property of gradient descent for L-smooth functions, if the step length $1/L$ is used.

Proposition 2 *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth, and $x \in \mathbb{R}^n$, then*

$$f \left(x - \frac{1}{L} \nabla f(x) \right) \le f(x) - \frac{1}{2L} \| \nabla f(x) \|^2.$$

The following theorem plays a key role in our analysis. Indeed, it provides necessary and sufficient conditions for the interpolation of L-smooth functions. Using this theorem, we will formulate problem (6) as a finite dimensional optimization problem.

Theorem 1 ([12, Lemma 3.9.], [7, Theorem 7 in Appendix]) *Let $\{(x^i; g^i; f^i)\}_{i \in I} \subseteq \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ with a given index set I and $L > 0$. There exists an L-smooth function f with*

$$f(x^i) = f^i, \nabla f(x^i) = g^i \quad i \in I, \tag{8}$$

if and only if

$$\frac{1}{2L}\left\|g^i - g^j\right\|^2 - \frac{L}{4}\left\|x^i - x^j - \frac{1}{L}(g^i - g^j)\right\|^2 \le f^i - f^j - \langle g^j, x^i - x^j \rangle \quad i,j \in I. \quad (9)$$

In addition, if the triple $\{(x^i; g^i; f^i)\}_{i \in I}$ satisfies (9), then there exists L-smooth function f for which (8) holds and $\min_{x \in \mathbb{R}^n} f(x) = \min_{i \in I} f_i - \frac{1}{2L}\|g^i\|^2$. Moreover, letting $i^* \in \arg\min_{i \in I} f_i - \frac{1}{2L}\|g^i\|^2$, a global minimizer of this function is given by $x^\star = x_{i^*} - \frac{1}{L}g^{i^*}$.

Another proof of the first part of Theorem 1 may be found in [15, Theorem 2, Page 148]. By virtue of Theorem 1, problem (6) may be reformulated as follows,

$$\max \left(\min_{1 \le k \le N+1} \left\|g^k\right\| \right)$$
$$\text{s.t. } \frac{1}{2L}\left\|g^i - g^j\right\|^2 - \frac{L}{4}\left\|x^i - x^j - \frac{1}{L}(g^i - g^j)\right\|^2 \le f^i - f^j -$$
$$\langle g^j, x^i - x^j \rangle \quad i,j \in \{1, \dots, N+1\}$$
$$x^{k+1} = x^k - t_k g^k \quad k \in \{1, \dots, N\} \quad (10)$$
$$f^k \ge f^\star \quad k \in \{1, \dots, N+1\}$$
$$f^1 - f^\star \le \Delta.$$

In the above formulation, x^k, g^k, f^k, $k \in \{1, \dots, N+1\}$, are decision variables. Note that in the above formulation, the constraints $f(x) \ge f^\star$ for each $x \in \mathbb{R}^n$ are replaced by $f^k \ge f^\star$, $k \in \{1, \dots, N+1\}$. These constraints do not necessarily impose a given L-Lipschitz function f with

$$f(x^i) = f^i, \nabla f(x^i) = g^i \quad i \in \{1, \dots, N+1\},$$

which has the lower bound $f^\star$. Therefore, the optimal value of (6) and (10) may not be equal in general. However, if an optimal solution of problem (10) satisfies $f^\star = \min_{1 \le k \le N+1} f^k - \frac{1}{2L}\|g^k\|^2$, the formulation will be exact; see the second part of Theorem 1. By Proposition 2, we have $f(x) - \frac{1}{2L}\|\nabla f(x)\|^2 \ge f^\star$ for $x \in \mathbb{R}^n$. Hence, we replace the constraint $f^k \ge f^\star$ by $f^k - \frac{1}{2L}\|g^k\|^2 \ge f^\star$ and consider the following problem:

$$\max \left(\min_{1 \le k \le N+1} \left\|g^k\right\| \right)$$
$$\text{s.t. } \frac{1}{2L}\left\|g^i - g^j\right\|^2 - \frac{L}{4}\left\|x^i - x^j - \frac{1}{L}(g^i - g^j)\right\|^2 \le f^i - f^j -$$
$$\langle g^j, x^i - x^j \rangle \quad i,j \in \{1, \dots, N, N+1\}$$
$$x^{k+1} = x^k - t_k g^k \quad k \in \{1, \dots, N+1\} \quad (11)$$
$$f^k - \frac{1}{2L}\|g^k\|^2 - f^\star \ge 0 \quad k \in \{1, \dots, N+1\}$$
$$f^1 - f^\star \le \Delta.$$

From the constraint $x^{k+1} = x^k - t_k g^k$, we get $x^i = x^1 + \sum_{k=1}^{i-1} g^k$, $i \in \{2, \ldots, N\}$. By using this relation to eliminate the x^i ($i \in \{2, \ldots, N+2\}$), problem (11) may be written as follows:

$$\max \ell$$

$$\text{s.t. } f^i - f^j - \frac{1}{2L} \left\| g^i - g^j \right\|^2 + \frac{L}{4} \left\| -\sum_{k=j}^{i-1} t_k g^k + \frac{1}{L}(g^i - g^j) \right\|^2 + \left\langle g^j, \sum_{k=j}^{i-1} t_k g^k \right\rangle \geq 0 \quad i > j$$

$$f^i - f^j - \frac{1}{2L} \left\| g^i - g^j \right\|^2 + \frac{L}{4} \left\| \sum_{k=i}^{j-1} t_k g^k - \frac{1}{L}(g^i - g^j) \right\|^2 - \left\langle g^j, \sum_{k=i}^{j-1} t_k g^k \right\rangle \geq 0 \quad i < j$$

$$f^k - \frac{1}{2L} \|g^k\|^2 - f^\star \geq 0 \quad k \in \{1, \ldots, N+1\} \tag{12}$$

$$f^\star - f^1 + \Delta \geq 0$$

$$\left\| g^k \right\|^2 - \ell \geq 0 \quad k \in \{1, \ldots, N+1\},$$

where ℓ is an auxiliary variable to convert problem (11) into a quadratic program. Problem (12) is a non-convex quadratic program with quadratic constraints. In the following proposition, we show that the optimal values of problems (6) and (11) (or equivalently problem (12)) are the same for step lengths in the interval $(0, \frac{1}{2L})$.

Proposition 3 *If $t_k \in (0, \frac{2}{L})$, $k \in \{1, \ldots, N\}$, then problems (6) and (11) (or equivalently problem (12)) share the same optimal value.*

Proof. Clearly, problem (11) is a relaxation of problem (6). Therefore, we only need to show that, for any feasible solution of (11), say $\{(\bar{x}^i; \bar{g}^i; \bar{f}^i)\}_1^{N+1}$, there exists an L-smooth function f with

$$f(\bar{x}^i) = \bar{f}^i, \ \nabla f(\bar{x}^i) = \bar{g}^i, \quad 1 \leq i \leq N+1,$$

and $\min_{x \in \mathbb{R}^n} f(x) \geq f^\star$. The existence such of a function follows from Theorem 1, as all assumptions of Theorem 1 are satisfied. $\qquad\square$

To obtain a tractable form of problem (12), we relax it to a semidefinite program, in the spirit of [8]. To this end, we define the $(N+1) \times (N+1)$ positive semi-definite matrix G as,

$$G = \begin{pmatrix} (g^1)^T \\ \vdots \\ (g^{N+1})^T \end{pmatrix} \begin{pmatrix} g^1 \ldots g^{N+1} \end{pmatrix} = \begin{pmatrix} \|g^1\|^2 & \cdots & \langle g^1, g^{N+1} \rangle \\ \vdots & \ddots & \vdots \\ \langle g^1, g^{N+1} \rangle & \cdots & \|g^{N+1}\|^2 \end{pmatrix}.$$

We may now formulate the following semidefinite program,

$$\max \ell$$
$$\text{s.t. } f^i - f^j + \text{tr}(A^{ij}G) \geq 0 \quad i \neq j \in \{1, \dots, N+1\}$$
$$f^k - \frac{1}{2L}G_{kk} - f^\star \geq 0 \quad k \in \{1, \dots, N+1\}$$
$$f^\star - f^1 + \Delta \geq 0 \tag{13}$$
$$G_{kk} - \ell \geq 0 \quad k \in \{1, \dots, N+1\}$$
$$G \succeq 0,$$

where the $(N+1) \times (N+1)$ matrices A^{ij}, $i \neq j \in \{1, \dots, N+1\}$, are formed according to the constraints (12), and G, ℓ, f^i, $i \in \{1, \dots, N+1\}$, are decision variables. Problem (13) is a relaxation of (12), but if $n \geq N+1$ the relaxation is exact, that is the optimal values of (12) and (13) are the same. Indeed, if $n \geq N+1$, and G is a feasible matrix in (13), then G is the Gram matrix of $N+1$ vectors in $\mathbb{R}^n$, and these vectors may be identified with $g^1, \dots, g^{N+1}$; a similar argument is used in [14, Theorem 5].

3 Worst-case convergence rate

In this section, we investigate the convergence rate of gradient method with fixed step lengths. The next theorem gives the worst-case convergence rate of Algorithm 1 to a stationary point of an L-smooth function. The technique of the proof, as is usual for SDP performance estimation, is to use weak duality. In particular, we will in fact construct a feasible solution to the dual SDP problem of (13), and thus derive an upper bound for problem (12).

In practice, this dual feasible solution is constructed in a computer-assisted manner, by solving the primal and dual SDP problems for different fixed values of the parameters, and subsequently guessing the values of the dual multipliers. (There is also dedicated software for this purpose, namely 'PESTO' by Taylor, Glineur, and Hendrickx [13].) In the proof of Theorem 2, we simply verify that these 'guesses' are correct.

Theorem 2 *Let $t_k \in (0, \frac{\sqrt{3}}{L})$ for $k \in \{1, \dots, N\}$. Consider N iterations of Algorithm 1 with step lengths t_k ($k \in \{1, \dots, N\}$), applied to some L-smooth function f with minimum value $f^\star$, with the starting point x^1 satisfying $f(x^1) - f^\star \leq \Delta$, for some given $\Delta > 0$.*

Then, if $x^1, \dots, x^{N+1}$ denote the iterates of Algorithm 1, one has

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(x^k) \right\| \leq \left(\frac{4\Delta}{\sum_{k=1}^{N} \min(-L^2 t_k^3 + 4t_k, -Lt_k^2 + 4t_k) + \frac{2}{L}} \right)^{\frac{1}{2}}. \tag{14}$$

In particular, if $t_k = \frac{\sqrt{4/3}}{L}$ for $k \in \{1, \dots, N\}$, we get

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(x^k) \right\| \leq \left(\frac{6\sqrt{3}L(f(x^1) - f^\star)}{8N + 3\sqrt{3}} \right)^{\frac{1}{2}}. \tag{15}$$

Similarly, if $t_k = \frac{1}{L}$ for $k \in \{1, \ldots, N\}$, one has

$$\min_{1 \le k \le N+1} \left\| \nabla f(x^k) \right\| \le \left(\frac{4L(f(x^1) - f^\star)}{3N+2} \right)^{\frac{1}{2}}. \tag{16}$$

Proof. Let U denote the square of the right-side of inequality (14) and let $B = \frac{U}{\Delta}$. To establish this bound, we show that U is an upper bound for problem (12). Consider the feasible point $\left(\{g^k; f^k\}_1^{N+1}; \ell \right)$ for problem (12). Suppose that

$$\alpha_k = \frac{B}{2} \max \{2, t_k L + 1\} \quad k \in \{1, \ldots, N\}.$$

In addition, we define σ_1 and σ_k, respectively, as follows:

$$\sigma_1 = \frac{B}{4} \min \left\{ -Lt_1^2 + 3t_1, -L^2 t_1^3 + 3t_1 \right\},$$
$$\sigma_k = \frac{B}{4} \min \left\{ -Lt_k^2 + 3t_k + t_{k-1}, -L^2 t_k^3 + 3t_k + t_{k-1} \right\} \quad k \in \{2, \ldots, N\},$$

and $\sigma_{N+1} = 1 - \sum_{k=1}^N \sigma_k = \frac{B}{4L}(2 + Lt_N)$. As $t_k \in (0, \frac{\sqrt{3}}{L})$ for $k \in \{1, \ldots, N\}$, the σ_k's will be non-negative. It is seen that

$$\sigma_k + (2\alpha_k - B) \frac{Lt_k^2}{4} - \frac{Bt_k}{2} = \frac{B}{4}(t_k + t_{k-1}) \quad k \in \{2, \ldots, N\}.$$

By using the last equality, one may verify directly through elementary algebra that

$$\ell - U + \sum_{k=1}^{N+1} \sigma_k \left(\left\| g^k \right\|^2 - \ell \right) + B \left(f^\star - f^1 + \Delta \right) + B \left(f^{N+1} - \frac{1}{2L} \| g^{N+1} \|^2 - f^\star \right)$$

$$+ \sum_{k=1}^N \alpha_k \left(f^k - f^{k+1} - \frac{1}{2L} \left\| g^k - g^{k+1} \right\|^2 + \frac{L}{4} \left\| t_k g^k - \frac{1}{L} \left(g^k - g^{k+1} \right) \right\|^2 \right.$$

$$- \left\langle g^{k+1}, t_k g^k \right\rangle \right) + \sum_{k=1}^N (\alpha_k - B) \left(f^{k+1} - f^k - \frac{1}{2L} \left\| g^{k+1} - g^k \right\|^2 \right.$$

$$+ \frac{L}{4} \left\| -t_k g^k - \frac{1}{L} \left(g^{k+1} - g^k \right) \right\|^2 - \left\langle g^k, -t_k g^k \right\rangle \right) = \frac{-(2\alpha_1 - B)}{4L} \left\| g^1 - g^2 \right\|^2$$

$$+ \frac{Bt_1}{4} \left\| g^1 \right\|^2 - \frac{Bt_1}{2} \left\langle g^1, g^2 \right\rangle + \frac{Bt_N}{4} \left\| g^{N+1} \right\|^2$$

$$+ \sum_{k=2}^N \left(\frac{-(2\alpha_k - B)}{4L} \left\| g^k - g^{k+1} \right\|^2 + \frac{B(t_k + t_{k-1})}{4} \left\| g^k \right\|^2 - \frac{Bt_k}{2} \left\langle g^k, g^{k+1} \right\rangle \right) =$$

$$- \sum_{k=1}^N Q_k,$$

where

$$Q_k = \begin{cases} \frac{B}{4} \left(\frac{1}{L} - t_k \right) \left\| g^k - g^{k+1} \right\|^2 & t_k < \frac{1}{L} \\ 0 & t_k \ge \frac{1}{L}. \end{cases}$$

Since $\sum_{k=1}^N Q_k$ is a non-negative quadratic function and the given dual multipliers are non-negative, we have $\ell \le U$ for any feasible solution of (12). $\quad\square$

The special step length $t_k = \frac{\sqrt{4/3}}{L}$ for $k \in \{1, \ldots, N\}$ used to obtain (15) will be motivated later in Theorem 3. Note that (16) gives a formal proof (with a small improvement) of the bound claimed by Taylor [11, page 190]; see (3).

An important question concerning the bound (14) is its difference with the optimal value of (6). It is known that the lower bound for Algorithm 1 is of the order $\Omega\left(\frac{1}{\sqrt{N}}\right)$ [2,3]. In what follows, we establish that the bound (14) is exact in some cases.

Proposition 4 *The value*

$$\left(\frac{4\Delta}{\sum_{k=1}^{N} \min\left(-L^2 t_k^3 + 4t_k, -Lt_k^2 + 4t_k\right) + \frac{2}{L}}\right)^{\frac{1}{2}}$$

is the optimal value of (6) *when all step lengths satisfy* $t_k \in (0, \frac{1}{L}]$, $k \in \{1, \ldots, N\}$.

Proof. It suffices for a given N to demonstrate an L-smooth function f and a point x^1 such that

$$\min_{1 \le k \le N+1} \left\|\nabla f(x^k)\right\| = \left(\frac{4\Delta}{\sum_{k=1}^{N} \min\left(-L^2 t_k^3 + 4t_k, -Lt_k^2 + 4t_k\right) + \frac{2}{L}}\right)^{\frac{1}{2}}. \quad (17)$$

Suppose now that $t_k \in (0, \frac{1}{L}]$, $k \in \{1, \ldots, N\}$, and U denotes the right-hand-side of equality (17). As before, we set $t_{N+1} = \frac{1}{L}$. Let

$$l_i = U\left(\sum_{k=i}^{N+1} t_k\right), \quad f^i = \Delta - \frac{U^2}{4}\left(\sum_{k=1}^{i-1} -Lt_k^2 + 4t_k\right) \quad i \in \{1, \ldots, N+1\},$$

and $l_{N+2} = 0$. By elementary calculus, one can check that the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} \frac{L}{2}(x - l_1)^2 + U(x - l_1) + f^1 & x \in \left[\frac{1}{2}(l_1 + l_2), \infty\right) \\ \frac{-L}{2}(x - l_{i+1})^2 + U(x - l_{i+1}) + f^{i+1} & x \in \left[l_{i+1}, \frac{1}{2}(l_i + l_{i+1})\right] \\ \frac{L}{2}(x - l_{i+1})^2 + U(x - l_{i+1}) + f^{i+1} & x \in \left[\frac{1}{2}(l_{i+1} + l_{i+2}), l_{i+1}\right] \\ \frac{L}{2}x^2 & x \in \left(-\infty, \frac{1}{2}l_{N+1}\right] \end{cases} \quad (18)$$

for $i \in \{1, \ldots, N\}$, is L-smooth with the optimal value $f^\star = 0$ and the optimal solution $x^\star = 0$. In addition, we have equality (17) for $x^1 = l_1$. Indeed,

$$\begin{aligned} x^i &= l_i & i \in \{1, \ldots, N+1\} \\ \nabla f(x^i) &= U & i \in \{1, \ldots, N+1\} \\ f(x^i) &= f^i & i \in \{1, \ldots, N+1\}. \end{aligned}$$

$\square$

(a) $N = 4, \ \Delta = 2, \ L = 1$
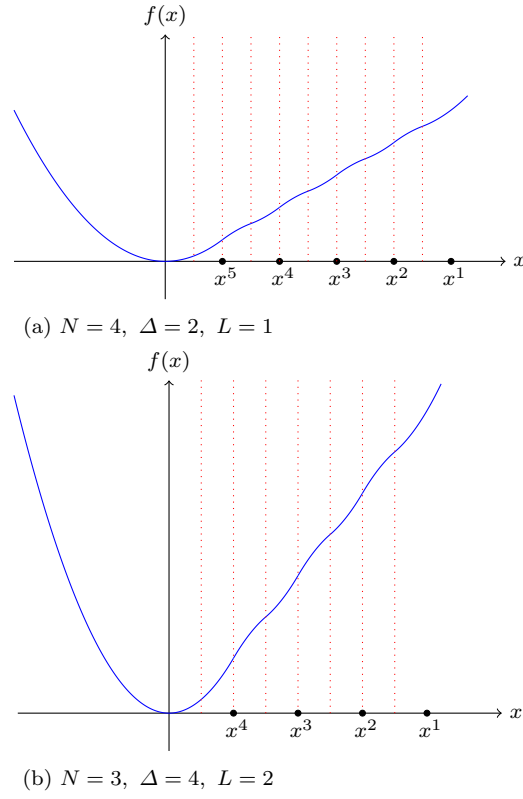

(b) $N = 3, \ \Delta = 4, \ L = 2$

Fig. 1: Plot of the function f in (18) for different parameters and $t_k = \frac{1}{L}$. (Dotted lines denote the endpoints of intervals.)

Figure 1 represents the plot of function f as constructed in the proof of Proposition 4 for different parameters and the fixed step length $t_k = \frac{1}{L}$ for all k.

Note that, though we have only shown the exactness of the bound (14) for step lengths in the interval $(0, \frac{1}{L}]$, we also conjecture that the bound (14) is in fact exact for all step lengths in the interval $(0, \frac{\sqrt{3}}{L})$.

By minimizing the right-hand-side of (14), the next theorem gives the 'optimal' step lengths with respect to the bound.

Theorem 3 *Let f be an L-smooth function. Then the optimal step size for gradient method with respect to bound* (14) *is given by*

$$t_k = \frac{\sqrt{\frac{4}{3}}}{L} \quad \forall k \in \{1, \dots, N\},$$

provided that $t_k \in (0, \frac{\sqrt{3}}{L})$ for all $k \in \{1, \dots, N\}$.

Proof. We minimize the right-hand-side of (14), that is

$$\min_{t_k \in \left(0, \frac{\sqrt{3}}{L}\right)} \left(\frac{4\Delta}{\sum_{k=1}^{N} \min\left(-L^2 t_k^3 + 4t_k, -Lt_k^2 + 4t_k\right) + \frac{2}{L}} \right)^{\frac{1}{2}},$$

which is equivalent to maximizing

$$\max_{t \in \left(0, \frac{\sqrt{3}}{L}\right)^N} H(t) := \sum_{k=1}^{N} \min\left(-L^2 t_k^3 + 4t_k, -Lt_k^2 + 4t_k\right).$$

Since H is a strictly concave function on $\left(0, \frac{\sqrt{3}}{L}\right)^N$ and at $\bar{t}$ given by

$$\bar{t}_k = \frac{\sqrt{\frac{4}{3}}}{L} \quad \forall k \in \{1, \dots, N\},$$

we have $\nabla H(\bar{t}) = 0$, which shows that $\bar{t}$ is the unique maximum solution of H over $\left(0, \frac{\sqrt{3}}{L}\right)^N$. $\qquad\square$

The step length $\frac{1}{L}$ commonly is regarded as the optimal step length in the literature; see [10, Chapter 1]. Due to the example introduced in (18), we see that the worst-case convergence rate for the step length $\frac{1}{L}$ cannot be better than $\left(\frac{4L(f(x^1) - f^\star)}{3N+2} \right)^{\frac{1}{2}}$. By our analysis, it follows that, for the step length $\frac{2\sqrt{3}}{3L}$, we get the convergence rate (15), which is better than $\left(\frac{4L(f(x^1) - f^\star)}{3N+2} \right)^{\frac{1}{2}}$, since the constant in the bound improves from ca. $\frac{4}{3} \approx 1.333$ to $\frac{6\sqrt{3}}{8} \approx 1.299$.

4 Concluding remarks

In this paper, we studied the convergence rate of gradient method for L-smooth functions and we provided a new convergence rate when the step lengths belong to the interval $(0, \frac{\sqrt{3}}{L})$. Moreover, we have shown that this convergence rate is tight for step lengths in the interval $(0, \frac{1}{L}]$. As mentioned in the introduction, Algorithm 1 is convergent for the step lengths in the larger interval $(0, \frac{2}{L})$. Following extensive numerical experiments, where we solved the semidefinite program (13) for different parameter values, we conjecture that when $t_k \in \left(0, \frac{2}{L}\right)$ for $k \in \{1, \dots, N\}$, we have

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(x^k) \right\| \leq \left(\frac{4\Delta}{\sum_{k=1}^{N} \min\left(-L^2 t_k^3 + 4t_k, -Lt_k^2 + 4t_k\right)} \right)^{\frac{1}{2}},$$

under the same conditions as for Theorem 2. The right-hand-side is again minimized by the constant step length $t_k = \frac{\sqrt{4/3}}{L}$ for all $k \in \{1, \dots, N\}$.

References

1. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Review **60**(2), 223–311 (2018)
2. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points I. Mathematical Programming **184**, 1–50 (2019)
3. Cartis, C., Gould, N.I., Toint, P.L.: On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. SIAM Journal on Optimization **20**(6), 2833–2852 (2010)
4. Cauchy, A.: Méthode générale pour la résolution des systemes déquations simultanées. Comp. Rend. Sci. Paris **25**(1847), 536–538 (1847)
5. De Klerk, E., Glineur, F., Taylor, A.B.: 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)
6. De Klerk, E., Glineur, F., Taylor, A.B.: Worst-case convergence analysis of inexact gradient and Newton methods through semidefinite programming performance estimation. SIAM Journal on Optimization **30**(3), 2053–2082 (2020)
7. Drori, Y., Shamir, O.: The complexity of finding stationary points with stochastic gradient descent. In: Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 2658–2667. Proceedings of Machine Learning Research (2020)
8. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: a novel approach. Mathematical Programming **145**(1), 451–482 (2014)
9. Jain, P., Kar, P.: Non-convex optimization for machine learning. Foundations and Trends in Machine Learning **10**(3-4), 142–363 (2017)
10. Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer Science & Business Media (2003)
11. Taylor, A.B.: Convex interpolation and performance estimation of first-order methods for convex optimization. Ph.D. thesis, Catholic University of Louvain, Louvain-la-Neuve, Belgium (2017)
12. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Exact worst-case performance of first-order methods for composite convex optimization. SIAM Journal on Optimization **27**(3), 1283–1313 (2017)
13. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Performance estimation toolbox (PESTO): Automated worst-case analysis of first-order optimization methods. In: 2017 IEEE 56th Annual Conference on Decision and Control (CDC), pp. 1278–1283 (2017). DOI 10.1109/CDC.2017.8263832
14. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Mathematical Programming **161**(1-2), 307–345 (2017)
15. Wells, J.C.: Differentiable functions on Banach spaces with Lipschitz derivatives. Journal of Differential Geometry **8**(1), 135–152 (1973)