Optimization Letters (2023) 17:1105–1125
https://doi.org/10.1007/s11590-023-01981-2

ORIGINAL PAPER

Conditions for linear convergence of the gradient method for non-convex optimization

Hadi Abbaszadehpeivasti[1] · Etienne de Klerk[1] · Moslem Zamani[1]

Received: 31 March 2022 / Accepted: 20 January 2023 / Published online: 25 February 2023

Abstract
In this paper, we derive a new linear convergence rate for the gradient method with fixed step lengths for non-convex smooth optimization problems satisfying the Polyak-Łojasiewicz (PŁ) inequality. We establish that the PŁ inequality is a necessary and sufficient condition for linear convergence to the optimal value for this class of problems. We list some related classes of functions for which the gradient method may enjoy linear convergence rate. Moreover, we investigate their relationship with the PŁ inequality.

Keywords Weakly convex optimization · Gradient method · Performance estimation problem · Polyak-Łojasiewicz inequality · Semidefinite programming

1 Introduction

We consider the gradient method for the unconstrained optimization problem

$$f^\star := \inf_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, and $f^\star$ is finite. The gradient method with fixed step lengths may be described as follows.

✉ Moslem Zamani
m.zamani_1@tilburguniversity.edu

Hadi Abbaszadehpeivasti
h.abbaszadehpeivasti@tilburguniversity.edu

Etienne de Klerk
e.deklerk@tilburguniversity.edu

[1] Department of Econometrics and Operations Research, Tilburg University, Tilburg, The Netherlands

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. $x^{k+1} = x^k - t_k \nabla f(x^k)$

In addition, we assume that f has a maximum curvature $L \in (0, \infty)$ and a minimum curvature $\mu \in (-\infty, L)$. Recall that f has a maximum curvature L if $\frac{L}{2}\|.\|^2 - f$ is convex. Similarly, f has a minimum curvature μ if $f - \frac{\mu}{2}\|.\|^2$ is convex. We denote smooth functions with curvature belonging to the interval $[\mu, L]$ by $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$. The class $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$ includes all smooth functions with Lipschitz gradient (note that $\mu \geq 0$ corresponds to convexity). Indeed, f is L-smooth on $\mathbb{R}^n$ if and only if f has a maximum and minimum curvature $\bar{L} > 0$ and $\bar{\mu}$, respectively, with $\max(\bar{L}, |\bar{\mu}|) \leq L$. This class of functions is broad and appears naturally in many models in machine learning, see [8] and the references therein.

For $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$, we have the following inequalities for $x, y \in \mathbb{R}^n$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \tag{2}$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2; \tag{3}$$

see Lemma 2.5 in [21].

It is known that the lower bound of first order methods for obtaining an ϵ-stationary point is of the order $\Omega\left(\epsilon^{-2}\right)$ for L-smooth functions [6]. Hence, it is of interest to investigate the classes of functions for which the gradient method enjoys linear convergence rate. This subject has been investigated by some scholars and some classes of functions have been introduced where linear convergence is possible; see [7, 14–16] and the references therein. This includes the class of functions satisfying the Polyak-Łojasiewicz (PŁ) inequality [16, 20].

Definition 1 A function f is said to satisfy the PŁ inequality on $X \subseteq \mathbb{R}^n$ if there exists $\mu_p > 0$ such that

$$f(x) - f^\star \leq \frac{1}{2\mu_p}\|\nabla f(x)\|^2, \quad \forall x \in X. \tag{4}$$

Note that the PŁ inequality is also known as *gradient dominated*; see [19, Definition 4.1.3]. Strongly convex functions satisfy the PŁ inequality, but some classes of non-convex functions also fulfill this inequality. For instance, consider a differentiable function $G : \mathbb{R}^n \to \mathbb{R}^m$ with $m \leq n$. Suppose that the non-linear system $G(x) = 0$ has some solution. If

$$\min_{x \in \mathbb{R}^n} \lambda_{\min}\left(J_G(x) J_G(x)^T\right) = \sigma > 0,$$

where $J_G(x)$ is the Jacobian matrix of G at x, then the function $f(x) = \|G(x)\|^2$ fulfils the PŁ inequality; see [19, Example 4.1.3]. Here, $\lambda_{\min}(A)$ denotes the smallest

eigenvalue of symmetric matrix A. In other words, nonlinear least squares problems often correspond to instances of (1) where the objective satisfies the PŁ inequality.

The following classical theorem provides a linear convergence rate for Algorithm 1 under the PŁ inequality.

Theorem 1 [20, *Theorem* 4] *Let f be L-smooth and let f satisfy PŁ inequality on $X = \{x : f(x) \leq f(x^1)\}$. If $t_1 \in (0, \frac{2}{L})$ and x^2 is generated by Algorithm* 1, *then*

$$f(x^2) - f^\star \leq \big(1 - t_1\mu_p(2 - t_1L)\big)\big(f(x^1) - f^\star\big). \tag{5}$$

In particular, if $t_1 = \frac{1}{L}$, we have

$$f(x^2) - f^\star \leq \Big(1 - \frac{\mu_p}{L}\Big)\big(f(x^1) - f^\star\big). \tag{6}$$

In this paper we will sharpen this bound; see Theorem 3. Under the assumptions of Theorem 1, Karimi et al. [16] established linear convergence rates for some other methods including the randomized coordinate descent. We refer the interested reader to the recent survey [7] for more details on the convergence of non-convex algorithms under the PŁ inequality.

In this paper, we analyse the gradient method from black-box perspective, which means that we have access to the gradient and the function value at the given point. Furthermore, we study the convergence rate of Algorithm 1 by using performance estimation.

In recent years, performance estimation has been used to find worst-case convergence rates of first order methods [1, 2, 9, 10, 13, 23], to name but a few. This strong tool first has been introduced by Drori and Teboulle in their seminal paper [12]. The idea of performance estimation is that the infinite dimensional optimization problem concerning the computation of convergence rate may be formulated as a finite dimensional optimization problem (often semidefinite programs) by using interpolation theorems. The interested reader may consult the PhD theses of Drori [11] and Taylor [22] for an introduction to, and review of the topic.

The rest of the paper is organized as follows. In Sect. 2, we consider problem (1) when f satisfies the PŁ inequality. We derive a new linear convergence rate for Algorithm 1 by using performance estimation. Furthermore, we provide an optimal step length with respect to the given bound. We also show that the PŁ inequality is necessary and sufficient for linear convergence, in a well-defined sense. Sect. 3 lists some other situations where Algorithm 1 is linearly convergent. Moreover, we study the relationships between these situations. Finally, we conclude the paper with some remarks and questions for future research.

Notation

The n-dimensional Euclidean space is denoted by $\mathbb{R}^n$. Vectors are considered to be column vectors and the superscript T denotes the transpose operation. 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. The notation $A \succeq 0$ means the matrix A is symmetric positive semi-definite, and $\mathrm{tr}(A)$ stands for the trace of A.

2 Linear convergence under the PŁ inequality

This section studies linear convergence of the gradient descent under the PŁ inequality. It is readily seen that the PŁ inequality implies that every stationary point is a global minimum on X. By virtue of the descent lemma [19, Page 29], we have

$$f(x) - f^\star \geq \tfrac{1}{2L}\|\nabla f(x)\|^2, \ \forall x \in \mathbb{R}^n.$$

Hence, μ_p can take value in $(0, L]$. On the other hand, we may assume without loss of generality $\mu \leq \mu_p$. The inequality is trivial if $\mu \leq 0$, and we therefore assume that $\mu > 0$. By taking the minimum with respect to y from both side of inequality (3), we get

$$f(x) - f^\star \leq \tfrac{1}{2\mu}\|\nabla f(x)\|^2.$$

Hence, one may assume without loss of generality $\mu_p = \max\{\mu, \mu_p\}$ in inequality (4).

In what follows, we employ performance estimation to get a new bound under the assumptions of Theorem 1. In this setting, the worst-case convergence rate of Algorithm 1 may be cast as the following optimization problem,

$$\begin{aligned} \max \ & \tfrac{f(x^2)-f^\star}{f(x^1)-f^\star} \\ & x^2 \text{ is generated by Algorithm 1 w.r.t. } f, x^1 \\ & f(x) \geq f^\star \ \forall x \in \mathbb{R}^n \\ & f(x) - f^\star \leq \tfrac{1}{2\mu_p}\|\nabla f(x)\|^2, \quad \forall x \in X \\ & f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n) \\ & x^1 \in \mathbb{R}^n. \end{aligned} \tag{7}$$

In problem (7), f and x^1 are decision variables and $X = \{x : f(x) \leq f(x^1)\}$. We may replace the infinite dimensional condition $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ by a finite set of constraints, by using interpolation. Theorem 2 gives some necessary and sufficient conditions for the interpolation of given data by some $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$.

Theorem 2 [21, *Theorem* 3.1] *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 let* $L \in (0, \infty]$ *and* $\mu \in (-\infty, L)$. *There exists a function* $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ *with*

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

if and only if for every $i, j \in I$

$$\begin{aligned} & \tfrac{1}{2\left(1-\frac{\mu}{L}\right)}\left(\tfrac{1}{L}\left\|g^i - g^j\right\|^2 + \mu\left\|x^i - x^j\right\|^2 - \tfrac{2\mu}{L}\left\langle g^j - g^i, x^j - x^i\right\rangle\right) \\ & \quad \leq f^i - f^j - \left\langle g^j, x^i - x^j\right\rangle. \end{aligned} \tag{9}$$

It is worth noting that Theorem 2 addresses non-smooth functions as well. In fact, $L = \infty$ covers non-smooth functions. Note that we only investigate the smooth case in this paper, that is $L \in (0, \infty)$ and $\mu \in (-\infty, 0]$.

By Theorem 2, problem (7) may be relaxed as follows,

$$\begin{array}{ll} \max & \frac{f^2-f^\star}{f^1-f^\star} \\ \text{s.t.} & \frac{1}{2(1-\frac{\mu}{L})}\left(\frac{1}{L}\left\|g^i-g^j\right\|^2+\mu\left\|x^i-x^j\right\|^2-\frac{2\mu}{L}\left\langle g^j-g^i,x^j-x^i\right\rangle\right)\leq \\ & f^i-f^j-\left\langle g^j,x^i-x^j\right\rangle \quad i,j\in\{1,2\} \\ & x^2=x^1-t_1g^1 \\ & f^k\geq f^\star \quad k\in\{1,2\} \\ & f^k-f^\star\leq\frac{1}{2\mu_p}\|g^k\|^2, \quad k\in\{1,2\}. \end{array} \tag{10}$$

As we replace the constraint $f(x)-f^\star \leq \frac{1}{2\mu_p}\|\nabla f(x)\|^2$ for each $x \in X$ by $f^1-f^\star \leq \frac{1}{2\mu_p}\|g^1\|^2$ and $f^2-f^\star \leq \frac{1}{2\mu_p}\|g^2\|^2$, problem (10) is a relaxation of problem (7). By using the constraint $x^2 = x^1 - t_1 g^1$, problem (10) may be reformulated as,

$$\begin{array}{ll} \max & \frac{f^2-f^\star}{f^1-f^\star} \\ \text{s.t.} & \frac{1}{2(L-\mu)}\left(\|g^2\|^2+(1+\mu Lt_1^2-2\mu t_1)\|g^1\|^2+2(\mu t_1-1)\langle g^1,g^2\rangle\right) \\ & -f^2+f^1-\left\langle g^1,t_1g^1\right\rangle\leq 0 \\ & \frac{1}{2(L-\mu)}\left(\|g^2\|^2+(1+\mu Lt_1^2-2\mu t_1)\|g^1\|^2+2(\mu t_1-1)\langle g^1,g^2\rangle\right) \\ & -f^1+f^2+\left\langle g^2,t_1g^1\right\rangle\leq 0 \\ & f^\star-f^k\leq 0 \quad k\in\{1,2\} \\ & f^k-f^\star-\frac{1}{2\mu_p}\|g^k\|^2\leq 0, \quad k\in\{1,2\}. \end{array} \tag{11}$$

By using the Gram matrix,

$$X=\begin{pmatrix}(g^1)^T\\(g^2)^T\end{pmatrix}\begin{pmatrix}g^1 & g^2\end{pmatrix}=\begin{pmatrix}\|g^1\|^1 & \langle g^1,g^2\rangle\\ \langle g^1,g^2\rangle & \|g^2\|^2\end{pmatrix},$$

problem (11) can be relaxed as follows,

$$\begin{array}{ll} \max & \frac{f^2-f^\star}{f^1-f^\star} \\ \text{s.t.} & \operatorname{tr}(A_1X)-f^2+f^1\leq 0 \\ & \operatorname{tr}(A_2X)-f^1+f^2\leq 0 \\ & f^1-f^\star+\operatorname{tr}(A_3X)\leq 0 \\ & f^2-f^\star+\operatorname{tr}(A_4X)\leq 0 \\ & f^1,f^2\geq f^\star, X\succeq 0, \end{array} \tag{12}$$

where

$$A_1 = \begin{pmatrix} \frac{1+\mu L t_1^2 - 2\mu t_1}{2(L-\mu)} - t_1 & \frac{\mu t_1 - 1}{2(L-\mu)} \\ \frac{\mu t_1 - 1}{2(L-\mu)} & \frac{1}{2(L-\mu)} \end{pmatrix} \quad A_2 = \begin{pmatrix} \frac{1+\mu L t_1^2 - 2\mu t_1}{2(L-\mu)} & \frac{\mu t_1 - 1}{2(L-\mu)} + \frac{t_1}{2} \\ \frac{\mu t_1 - 1}{2(L-\mu)} + \frac{t_1}{2} & \frac{1}{2(L-\mu)} \end{pmatrix}$$
$$A_3 = \begin{pmatrix} \frac{-1}{\mu_p^2} & 0 \\ 0 & 0 \end{pmatrix} \quad A_4 = \begin{pmatrix} 0 & 0 \\ 0 & \frac{-1}{\mu_p^2} \end{pmatrix}.$$

In addition, X, f^1, f^2 are decision variables in this formulation. In the next theorem, we obtain an upper bound for problem (11) by using weak duality. This bound gives a new convergence rate for Algorithm 1 for a wide variety of functions.

Theorem 3 *Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ with $L \in (0, \infty)$, $\mu \in (-\infty, 0]$ and let f satisfy the PŁ inequality on $X = \{x : f(x) \leq f(x^1)\}$. Suppose that x^2 is generated by Algorithm* 1.

i) If $t_1 \in \left(0, \frac{1}{L}\right)$, then

$$\frac{f(x^2) - f^\star}{f(x^1) - f^\star} \leq \left(\frac{\mu_p(1 - Lt_1) + \sqrt{(L-\mu)(\mu - \mu_p)(2 - Lt_1)\mu_p t_1 + (L-\mu)^2}}{L - \mu + \mu_p} \right)^2.$$

ii) If $t_1 \in \left[\frac{1}{L}, \frac{3}{\mu + L + \sqrt{\mu^2 - L\mu + L^2}}\right]$, then

$$\frac{f(x^2) - f^\star}{f(x^1) - f^\star} \leq \left(\frac{(Lt_1 - 2)(\mu t_1 - 2)\mu_p t_1}{(L + \mu - \mu_p)t_1 - 2} + 1 \right).$$

iii) If $t_1 \in \left(\frac{3}{\mu + L + \sqrt{\mu^2 - L\mu + L^2}}, \frac{2}{L}\right)$, then

$$\frac{f(x^2) - f^\star}{f(x^1) - f^\star} \leq \frac{(Lt_1 - 1)^2}{(Lt_1 - 1)^2 + \mu_p t_1 (2 - Lt_1)}.$$

In particular, if $t_1 = \frac{1}{L}$ and $\mu = -L$, we have

$$f(x^2) - f^\star \leq \left(\frac{2L - 2\mu_p}{2L + \mu_p} \right)(f(x^1) - f^\star). \tag{13}$$

Proof First we consider $t_1 \in \left(0, \frac{1}{L}\right)$. Let

$$b_1 = \frac{(L-\mu)\big(\alpha + \mu_p\big(1 - Lt_1\big)\big)}{\alpha\big(L - \mu + \mu_p\big)}$$
$$b_2 = b_1 - \left(\frac{\alpha}{L-\mu} b_1\right)^2,$$

where

$$\alpha = \sqrt{(L-\mu)\big(\mu_p t_1\big(\mu_p - \mu\big)\big(Lt_1 - 2\big) + (L-\mu)\big)}.$$

It is readily seen that $b_1, b_2 \geq 0$. Furthermore,

$$\begin{aligned}
&f^2 - f^\star - (b_1 - b_2)\big(f^1 - f^\star\big) - b_2\left(-\frac{1}{2\mu_p}\left\|g^1\right\|^2 + f^1 - f^\star\right) \\
&\quad - \big(1 - b_1\big)\left(-\frac{1}{2\mu_p}\left\|g^2\right\|^2 + f^2 - f^\star\right) - b_1\left(\frac{1}{2(L-\mu)}\big(\|g^2\|^2\right. \\
&\qquad \left. +(1 + \mu L t_1^2 - 2\mu t_1)\|g^1\|^2 + 2(\mu t_1 - 1)\langle g^1, g^2\rangle\big) - f^1 + f^2 + \left\langle g^2, t_1 g^1\right\rangle\right) \\
&\quad = -\frac{1 - Lt_1}{2\alpha}\left\|\frac{\alpha b_1}{L-\mu} g^1 - g^2\right\|^2 \leq 0.
\end{aligned}$$

Therefore, for any feasible solution of problem (11), we have

$$\begin{aligned}
&\frac{f(x^2) - f^\star}{f(x^1) - f^\star} \\
&\quad \leq \left(\frac{\mu_p\big(1 - Lt_1\big) + \sqrt{(L-\mu)\big(\mu - \mu_p\big)\big(2 - Lt_1\big)\mu_p t_1 + (L-\mu)^2}}{L - \mu + \mu_p}\right)^2,
\end{aligned}$$

and the proof of this part is complete. Now, we consider the case that $t_1 \in \left[\frac{1}{L}, \frac{3}{\mu + L + \sqrt{\mu^2 - L\mu + L^2}}\right]$. Suppose that

$$a_1 = \frac{\mu t_1 - 1}{\big(L + \mu - \mu_p\big)t_1 - 2}, \quad a_2 = \frac{1 - Lt_1}{\big(L + \mu - \mu_p\big)t_1 - 2},$$
$$a_3 = -\frac{\big((Lt_1 - 2)(\mu t_1 - 2) - 1\big)\mu_p t_1}{\big(L + \mu - \mu_p\big)t_1 - 2}, \quad a_4 = -\frac{\mu_p t_1}{\big(L + \mu - \mu_p\big)t_1 - 2}.$$

It is readily seen that $a_1, a_2, a_3, a_4 \geq 0$. Furthermore,

$$f^2 - f^\star - (1 - a_3 - a_4)(f^1 - f^\star) - a_3\left(-\frac{1}{2\mu_p}\left\|g^1\right\|^2 + f^1 - f^\star\right)$$
$$- a_4\left(-\frac{1}{2\mu_p}\left\|g^2\right\|^2 + f^2 - f^\star\right) - a_1\left(\frac{1}{2(L-\mu)}\left(\|g^2\|^2 + (1+\mu L t_1^2 - 2\mu t_1)\|g^1\|^2\right.\right.$$
$$\left.\left.+2(\mu t_1 - 1)\langle g^1, g^2\rangle\right) - f^1 + f^2 + \left\langle g^2, t_1 g^1\right\rangle\right) - a_2\left(\frac{1}{2(L-\mu)}\left(\|g^2\|^2\right.\right.$$
$$\left.\left.+(1+\mu L t_1^2 - 2\mu t_1)\|g^1\|^2 + 2(\mu t_1 - 1)\langle g^1, g^2\rangle\right) - f^2 + f^1 - \left\langle g^1, t_1 g^1\right\rangle\right) = 0.$$

Therefore, for any feasible solution of problem (11), we have

$$f(x^2) - f^\star - \left(\frac{L\mu_p \mu t_1^3 - 2\mu_p(L+\mu)t_1^2 + 4\mu_p t_1}{(L+\mu-\mu_p)t_1 - 2} + 1\right)\left(f(x^1) - f^\star\right) \le 0.$$

Now, we prove the last part. Assume that $t_1 \in \left(\frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}, \frac{2}{L}\right)$. With some algebra, one can show

$$f^2 - f^\star - \left(\frac{(Lt_1-1)^2}{\beta}\right)(f^1 - f^\star) - \left(\frac{\mu_p t_1(2-Lt_1)}{\beta}\right)\left(-\frac{1}{2\mu_p}\left\|g^2\right\|^2 + f^2 - f^\star\right)$$
$$- \left(\frac{(Lt_1-1)(2-Lt_1)}{\beta}\right)$$
$$\left(\frac{1}{2(L-\mu)}\left(\|g^2\|^2 + (1+\mu L t_1^2 - 2\mu t_1)\|g^1\|^2 + 2(\mu t_1 - 1)\langle g^1, g^2\rangle\right)\right.$$
$$\left.-f^2 + f^1 - \left\langle g^1, t_1 g^1\right\rangle\right) - \left(\frac{Lt_1 - 1}{\beta}\right)\left(\frac{1}{2(L-\mu)}\left(\|g^2\|^2 + (1+\mu L t_1^2 - 2\mu t_1)\|g^1\|^2\right.\right.$$
$$\left.\left.+2(\mu t_1 - 1)\langle g^1, g^2\rangle\right) - f^1 + f^2 + \left\langle g^2, t_1 g^1\right\rangle\right)$$
$$= -\frac{(1-Lt_1)\left(L\mu t^2 - 2(\mu+L)t + 3\right)}{2\beta(L-\mu)}\left\|\sqrt{Lt_1 - 1}\, g^1 + \frac{1}{\sqrt{Lt_1-1}} g^2\right\|^2 \le 0,$$

where,

$$\beta = (Lt_1 - 1)^2 + \mu_p t_1(2 - Lt_1).$$

The rest of the proof is similar to that of the former cases. □

One may wonder how we obtain Lagrange multipliers (dual variables) in Theorem 3. The multipliers are computed by solving the dual of problem (12) by hand. Furthermore, Theorem 3 provides a tighter bound in comparison with the convergence rate given in Theorem 1 for L-smooth functions with $t_1 \in (0, \frac{2}{L})$. To show this, we need investigate three subintervals:

i) Suppose that $t_1 \in \left(0, \frac{1}{L}\right)$. As $1 - Lt_1 \le 0$,

$$\left(\frac{\mu_p(1-Lt_1)+\sqrt{2L(-L-\mu_p)(2-Lt_1)\mu_p t_1+4L^2}}{2L+\mu_p}\right)^2$$
$$\leq \frac{4L^2+2L\mu_p t_1(L+\mu_p)(Lt_1-2)+(\mu_p-L\mu_p t_1)^2}{(2L+\mu_p)^2} \leq 1-t_1\mu_p(2-t_1L),$$

where the last inequality follows from non-positivity of the quadratic function $T_1(t_1) = -Lt_1^2\left(2L^2+L\mu_p+\mu_p^2\right)+2t_1\left(2L^2+L\mu_p+\mu_p^2\right)-4L$ on the given interval.

ii) Let $t_1 \in \left[\frac{1}{L}, \frac{\sqrt{3}}{L}\right]$. Since $\mu_p \leq L$ and $(2-Lt_1) > 0$, we have

$$1 \leq \frac{Lt_1+2}{\mu_p t_1+2} \Rightarrow 1-\frac{(2-Lt_1)(Lt_1+2)\mu_p t_1}{\mu_p t_1+2} \leq 1-t_1\mu_p(2-Lt_1).$$

iii) Assume that $t_1 \in (\frac{\sqrt{3}}{L}, \frac{2}{L})$. It is readily verified that the quadratic function $T_2(t_1) = (Lt_1-1)^2+\mu_p t_1(2-Lt_1)-1$ is non-positive on the given interval. Hence,

$$\frac{(Lt_1-1)^2}{(Lt_1-1)^2+\mu_p t_1(2-Lt_1)} = 1-\frac{\mu_p t_1(2-Lt_1)}{(Lt_1-1)^2+\mu_p t_1(2-Lt_1)} \leq 1-t_1\mu_p(2-Lt_1).$$

Therefore, for $t_1 \in \left(0, \frac{2}{L}\right)$ the bound provided by Theorem 3 is tighter than that given by Theorem 2.

In most problems, the smoothness constant, L, is unknown. By using (2), any estimation of the smoothness constant L, say $\tilde{L}$, should satisfy the following inequality,

$$f\left(x-\tfrac{1}{\tilde{L}}\nabla f(x)\right) \leq f(x)-\tfrac{1}{2\tilde{L}}\|\nabla f(x)\|^2.$$

Thus one may try to obtain a suitable estimate by searching for a sufficiently large value of $\tilde{L}$ that satisfies this inequality. This technique is due to Nesterov; see [18, Section 3] for details.

The next proposition gives the optimal step length with respect to the worst-case convergence rate.

Proposition 1 *Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ with $L \in (0,\infty)$, $\mu \in (-\infty, 0]$ and let f satisfy the PŁ inequality on $X = \{x : f(x) \leq f(x^1)\}$. Suppose that $r(t) = L\mu(L+\mu-\mu_p)t^3-\left(L^2-\mu_p(L+\mu)+5L\mu+\mu^2\right)t^2+4(L+\mu)t-4$ and $\bar{t}$ is the unique root of r in $\left[\frac{1}{L}, \frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}\right]$ if it exists. Then $t^\star$ given by*

$$t^\star = \begin{cases} \bar{t} & \text{if } \bar{t} \text{ exists} \\ \frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}} & \text{otherwise,} \end{cases}$$

is the optimal step length for Algorithm 1 *with respect to the worst-case convergence rate.*

Proof To obtain an optimal step length, we need to solve the optimization problem

$$\min_{t\in\left(0,\frac{2}{L}\right)} h(t),$$

where h is given by

$$h(t)=\begin{cases}\left(\dfrac{\mu_p(1-Lt)+\sqrt{(L-\mu)(\mu-\mu_p)(2-Lt)\mu_p t+(L-\mu)^2}}{L-\mu+\mu_p}\right)^2 & t\in\left(0,\frac{1}{L}\right)\\ \dfrac{(Lt-2)(\mu t-2)\mu_p t}{(L+\mu-\mu_p)t-2}+1 & t\in\left[\frac{1}{L},\frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}\right]\\ \dfrac{(Lt-1)^2}{(Lt-1)^2+(2-Lt)\mu_p t} & t\in\left(\frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}},\frac{2}{L}\right).\end{cases}$$

It is easily seen that h is decreasing on $\left(0,\frac{1}{L}\right)$ and is increasing on $\left(\frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}},\frac{2}{L}\right)$. Hence, we need investigate the closed interval $\left[\frac{1}{L},\frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}\right]$. We will show that h is convex on the interval in question. First, we consider the case $L+\mu-\mu_p\leq 0$. Let $p(t)=\frac{\mu t-2}{(L+\mu-\mu_p)t-2}$ and $q(t)=(Lt-2)\mu_p t$. By some algebra, one can show the following inequalities for $t\in\left[\frac{1}{L},\frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}\right]$:

$$\begin{aligned}&p(t)\geq 0 \quad q(t)\leq 0\\ &p'(t)\geq 0 \quad q'(t)\geq 0\\ &p''(t)\leq 0 \quad q''(t)\geq 0.\end{aligned}$$

Hence, the convexity of h follows from $h''=p''q+2p'q'+pq''$. Now, we investigate the case that $L+\mu-\mu_p>0$. Suppose that $p(t)=\frac{\mu_p t}{(L+\mu-\mu_p)t-2}$ and $q(t)=(Lt-2)(\mu t-2)$. For these functions, we have the following inequalities

$$\begin{aligned}&p(t)\leq 0 \quad q(t)\geq 0\\ &p'(t)\leq 0 \quad q'(t)\leq 0\\ &p''(t)\geq 0 \quad q''(t)\leq 0,\end{aligned}$$

which analogous to the former case one can infer the convexity of h on the given interval. Hence, if h has a root in $\left[\frac{1}{L},\frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}\right]$, it will be the minimum. Otherwise, the point $t^\star=\frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}$ will be the minimum. This follows from the point that $h'(\frac{1}{L})=\frac{2L\mu_p(\mu_p-L)}{(L+\mu_p-\mu)^2}\leq 0$ and the convexity of h on the interval in question.

□

Thanks to Proposition 1, the following corollary gives the optimal step length for L-smooth convex functions satisfying the PŁ inequality.

Corollary 1 *If f is an L-smooth convex function satisfying the PŁ inequality, then the optimal step length with respect to the worst-case convergence rate is* $\min\left\{\frac{2}{L+\sqrt{L\mu_p}}, \frac{3}{2L}\right\}$.

The constant $\frac{2}{L+\sqrt{L\mu_p}}$ also appears in the the fast gradient algorithm introduced in [17] for L-smooth convex functions which are $(1, \mu_s)$-quasar-convex, see Definition 4. By Theorem 9, $(1, \mu_s)$-quasar-convexity implies the PŁ inequality with the same constant. Algorithm 2 describes the method in question.

Algorithm 2 Fast gradient method

Pick $x^1 \in \mathbb{R}^n$, set N and $y^1 = x^1$.
For $k = 1, 2, \ldots, N$ perform the following step:
1. $y^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$
2. $x^{k+1} = y^{k+1} + \frac{\sqrt{L}-\sqrt{\mu_p}}{\sqrt{L}+\sqrt{\mu_p}}\left(y^{k+1} - y^k\right)$

One can verify that Algorithm 2, at the first iteration, generates $x^2 = x^1 - \frac{2}{L+\sqrt{L\mu_p}}\nabla f(x^1)$.

A more general form of the PŁ inequality, called the Łojasiewicz inequality, may be written as

$$\left(f(x) - f^\star\right)^{2\theta} \leq \frac{1}{2\mu_p}\|\nabla f(x)\|^2, \quad \forall x \in X, \tag{14}$$

where $\theta \in (0, 1)$. It is known that when $\theta \in (0, \frac{1}{2}]$ some algorithms, including Algorithm 1, is linearly convergent; see [3, 4]. In the next theorem, we show that for functions with finite maximum and minimum curvature the Łojasiewicz inequality cannot hold for $\theta \in (0, \frac{1}{2})$.

Theorem 4 *Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ be a non-constant function. If f satisfies the Łojasiewicz inequality on $X = \{x : f(x) \leq f(x^1)\}$, then $\theta \geq \frac{1}{2}$.*

Proof To the contrary, assume that $\theta \in (0, \frac{1}{2})$. Without loss of generality, we may assume that $\mu = -L$. It is known that Algorithm 1 generates a decreasing sequence $\{f(x^k)\}$ and it is convergent, that is $\|\nabla f(x^k)\| \to 0$; see [19, page 28]. Furthermore, (14) implies that $f(x^k) \to f^\star$. Without loss of generality, we may assume that $f^\star = 0$. First, we investigate the case that $f(x^1) = 1$. The semi-definite programming problem corresponding to performance estimation in this case may be formulated as follows,

$$\begin{aligned} &\max\ f^2 \\ &\text{s.t.}\quad \operatorname{tr}(A_1X) - f^2 + 1 \le 0 \\ &\qquad \operatorname{tr}(A_2X) - 1 + f^2 \le 0 \\ &\qquad 1 + \operatorname{tr}(A_3X) \le 0 \\ &\qquad (f^2)^{2\theta} + \operatorname{tr}(A_4X) \le 0 \\ &\qquad f^2 \ge 0, X \succeq 0. \end{aligned} \tag{15}$$

Since Algorithm 1 is a monotone method, f^2 can take value in $[0, 1]$. In addition, we have $f^2 \le (f^2)^{2\theta}$ on this interval. Hence, by using Theorem 3, we get the following bound,

$$f^2 \le \frac{2L - 2\mu_p}{2L + \mu_p}.$$

Now, suppose that $f(x^1) = f^1 > 0$. Consider the function $h : \mathbb{R}^n \to \mathbb{R}$ given by $h(x) = \frac{f(x)}{f^1}$. It is seen that h is $\frac{L}{f^1}$-smooth and

$$h(x)^{2\theta} \le \frac{1}{2\mu_p (f^1)^{2\theta-2}} \|\nabla h(x)\|^2, \quad \forall x \in X.$$

As Algorithm 1 generates the same x^2 for both functions, by using the first part, we obtain

$$\frac{f(x^2)}{f(x^1)} \le \frac{2L(f^1)^{-1} - 2\mu_p (f^1)^{2\theta-2}}{2L(f^1)^{-1} + \mu_p (f^1)^{2\theta-2}} = \frac{2L - 2\mu_p (f^1)^{2\theta-1}}{2L + \mu_p (f^1)^{2\theta-1}}.$$

For f^1 sufficiently small, we have $\frac{2L - 2\mu_p (f^1)^{2\theta-1}}{2L + \mu_p (f^1)^{2\theta-1}} < 0$, which contradicts $f^\star \ge 0$ and the proof is complete. □

Necoara et al. gave necessary and sufficient conditions for linear convergence of the gradient method with constant step lengths when f is a smooth convex function; see [17, Theorem 13]. Indeed, the theorem says that Algorithm 1 is linearly convergent if and only if f has a quadratic functional growth on $\{x : f(x) \le f(x^1)\}$; see Definition 3. However, this theorem does not hold necessarily for non-convex functions. The next theorem provides necessary and sufficient conditions for linear convergence of Algorithm 1.

Theorem 5 *Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$. Algorithm 1 is linearly convergent to the optimal value if and only if f satisfies PŁ inequality on $\{x : f(x) \le f(x^1)\}$.*

Proof Let $\bar{x} \in \{x : f(x) \le f(x^1)\}$. Linear convergence implies the existence of $\gamma \in [0, 1)$ with

$$f(\hat{x}) - f^\star \le \gamma \big(f(\bar{x}) - f^\star\big), \tag{16}$$

where $\hat{x} = \bar{x} - \frac{1}{L}\nabla f(\bar{x})$. By (3), we have $f(\bar{x}) - f(\hat{x}) \le \frac{2L-\mu}{2L^2} \|\nabla f(\bar{x})\|^2$. By using this inequality with (16), we get

$$f(\bar{x}) - f^\star \leq \frac{1}{1-\gamma}(f(\bar{x}) - f(\hat{x})) \leq \frac{2L-\mu}{2L^2(1-\gamma)}\|\nabla f(\bar{x})\|^2,$$

which shows that f satisfies PŁ inequality on $\{x : f(x) \leq f(x^1)\}$. The other implication follows from Theorem 3. □

3 The PŁ inequality: relation to some classes of functions

In this section, we study some classes of functions for which Algorithm 1 may be linearly convergent. We establish that these classes of functions satisfy the PŁ inequality under mild assumptions, and we infer the linear convergence by using Theorem 3. Moreover, one can get convergence rates by applying performance estimation.

Throughout the section, we denote the optimal solution set of problem (1) by $X^\star$ and we assume that $X^\star$ is non-empty. We denote the distance function to $X^\star$ by $d_{X^\star}(x) := \inf_{y \in X^\star} \|y - x\|$. The set-valued mapping $\Pi_{X^\star}(x)$ stands for the projection of x on $X^\star$, that is, $\Pi_{X^\star}(x) = \{y : \|y - x\| = d_{X^\star}(x)\}$. Note that, as $X^\star$ is non-empty closed set, $\Pi_{X^*}(x)$ exists and is well-defined.

Definition 2 Let $\mu_g > 0$. A function f has a quadratic gradient growth on $X \subseteq \mathbb{R}^n$ if

$$\langle \nabla f(x), x - x^\star \rangle \geq \mu_g d^2_{X^\star}(x), \quad \forall x \in X, \tag{17}$$

for some $x^\star \in \Pi_{X^\star}(x)$.

Note that inequality (2) implies that $\mu_g \leq L$. Hu et al. [15] investigated the convergence rate $\{x^k\}$ when f satisfies (17) and $X^\star$ is singleton. To the best knowledge of the authors, there is no convergence rate result in terms of $\{f(x^k)\}$ for functions with a quadratic gradient growth. The next proposition states that quadratic gradient growth property implies the PŁ inequality.

Proposition 2 *Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$. If f has a quadratic gradient growth on $X \subseteq \mathbb{R}^n$ with $\mu_g > 0$, then f satisfies the PŁ inequality with $\mu_p = \frac{\mu_g^2}{L}$.*

Proof Suppose that $x^\star \in \Pi_{X^\star}(x)$ satisfies (17). By the Cauchy-Schwarz inequality, we have

$$\mu_g \|x - x^\star\| \leq \|\nabla f(x)\|. \tag{18}$$

On the other hand, (2) implies that

$$f(x) \leq f(x^\star) + \tfrac{L}{2}\|x - x^\star\|^2. \tag{19}$$

The PŁ inequality follows from (18) and (19). □

By Proposition 2 and Theorem 3, one can infer the linear convergence of Algorithm 1 when f has a quadratic gradient growth on $X = \{x : f(x) \leq f(x^1)\}$. Indeed, one can derive the following bound if $t_1 = \frac{1}{L}$ and $\mu = -L$,

$$f(x^2) - f^\star \leq \left(\frac{2L^2 - 2\mu_g^2}{2L^2 + \mu_g^2}\right)\big(f(x^1) - f^\star\big). \tag{20}$$

Nevertheless, by using the performance estimation method, one can derive a better bound than the bound given by (20). The performance estimation problem for $t_1 = \frac{1}{L}$ in this case may be formulated as

$$\begin{aligned}
\max\ & \tfrac{f^2 - f^\star}{f^1 - f^\star} \\
\text{s.t.}\ & \{x^k, g^k, f^k\} \cup \{y^k, 0, f^\star\} \text{ satisfy interpolation constraints (9) for } k \in \{1,2\} \\
& x^2 = x^1 - \tfrac{1}{L} g^1 \\
& f^k \geq f^\star \quad k \in \{1,2\} \\
& \langle g^k, x^k - y^k \rangle \geq \mu_g \|y^k - x^k\|^2, \quad k \in \{1,2\} \\
& \|x^1 - y^1\|^2 \leq \|x^1 - y^2\|^2 \\
& \|x^2 - y^2\|^2 \leq \|x^2 - y^1\|^2.
\end{aligned} \tag{21}$$

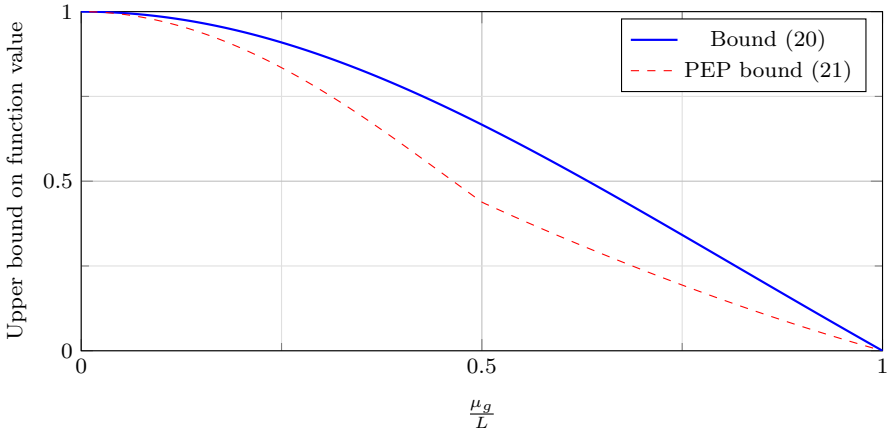

Fig. 1 Convergence rate computed by performance estimation (*red line*) and the bound given by (20) (*blue line*) for $\frac{\mu_g}{L} \in (0,1)$. (color figure online)

Analogous to Sect. 2, one can obtain an upper bound for problem (21) by solving a semidefinite program. Our numerical results show that the bounds generated by performance estimation is tighter than bound (20); see Fig. 1. We do not have a closed-form bound on the optimal value of (21), though.

Definition 3 [17, Definition 4], [19, Definition 4.1.2] Let $\mu_q > 0$. A function f has a quadratic functional growth on $X \subseteq \mathbb{R}^n$ if

$$\frac{\mu_q}{2} d^2_{X^\star}(x) \leq f(x) - f^\star, \quad \forall x \in X. \tag{22}$$

It is readily seen that, contrary to the previous situations, the quadratic functional growth property does not necessarily imply that each stationary point is a global optimal solution. The next theorem investigates the relationship between quadratic functional growth property and other notions.

Theorem 6 *Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ and let $X = \{x : f(x) \leq f(x^1)\}$. We have the following implications*:

i) (4) $\Rightarrow$ (22) *with* $\mu_q = \mu_p$.
ii) *If* $\mu_q > \frac{-\mu L}{L-\mu}$, *then* (22) $\Rightarrow$ (17) *with* $\mu_g = \frac{\mu_q}{2}(1 - \frac{\mu}{L}) + \frac{\mu}{2}$.
iii) *If*

$$f(x) - f(x^\star) \leq \langle \nabla f(x), x - x^\star \rangle, \ \forall x \in X,$$

for some $x^\star \in \Pi_{X^\star}(x)$ then (22) $\Rightarrow$ (17) with $\mu_g = \frac{\mu_q}{2}$.

Proof One can establish *i*) similarly to the proof of [16, Theorem 2]. Consider part *ii*). Let $x \in X$ and $x^\star \in \Pi_{X^\star}(x)$ with $d_{X^\star}(x) = \|x - x^\star\|$. By (9), we have

$$f(x) - f(x^\star) \leq \frac{-1}{2(L-\mu)} \|\nabla f(x)\|^2 - \frac{\mu L}{2(L-\mu)} \|x - x^\star\|^2 + \frac{L}{L-\mu} \langle \nabla f(x), x - x^\star \rangle.$$

As $\frac{\mu_q}{2} \|x - x^\star\|^2 \leq f(x) - f(x^\star)$, we get

$$\left(\frac{\mu_q}{2} \left(1 - \frac{\mu}{L} \right) + \frac{\mu}{2} \right) \|x - x^\star\|^2 \leq \langle \nabla f(x), x - x^\star \rangle,$$

which establishes the desired inequality. Part *iii*) is proved similarly to the former case. □

By Theorem 3, it is clear that Algorithm 1 enjoys linear convergence rate if f has a quadratic gradient growth on $X = \{x : f(x) \leq f(x^1)\}$ and if f satisfies assumptions *ii*) or *iii*) in Theorem 6. For instance, if $\mu = -L$ and $\mu_q \in (\frac{L}{2}, L)$, one can derive the following convergence rate for Algorithm 1 for fixed step length $t_k = \frac{1}{L}$, $k \in \{1, ..., N\}$,

$$f(x^{N+1}) - f(x^1) \leq \left(\frac{2L^2 - 2(\mu_q - \frac{L}{2})^2}{2L^2 + (\mu_q - \frac{L}{2})^2} \right)^N \big(f(x^1) - f^\star\big). \tag{23}$$

It is interesting to compare the convergence rate (23) to the convergence rate obtained by using the performance estimation framework. In this case, the performance estimation problem may be cast as follows,

$$\begin{aligned}
&\max \ \tfrac{f^{N+1}-f^\star}{f^1-f^\star} \\
&\text{s.t.} \quad \{x^k, g^k, f^k\} \cup \{y^k, 0, f^\star\} \text{ satisfy inequality (9) for } k \in \{1, ..., N+1\} \\
&\qquad x^{k+1} = x^k - \tfrac{1}{L} g^k, \ \ k \in \{1, ..., N\} \\
&\qquad f^k \geq f^\star \ \ k \in \{1, ..., N\} \\
&\qquad f^k - f^\star \geq \tfrac{\mu_q}{2} \|x^k - y^k\|^2, \ \ k \in \{1, ..., N+1\} \\
&\qquad \|x^k - y^k\|^2 \leq \|x^k - y^{k'}\|^2, \ \ k \in \{1, ..., N+1\}, k' \in \{1, ..., N+1\}.
\end{aligned} \tag{24}$$

Since $x^{k+1} = x^k - \frac{1}{L} g^k$, we get $x^{k+1} = x^1 - \frac{1}{L} \sum_{l=1}^k g^l$. Hence, problem (24) may be reformulated as follows,

$$\begin{aligned}
&\max \ \tfrac{f^{N+1}-f^\star}{f^1-f^\star} \\
&\text{s.t.} \quad \left\{ x^1 - \tfrac{1}{L} \sum_{l=1}^{k-1} g^l, g^k, f^k \right\} \cup \{y^k, 0, f^\star\} \text{ satisfy interpolation constraints (9)} \\
&\qquad f^k \geq f^\star \ \ k \in \{1, ..., N\} \\
&\qquad f^k - f^\star \geq \tfrac{\mu_q}{2} \left\| x^1 - \tfrac{1}{L} \sum_{l=1}^{k-1} g^l - y^k \right\|^2, \ \ k \in \{1, ..., N+1\} \\
&\qquad \left\| x^1 - \tfrac{1}{L} \sum_{l=1}^{k-1} g^l - y^k \right\|^2 \leq \left\| x^1 - \tfrac{1}{L} \sum_{l=1}^{k-1} g^l - y^{k'} \right\|^2, \ \ k, k' \in \{1, ..., N+1\}.
\end{aligned} \tag{25}$$

The next theorem provides an upper bound for problem (25) by using weak duality.

Theorem 7 *Let $f \in \mathcal{F}_{-L,L}(\mathbb{R}^n)$ and let f have a quadratic functional growth on $X = \{x : f(x) \leq f(x^1)\}$ with $\mu_q \in (\frac{L}{2}, L)$. If $t_k = \frac{1}{L}$, $k \in \{1, ..., N\}$, then we have the following convergence rate for Algorithm* 1,

$$f(x^{N+1}) - f(x^1) \leq \frac{L}{\mu_q} \left(2 - \frac{2\mu_q}{L}\right)^N \big(f(x^1) - f^\star\big). \tag{26}$$

Proof The proof is analogous to that of Theorem 3. Without loss of generality, we may assume that $f^\star = 0$. By some algebra, one can show that

$$
\begin{aligned}
&f^{N+1}-f^\star-\tfrac{L}{\mu_q}\Big(2-\tfrac{2\mu_q}{L}\Big)^N\big(f^1-f^\star\big)+\sum_{j=1}^{N+1}\left(2^{N+1-j}\Big(1-\tfrac{\mu_q}{L}\Big)^{N-1}\right)\\
&\times\left(f^\star-f^j-\left\langle g^j,y^1-x^1+\tfrac{1}{L}\sum_{l=1}^{j-1}g^l\right\rangle-\tfrac{1}{2L}\|g^j\|^2+\tfrac{L}{4}\left\|y^1-x^1+\tfrac{1}{L}\sum_{l=1}^{j-1}g^l+\tfrac{1}{L}g^j\right\|^2\right)\\
&+\sum_{i=2}^{N}\sum_{j=i}^{N+1}\left(2^{N+1-j}\Big(\tfrac{\mu_q}{L}\Big)\Big(1-\tfrac{\mu_q}{L}\Big)^{N-i}\right)\left(f^\star-f^j-\langle g^j,y^i-x^1+\tfrac{1}{L}\sum_{l=1}^{j-1}g^l\rangle\right.\\
&\left.-\tfrac{1}{2L}\|g^j\|^2+\tfrac{L}{4}\left\|y^i-x^1+\tfrac{1}{L}\sum_{l=1}^{j-1}g^l+\tfrac{1}{L}g^j\right\|^2\right)+\sum_{j=2}^{N}\left(2^{N+1-j}(1-\tfrac{\mu_q}{L})^{N-j}\right)\\
&\times\left(f^j-f^\star-\tfrac{\mu_q}{2}\left\|y^j-x^1+\tfrac{1}{L}\sum_{l=1}^{j-1}g^l\right\|^2\right)+\left(2^N\Big(1-\tfrac{\mu_q}{L}\Big)^{N-1}+\tfrac{L}{\mu_q}\Big(2-\tfrac{2\mu_q}{L}\Big)^N\right)\\
&\times\left(f^1-f^\star-\tfrac{\mu_q}{2}\|y^1-x^1\|^2\right)=-\left(\tfrac{L}{4}\Big(1-\tfrac{\mu_q}{L}\Big)^{N-1}\|y^1-x^1+\tfrac{1}{L}\sum_{l=1}^{N+1}g^l\|^2\right)\\
&-\sum_{i=2}^{N}\left(\tfrac{\mu_q}{4}\Big(1-\tfrac{\mu_q}{L}\Big)^{N-i}\left\|y^i-x^1+\tfrac{1}{L}\sum_{l=1}^{N+1}g^l\right\|^2\right)\le 0.
\end{aligned}
$$

By using the above inequality, we get

$$
f^{N+1}-f^\star\le\tfrac{L}{\mu_q}\Big(2-\tfrac{2\mu_q}{L}\Big)^N\big(f^1-f^\star\big),
$$

for any feasible point of (25), and the proof is complete. □

By doing some calculus, one can verify the following inequality

$$
\frac{2L^2-2\Big(\mu_q-\frac{L}{2}\Big)^2}{2L^2+\Big(\mu_q-\frac{L}{2}\Big)^2}\ge\Big(2-\tfrac{2\mu_q}{L}\Big),\quad \mu_q\in\Big(\tfrac{L}{2},L\Big).
$$

Hence, Theorem 7 provides a tighter bound than (23).

Definition 4 [14, Definition 1] Let $\gamma\in(0,1]$ and $\mu_s\ge 0$. A function f is called (γ,μ_s) -quasar-convex on $X\subseteq\mathbb{R}^n$ with respect to $x^\star\in\operatorname{argmin}_{x\in\mathbb{R}^n}f(x)$ if

$$
f(x)+\tfrac{1}{\gamma}\langle\nabla f(x),x^\star-x\rangle+\tfrac{\mu_s}{2}\|x^\star-x\|^2\le f^\star,\quad\forall x\in X. \tag{27}
$$

The class of quasar-convex functions is large. For instance, non-negative homogeneous functions are (1, 0)-quasar-convex on $\mathbb{R}^n$. (Recall that a function $f:\mathbb{R}^n\to\mathbb{R}$ is called homogeneous of degree k if $f(\alpha x)=\alpha^k f(x)$ for all $x\in\mathbb{R}^n,\alpha\in\mathbb{R}$.) Indeed, if f is non-negative homogeneous of degree $k\ge 1$, by the Euler identity, we have

$$f(x) + \langle \nabla f(x), x^\star - x \rangle = (1-k) f(x) \leq 0, \ \forall x \in \mathbb{R}^n,$$

where $x^\star = 0$. In what follows, we list some convergence results concerning quasar-convex functions for Algorithm 1.

Theorem 8 [5, *Remark* 4.3] *Let f be L-smooth and let f be (γ, μ_s)-quasar-convex on $X = \{x : f(x) \leq f(x^1)\}$. If $t_1 = \frac{1}{L}$ and if x^2 is from Algorithm* 1, *then*

$$f(x^2) - f^\star \leq \left(1 - \frac{\gamma^2 \mu_s}{L}\right)\left(f(x^1) - f^\star\right). \tag{28}$$

In the following theorem, we state the relationship between quasar-convexity and other concepts. Before we get to the theorem, we recall star convexity. A set X is called star convex at $x^\star$ if

$$\lambda x + (1-\lambda) x^\star \in X, \ \forall x \in X, \forall \lambda \in [0,1].$$

Theorem 9 *Let $x^\star$ be the unique solution of problem (1) and let $X = \{x : f(x) \leq f(x^1)\}$. If X is star convex at $x^\star$, then we have the following implications*:

i) (27) $\Rightarrow$ (17) *with* $\mu_g = \frac{\mu_s \gamma}{2} + \frac{\mu_s \gamma^2}{4}$.

ii) (17) $\Rightarrow$ (27) *with* $\mu_s = \ell - \frac{L}{2}$ *and* $\gamma = \frac{\mu_g}{\ell}$ *for each* $\ell \in (\max(\frac{L}{2}, \mu_g), \infty)$.

iii) (27) $\Rightarrow$ (4) *with* $\mu_p = \mu_s \gamma^2$.

Proof The proof of *i)* is similar in spirit to the proof of Theorem 1 in [17]. Let $x \in X$. By the fundamental theorem of calculus and (27),we have

$$\begin{aligned} f(x) - f(x^\star) &= \int_0^1 \tfrac{1}{\lambda} \langle \nabla f(\lambda x + (1-\lambda)x^\star), \lambda x + (1-\lambda)x^\star - x^\star \rangle d\lambda \\ &\geq \int_0^1 \tfrac{\gamma}{\lambda} \Big(f(\lambda x + (1-\lambda)x^\star) - f(x^\star) + \tfrac{\mu_s \lambda^2}{2} \|x - x^\star\|^2 \Big) d\lambda \\ &\geq \tfrac{\gamma \mu_s}{4} \|x - x^\star\|^2, \end{aligned}$$

where the last inequality follows from the global optimality of $x^\star$. By summing $f(x) - f(x^\star) \geq \frac{\gamma \mu_s}{4} \|x - x^\star\|^2$ and (27), we get the desired inequality. Now, we prove part *ii)*. Let $x \in \mathbb{R}^n$ and $\ell \in (\max(\frac{L}{2}, \mu_g), \infty)$. By (2), we have

$$f(x) \leq f(x^\star) + \frac{L}{2} \|x - x^\star\|^2. \tag{29}$$

By using (29) and (17), we get

$$f(x) + \left(\frac{\ell}{\mu_g}\right)\langle \nabla f(x), x^\star - x\rangle + \left(\ell - \frac{L}{2}\right)\|x - x^\star\|^2 \leq f(x^\star).$$

For the proof of *iii*), we refer the reader to [5, Lemma 3.2]. □

By combining Theorem 3 and Theorem 9, under the assumptions of Theorem 8, one can get the following convergence rate for Algorithm 1 with $t_1 = \frac{1}{L}$,

$$f(x^2) - f^\star \leq \left(\frac{2L - 2\mu_s\gamma^2}{2L + \mu_s\gamma^2}\right)\big(f(x^1) - f^\star\big),$$

which is tighter the bound given in Theorem 8.

4 Concluding remarks

In this paper we studied the convergence rate of the gradient method with fixed step lengths for smooth functions satisfying the PŁ inequality. We gave a new convergence rate, which is sharper than known bounds in the literature. One important question which remains to be addressed is the computation of the tightest bound for Algorithm 1. Moreover, the performance analysis of fast gradient methods, like Algorithm 2, for these classes of functions may also be of interest.

We only studied the linear convergence in terms of the convergence of objective values. However, one can also infer the linear convergence in terms of distance to the solution set or the norm of the gradient by using our results. For instance, under the assumption of Theorem 3, we have

$$\tfrac{\mu_p}{2} d^2_{X^\star}(x^{k+1}) \leq f(x^{k+1}) - f^\star \leq \gamma^k\big(f(x^1) - f^\star\big) \leq \tfrac{L\gamma^k}{2} d^2_{X^\star}(x^1),$$

where the first inequality follows from Theorem 6, γ is the linear convergence rate given in Theorem 3, and the last inequality resulted from (2). Hence,

$$d^2_{X^\star}(x^{k+1}) \leq \tfrac{L\gamma^k}{\mu_p} d^2_{X^\star}(x^1).$$

Moreover, the quadratic gradient growth is a necessary and sufficient conditions for the linear convergence in terms of distance to the solution set; see [24, Theorem 3.4]. Note that the PŁ inequality and the quadratic gradient growth are equivalent.

Acknowledgment This work was supported by the Dutch Scientific Council (NWO) Grant OCENW. GROOT.2019.015, *Optimization for and with Machine Learning (OPTIMAL)*.

References

1. Abbaszadehpeivasti, H., de Klerk, E., Zamani, M.: The exact worst-case convergence rate of the gradient method with fixed step lengths for L-smooth functions. Opt. Lett., pp. 1–13 (2021)
2. Abbaszadehpeivasti, H., de Klerk, E., Zamani, M.: On the rate of convergence of the difference-of-convex algorithm (DCA). arXiv preprint arXiv:2109.13566 (2021)
3. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**(1), 5–16 (2009)
4. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
5. Bu, J., Mesbahi, M.: A note on Nesterov's accelerated method in nonconvex optimization: a weak estimate sequence approach. arXiv preprint arXiv:2006.08548 (2020)
6. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points I. Math. Program. **184**, 1–50 (2019)
7. Danilova, M., Dvurechensky, P., Gasnikov, A., Gorbunov, E., Guminov, S., Kamzolov, D., Shibaev, I.: Recent theoretical advances in non-convex optimization. arXiv preprint arXiv:2012.06188 (2020)
8. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Opt. **29**(1), 207–239 (2019)
9. 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. Opt. Lett. **11**(7), 1185–1199 (2017)
10. De Klerk, E., Glineur, F., Taylor, A.B.: Worst-case convergence analysis of inexact gradient and Newton methods through semidefinite programming performance estimation. SIAM J. Opt. **30**(3), 2053–2082 (2020)
11. Drori, Y.: Contributions to the complexity analysis of optimization algorithms. Ph.D. thesis, Tel-Aviv University (2014)
12. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: a novel approach. Math. Program. **145**(1), 451–482 (2014)
13. Gupta, S.D., Van Parys, B.P., Ryu, E.K.: Branch-and-bound performance estimation programming: a unified methodology for constructing optimal optimization methods. arXiv preprint arXiv:2203.07305 (2022)
14. Hinder, O., Sidford, A., Sohoni, N.: Near-optimal methods for minimizing star-convex functions and beyond. In: Conference on Learning Theory, pp. 1894–1938. PMLR (2020)
15. Hu, B., Seiler, P., Lessard, L.: Analysis of biased stochastic gradient descent using sequential semidefinite programs. Math. Program. **187**(1), 383–408 (2021)
16. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 795–811. Springer (2016)
17. Necoara, I., Nesterov, Y., Glineur, F.: Linear convergence of first order methods for non-strongly convex optimization. Math. Program. **175**(1), 69–107 (2019)
18. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
19. Nesterov, Y.: Lectures on Convex Optimization, vol. 137. Springer (2018)
20. Polyak, B.T.: Gradient methods for the minimisation of functionals. USSR Computational Mathematics and Mathematical Physics **3**(4), 864–878 (1963)
21. Rotaru, T., Glineur, F., Panagiotis, P.: Tight convergence rates of the gradient method on hypoconvex functions. arXiv preprint arXiv:2203.00775 (2022)
22. 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)
23. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Math. Program. **161**(1), 307–345 (2017)

24. Zamani, M., Abbaszadehpeivasti, H., de Klerk, E.: Convergence rate analysis of the gradient descent-ascent method for convex-concave saddle-point problems. arXiv preprint arXiv:2209.01272 (2022)