CrossMark

On the Convergence Analysis of the Optimized Gradient Method

Donghwan Kim[1] · **Jeffrey A. Fessler**[1]

Abstract This paper considers the problem of unconstrained minimization of smooth convex functions having Lipschitz continuous gradients with known Lipschitz constant. We recently proposed the optimized gradient method for this problem and showed that it has a worst-case convergence bound for the cost function decrease that is twice as small as that of Nesterov's fast gradient method, yet has a similarly efficient practical implementation. Drori showed recently that the optimized gradient method has optimal complexity for the cost function decrease over the general class of first-order methods. This optimality makes it important to study fully the convergence properties of the optimized gradient method. The previous worst-case convergence bound for the optimized gradient method was derived for only the last iterate of a secondary sequence. This paper provides an analytic convergence bound for the primary sequence generated by the optimized gradient method. We then discuss additional convergence properties of the optimized gradient method, including the interesting fact that the optimized gradient method has two types of worst-case functions: a piecewise affine-quadratic function and a quadratic function. These results help complete the theory of an optimal first-order method for smooth convex minimization.

Keywords First-order algorithms · Optimized gradient method · Convergence bound · Smooth convex minimization · Worst-case performance analysis

✉ Donghwan Kim
kimdongh@umich.edu

Jeffrey A. Fessler
fessler@umich.edu

[1] University of Michigan, Ann Arbor, MI, USA

Mathematics Subject Classification 90C25 · 90C30 · 90C60 · 68Q25 · 49M25 · 90C22

1 Introduction

We recently proposed the optimized gradient method (OGM) [1] for unconstrained smooth convex minimization problems, building upon Drori and Teboulle [2]. We showed in [1] that OGM has a worst-case cost function convergence bound that is twice as small as that of Nesterov's fast gradient method (FGM) [3], yet has an efficient implementation that is similar to FGM. In addition, Drori [4] showed that OGM achieves the optimal worst-case convergence bound of the cost function decrease over the general class of first-order methods (for large-dimensional problems), making it important to further study the convergence properties of OGM.

The worst-case convergence bound for OGM was derived for only the last iterate of a secondary sequence in [1], and this paper additionally provides an analytic convergence bound for the primary sequence generated by OGM by extending the analysis in [1]. We further discuss convergence properties of OGM, including the interesting fact that OGM has two types of worst-case functions: a piecewise affine-quadratic function and a quadratic function. These results complement our understanding of an optimal first-order method for smooth convex minimization.

2 Problem, Algorithms and Contributions

We consider the unconstrained smooth convex minimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad f(\boldsymbol{x}) \tag{M}$$

with the following two conditions:

- $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function of the type $\mathcal{C}_L^{1,1}(\mathbb{R}^d)$, i.e., continuously differentiable with Lipschitz continuous gradient:

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

  where $L > 0$ is the Lipschitz constant.
- The optimal set $X_*(f) = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$ is nonempty, i.e., problem (M) is solvable.

We use $\mathcal{F}_L(\mathbb{R}^d)$ to denote the class of functions that satisfy the above conditions hereafter.

For large-scale optimization problems of type (M) that arise in various fields such as communications, machine learning and signal processing, general first-order algorithms that query only the cost function values and gradients are attractive because of their mild dependence on the problem dimension [5]. For simplicity, we initially focus on the class of *fixed-step* first-order (FO) algorithms having the following form:

Algorithm Class FO

Input: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$.

For $i = 0, \ldots, N - 1$

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \frac{1}{L} \sum_{k=0}^{i} h_{i+1,k} \nabla f(\boldsymbol{x}_k). \qquad (1)$$

FO updates use weighted sums of current and previous gradients $\{\nabla f(\boldsymbol{x}_k)\}_{k=0}^{i}$ with (predetermined) step sizes $\{h_{i+1,k}\}_{k=0}^{i}$ and the Lipschitz constant L. Class FO includes the (fixed-step) gradient method (GM), the heavy-ball method [6], Nesterov's fast gradient method (FGM) [3,7], and the recently introduced optimized gradient method (OGM) [1]. Those four methods have efficient recursive formulations rather than directly using (1) that would require storing all previous gradients and computing weighted summations every iteration. Among class FO, Nesterov's FGM has been used widely, since it achieves the optimal rate $O(1/N^2)$ for decreasing a cost function in N iterations [8] and has two efficient forms as shown below for smooth convex problems.

Algorithm FGM1 [3]

Input: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\boldsymbol{x}_0 = \boldsymbol{y}_0 \in \mathbb{R}^d$, $t_0 = 1$.

For $i = 0, \ldots, N - 1$

$$\boldsymbol{y}_{i+1} = \boldsymbol{x}_i - \frac{1}{L} \nabla f(\boldsymbol{x}_i)$$

$$t_{i+1} = \frac{1 + \sqrt{1 + 4t_i^2}}{2},$$

$$\boldsymbol{x}_{i+1} = \boldsymbol{y}_{i+1} + \frac{t_i - 1}{t_{i+1}} (\boldsymbol{y}_{i+1} - \boldsymbol{y}_i)$$

Algorithm FGM2 [7]

Input: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, $t_0 = 1$.

For $i = 0, \ldots, N - 1$

$$\boldsymbol{y}_{i+1} = \boldsymbol{x}_i - \frac{1}{L} \nabla f(\boldsymbol{x}_i)$$

$$\boldsymbol{z}_{i+1} = \boldsymbol{x}_0 - \frac{1}{L} \sum_{k=0}^{i} t_k \nabla f(\boldsymbol{x}_k)$$

$$t_{i+1} = \frac{1 + \sqrt{1 + 4t_i^2}}{2},$$

$$\boldsymbol{x}_{i+1} = \left(1 - \frac{1}{t_{i+1}}\right) \boldsymbol{y}_{i+1} + \frac{1}{t_{i+1}} \boldsymbol{z}_{i+1}$$

Both FGM1 and FGM2 produce identical sequences $\{\boldsymbol{y}_i\}$ and $\{\boldsymbol{x}_i\}$, where the primary sequence $\{\boldsymbol{y}_i\}$ satisfies the following convergence bound [3,7] for any $1 \leq i \leq N$:

$$f(\boldsymbol{y}_i) - f(\boldsymbol{x}_*) \leq \frac{LR^2}{2t_{i-1}^2} \leq \frac{2LR^2}{(i+1)^2}. \qquad (2)$$

In [1], we showed that the secondary sequence $\{\boldsymbol{x}_i\}$ of FGM satisfies the following convergence bound that is similar to (2) for any $1 \leq i \leq N$:

$$f(\boldsymbol{x}_i) - f(\boldsymbol{x}_*) \leq \frac{LR^2}{2t_i^2} \leq \frac{2LR^2}{(i+2)^2}. \qquad (3)$$

Taylor et al. [9] demonstrated that the upper bounds (2) and (3) are only asymptotically tight.

When the large-scale condition "$d \geq 2N + 1$" holds, Nesterov [8] showed that for any first-order method generating $\boldsymbol{x}_N$ after N iterations, there exists a function ϕ in $\mathcal{F}_L(\mathbb{R}^d)$ that satisfies the following lower bound:

$$\frac{3L||\boldsymbol{x}_0 - \boldsymbol{x}_*||^2}{32(N + 1)^2} \leq \phi(\boldsymbol{x}_N) - \phi(\boldsymbol{x}_*). \tag{4}$$

Although FGM achieves the optimal rate $O(1/N^2)$, one can still seek algorithms that improve upon the constant factor in (2) and (3), in light of the gap between the bounds (2), (3) of FGM and the lower complexity bound (4). Building upon Drori and Teboulle (hereafter "DT")'s approach [2] of seeking FO methods that are faster than Nesterov's FGM (reviewed in Sect. 3.3), we recently proposed following two efficient formulations of OGM [1].

Algorithm OGM1

Input: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\boldsymbol{x}_0 = \boldsymbol{y}_0 \in \mathbb{R}^d$, $\theta_0 = 1$.
For $i = 0, \ldots, N - 1$

$$\boldsymbol{y}_{i+1} = \boldsymbol{x}_i - \frac{1}{L}\nabla f(\boldsymbol{x}_i)$$

$$\theta_{i+1} = \begin{cases} \frac{1+\sqrt{1+4\theta_i^2}}{2}, & i \leq N - 2 \\ \frac{1+\sqrt{1+8\theta_i^2}}{2}, & i = N - 1 \end{cases}$$

$$\boldsymbol{x}_{i+1} = \boldsymbol{y}_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}(\boldsymbol{y}_{i+1} - \boldsymbol{y}_i)$$
$$\qquad\qquad + \frac{\theta_i}{\theta_{i+1}}(\boldsymbol{y}_{i+1} - \boldsymbol{x}_i)$$

Algorithm OGM2

Input: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\theta_0 = 1$.
For $i = 0, \ldots, N - 1$

$$\boldsymbol{y}_{i+1} = \boldsymbol{x}_i - \frac{1}{L}\nabla f(\boldsymbol{x}_i)$$

$$\boldsymbol{z}_{i+1} = \boldsymbol{x}_0 - \frac{1}{L}\sum_{k=0}^{i} 2\theta_k \nabla f(\boldsymbol{x}_k)$$

$$\theta_{i+1} = \begin{cases} \frac{1+\sqrt{1+4\theta_i^2}}{2}, & i \leq N - 2 \\ \frac{1+\sqrt{1+8\theta_i^2}}{2}, & i = N - 1 \end{cases}$$

$$\boldsymbol{x}_{i+1} = \left(1 - \frac{1}{\theta_{i+1}}\right)\boldsymbol{y}_{i+1} + \frac{1}{\theta_{i+1}}\boldsymbol{z}_{i+1}$$

OGM1 and OGM2 have computational efficiency comparable to FGM1 and FGM2, and produce identical primary sequence $\{\boldsymbol{y}_i\}$ and secondary sequence $\{\boldsymbol{x}_i\}$. The last iterate $\boldsymbol{x}_N$ of OGM satisfies the following analytical worst-case bound [1, Theorem 2]:

$$f(\boldsymbol{x}_N) - f(\boldsymbol{x}_*) \leq \frac{LR^2}{2\theta_N^2} \leq \frac{LR^2}{(N + 1)(N + 1 + \sqrt{2})}, \tag{5}$$

which is twice as small as those for FGM in (2) and (3). Recently for the condition "$d \geq N + 1$", Drori [4] showed that for any first-order method, there exists a function ψ in $\mathcal{F}_L(\mathbb{R}^d)$ that cannot be minimized faster than the following lower bound:

$$\frac{L||\boldsymbol{x}_0 - \boldsymbol{x}_*||^2}{2\theta_N^2} \leq \psi(\boldsymbol{x}_N) - \psi(\boldsymbol{x}_*), \tag{6}$$

where $\boldsymbol{x}_N$ is the Nth iterate of any first-order method. This lower complexity bound (6) improves on (4) and exactly matches the bound (5) of OGM, showing that OGM achieves the *optimal* worst-case bound of the cost function for first-order methods

when $d \geq N + 1$. What is remarkable about Drori's result is that OGM was derived by optimizing over the class FO having fixed step sizes, leading to (5), whereas Drori's lower bound in (6) is for the general class of first-order methods where the step sizes are arbitrary. It is interesting that OGM with its fixed-step sizes is optimal over the apparently much broader class.

Because OGM has such optimality, it is desirable to understand its properties thoroughly. For example, analytical bounds for the primary sequence $\{y_i\}$ of OGM have not been studied previously, although numerical bounds were discussed by Taylor et al. [9]. This paper provides analytical bounds for the primary sequence of OGM, augmenting the convergence analysis of x_N of OGM given in [1]. We also relate OGM to another version of Nesterov's accelerated first-order method in [10] that has a similar formulation as OGM2.

In [1, Theorem 3], we specified a worst-case function for which OGM achieves the first upper bound in (5) exactly. The corresponding worst-case function is the following piecewise affine-quadratic function:

$$f_{1,\text{OGM}}(\boldsymbol{x}; N) = \begin{cases} \frac{LR}{\theta_N^2}||\boldsymbol{x}|| - \frac{LR^2}{2\theta_N^4}, & \text{if } ||\boldsymbol{x}|| \geq \frac{R}{\theta_N^2}, \\ \frac{L}{2}||\boldsymbol{x}||^2, & \text{otherwise,} \end{cases} \qquad (7)$$

where OGM iterates remain in the affine region with the same gradient value (without overshooting) for all N iterations. Section 5 shows that a simple quadratic function is also a worst-case function for OGM and describes why it is interesting that the optimal OGM has these two types of worst-case functions.

Section 3 reviews DT's Performance Estimation Problem (PEP) framework in [2] that enables systematic worst-case performance analysis of optimization methods. Section 4 provides new convergence analysis for the primary sequence of OGM. Section 5 discusses the two types of worst-case functions for OGM, and Sect. 6 concludes.

3 Prior Work: Performance Estimation Problem (PEP)

Exploring the convergence performance of optimization methods including class FO has a long history. DT [2] were the first to cast the analysis of the worst-case performance of optimization methods into an optimization problem called PEP, reviewed in this section. We also review how we developed OGM [1] that is built upon DT's PEP.

3.1 Review of PEP

To analyze the worst-case convergence behavior of a method in class FO having given step sizes $\boldsymbol{h} = \{h_{i,k}\}_{0 \leq k < i \leq N}$, DT's PEP [2] bounds the decrease of the cost function after N iterations as

$$\mathcal{B}_{\mathrm{P}}(\boldsymbol{h}, N, d, L, R) := \max_{\substack{f \in \mathcal{F}_L(\mathbb{R}^d), \\ \boldsymbol{x}_0, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d, \\ \boldsymbol{x}_* \in X_*(f)}} f(\boldsymbol{x}_N) - f(\boldsymbol{x}_*) \tag{P}$$

$$\text{s.t.} \quad \begin{cases} ||\boldsymbol{x}_0 - \boldsymbol{x}_*|| \le R, \\ \boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \frac{1}{L}\sum_{k=0}^{i} h_{i+1,k} \nabla f(\boldsymbol{x}_k), \quad i = 0, \ldots, N-1, \end{cases}$$

for given dimension d, Lipschitz constant L and the distance R between an initial point $\boldsymbol{x}_0$ and an optimal point $\boldsymbol{x}_* \in X_*(f)$.

Since problem (P) is difficult to solve, DT [2] introduced a series of relaxations. Then the upper bound of the worst-case performance was found numerically in [2] by solving a relaxed PEP problem. For some cases, analytical worst-case bounds were revealed in [1,2], where some of those analytical bounds were even found to be exact despite the relaxations. On the other hand, Taylor et al. [9] studied the numerical tight worst-case bound of (P) by avoiding DT's one relaxation step that is not guaranteed to be tight and showing the tightness of the rest of DT's relaxations in [2] (for the condition "$d \ge N + 2$").

To summarize recent PEP studies, DT extended the PEP approach for nonsmooth convex problems [11], Drori's thesis [12] includes an extension of PEP to projected gradient methods for constrained smooth convex problems, and Taylor et al. [13] studied PEP for various first-order algorithms for solving composite convex problems. Similarly but using different relaxations of (P), Lessard et al. [14] applied the Integral quadratic constraints to (P), leading to simpler computation but slightly looser convergence upper bounds.

The next two sections review relaxations of DT's PEP and an approach for optimizing the choice of $\boldsymbol{h}$ for FO using PEP in [2].

3.2 Review of DT's Relaxation on PEP

This section reviews relaxations introduced by DT to make (P) into a simpler semidefinite programming (SDP) problem. DT first relax the functional constraint $f \in \mathcal{F}_L(\mathbb{R}^d)$ by a well-known property of the class of $\mathcal{F}_L(\mathbb{R}^d)$ functions in [8, Theorem 2.1.5] and then further relax as follows:

$$\mathcal{B}_{\mathrm{P1}}(\boldsymbol{h}, N, d, L, R) := \max_{\substack{\boldsymbol{G} \in \mathbb{R}^{(N+1)d}, \\ \boldsymbol{\delta} \in \mathbb{R}^{N+1}}} LR^2 \delta_N \tag{P1}$$

$$\text{s.t.} \quad \begin{cases} \frac{1}{2}||\boldsymbol{g}_{i-1} - \boldsymbol{g}_i||^2 \le \delta_{i-1} - \delta_i - \left\langle \boldsymbol{g}_i, \sum_{k=0}^{i-1} h_{i,k}\boldsymbol{g}_k \right\rangle, \quad i = 1, \ldots, N, \\ \frac{1}{2}||\boldsymbol{g}_i||^2 \le -\delta_i - \left\langle \boldsymbol{g}_i, \sum_{j=1}^{i}\sum_{k=0}^{j-1} h_{j,k}\boldsymbol{g}_k + \boldsymbol{v} \right\rangle, \quad i = 0, \ldots, N, \end{cases}$$

for any given unit vector $\boldsymbol{v} \in \mathbb{R}^d$, where we denote $\boldsymbol{g}_i := \frac{1}{L||\boldsymbol{x}_0 - \boldsymbol{x}_*||}\nabla f(\boldsymbol{x}_i)$ and $\delta_i := \frac{1}{L||\boldsymbol{x}_0 - \boldsymbol{x}_*||^2}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_*))$ for $i = 0, \ldots, N, *$, and group them into $\boldsymbol{G} = [\boldsymbol{g}_0 \ldots \boldsymbol{g}_N]^\top \in \mathbb{R}^{(N+1)\times d}$ and $\boldsymbol{\delta} = [\delta_0 \ldots \delta_N]^\top \in \mathbb{R}^{N+1}$.

Maximizing relaxed problem (P1) is still difficult, so DT [2] use a duality approach on (P1). Replacing $\max_{G,\delta} LR^2\delta_N$ by $\min_{G,\delta} -\delta_N$ for convenience, the Lagrangian of the corresponding constrained minimization problem (P1) with dual variables $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)^\top \in \mathbb{R}_+^N$ and $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_N)^\top \in \mathbb{R}_+^{N+1}$ becomes

$$
\mathcal{L}(\boldsymbol{G}, \boldsymbol{\delta}, \boldsymbol{\lambda}, \boldsymbol{\tau}; \boldsymbol{h}) := -\delta_N + \sum_{i=1}^N \lambda_i (\delta_i - \delta_{i-1}) + \sum_{i=0}^N \tau_i \delta_i
$$
$$
+ \operatorname{Tr}\left\{ \boldsymbol{G}^\top \boldsymbol{S}(\boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\tau}) \boldsymbol{G} + \boldsymbol{v}\boldsymbol{\tau}^\top \boldsymbol{G} \right\}, \tag{8}
$$

where

$$
\begin{cases}
\boldsymbol{S}(\boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\tau}) := \sum_{i=1}^N \lambda_i \boldsymbol{A}_{i-1,i}(\boldsymbol{h}) + \sum_{i=0}^N \tau_i \boldsymbol{D}_i(\boldsymbol{h}), \\
\boldsymbol{A}_{i-1,i}(\boldsymbol{h}) := \frac{1}{2}(\boldsymbol{u}_{i-1} - \boldsymbol{u}_i)(\boldsymbol{u}_{i-1} - \boldsymbol{u}_i)^\top + \frac{1}{2}\sum_{k=0}^{i-1} h_{i,k}\left(\boldsymbol{u}_i \boldsymbol{u}_k^\top + \boldsymbol{u}_k \boldsymbol{u}_i^\top\right), \\
\boldsymbol{D}_i(\boldsymbol{h}) := \frac{1}{2}\boldsymbol{u}_i \boldsymbol{u}_i^\top + \frac{1}{2}\sum_{j=1}^i \sum_{k=0}^{j-1} h_{j,k}\left(\boldsymbol{u}_i \boldsymbol{u}_k^\top + \boldsymbol{u}_k \boldsymbol{u}_i^\top\right),
\end{cases} \tag{9}
$$

and $\boldsymbol{u}_i = \boldsymbol{e}_{i+1} \in \mathbb{R}^{N+1}$ is the $(i+1)$th standard basis vector.

Using further derivations of a duality approach on (8) in [2], the dual problem of (P1) becomes the following SDP problem:

$$
\mathcal{B}_{\mathrm{P}}(\boldsymbol{h}, N, d, L, R)
$$
$$
\leq \mathcal{B}_{\mathrm{D}}(\boldsymbol{h}, N, L, R) := \min_{\substack{(\boldsymbol{\lambda}, \boldsymbol{\tau}) \in \Lambda, \\ \gamma \in \mathbb{R}}} \left\{ \frac{1}{2} LR^2 \gamma : \begin{pmatrix} \boldsymbol{S}(\boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\tau}) & \frac{1}{2}\boldsymbol{\tau} \\ \frac{1}{2}\boldsymbol{\tau}^\top & \frac{1}{2}\gamma \end{pmatrix} \succeq 0 \right\}, \tag{D}
$$

where $\Lambda = \left\{ (\boldsymbol{\lambda}, \boldsymbol{\tau}) \in \mathbb{R}_+^{2N+1} : \begin{array}{l} \tau_0 = \lambda_1, \quad \lambda_N + \tau_N = 1, \\ \lambda_i - \lambda_{i+1} + \tau_i = 0, \ i = 1, \ldots, N-1, \end{array} \right\}$. Then, for given $\boldsymbol{h}$, the bound $\mathcal{B}_{\mathrm{D}}(\boldsymbol{h}, N, L, R)$ (that is not guaranteed to be tight) can be numerically computed using any SDP solver, while analytical upper bounds $\mathcal{B}_{\mathrm{D}}(\boldsymbol{h}, N, L, R)$ for some choices of $\boldsymbol{h}$ were found in [1,2]. Section 4 finds a new analytical upper bound for a modified version of $\mathcal{B}_{\mathrm{D}}$.

3.3 Review of Optimizing the Step Sizes Using PEP

In addition to finding upper bounds for given FO methods, DT [2] searched for the best FO methods with respect to the worst-case performance. Ideally one would like to optimize $\boldsymbol{h}$ over problem (P):

$$
\hat{\boldsymbol{h}}_{\mathrm{P}} := \underset{\boldsymbol{h} \in \mathbb{R}^{N(N+1)/2}}{\arg\min} \mathcal{B}_{\mathrm{P}}(\boldsymbol{h}, N, d, L, R). \tag{HP}
$$

However, optimizing (HP) directly seems impractical, so DT minimized the dual
problem (D) using a SDP solver over the coefficients $\boldsymbol{h}$ as

$$\hat{\boldsymbol{h}}_{\mathrm{D}} := \underset{\boldsymbol{h} \,\in\mathbb{R}^{N(N+1)/2}}{\arg\min} \ \mathcal{B}_{\mathrm{D}}(\boldsymbol{h}\,, N, L, R). \tag{HD}$$

Due to relaxations, the computed $\hat{\boldsymbol{h}}_{\mathrm{D}}$ is not guaranteed to be optimal for problem (HP).
Nevertheless, we show in [1] that solving (HD) leads to an algorithm (OGM) having
a convergence bound that is twice as small as that of FGM. Interestingly, OGM is
optimal among first-order methods with $d \geq N + 1$ [4], i.e., $\hat{\boldsymbol{h}}_{\mathrm{D}}$ is a solution of
both (HP) and (HD) for $d \geq N + 1$. An optimal point $(\hat{\boldsymbol{h}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\tau}}, \hat{\gamma})$ of (HD) is given
in [1, Lemma 4 and Proposition 3] as follows:

$$\hat{h}_{i+1,k} = \begin{cases} \frac{\theta_i-1}{\theta_{i+1}}\hat{h}_{i,k}, & k = 0, \ldots, i - 2, \\ \frac{\theta_i-1}{\theta_{i+1}}(\hat{h}_{i,i-1} - 1), & k = i - 1, \\ 1 + \frac{2\theta_i-1}{\theta_{i+1}}, & k = i, \end{cases} \tag{10}$$

$$= \begin{cases} \frac{1}{\theta_{i+1}}\left(2\theta_k - \sum_{j=k+1}^{i} \hat{h}_{j,k}\right), & k = 0, \ldots, i - 1, \\ 1 + \frac{2\theta_i-1}{\theta_{i+1}}, & k = i, \end{cases} \tag{11}$$

$$\hat{\lambda}_i = \frac{2\theta_{i-1}^2}{\theta_N^2}, \quad i = 1, \ldots, N, \quad \hat{\tau}_i = \begin{cases} \frac{2\theta_i}{\theta_N^2}, & i = 0, \ldots, N - 1, \\ \frac{1}{\theta_N}, & i = N, \end{cases} \quad \hat{\gamma} = \frac{1}{\theta_N^2}. \tag{12}$$

Thus both OGM1 and OGM2 satisfy the convergence bound (5) [1, Theorem 2, Propo-
sitions 4 and 5].

4 New Convergence Analysis for the Primary Sequence of OGM

4.1 Relaxed PEP for the Primary Sequence of OGM

This section applies PEP to an iterate $\boldsymbol{y}_N$ of the class FO$'$ that generates a sequence
$\{\boldsymbol{y}_i\}$:

$$\boldsymbol{y}_{i+1} = \boldsymbol{x}_i - \frac{1}{L}\nabla f(\boldsymbol{x}_i), \quad i = 0, \ldots, N - 1, \tag{13}$$

using a sequence $\{\boldsymbol{x}_i\}$ generated by the class FO, complementing the worst-case per-
formance of $\boldsymbol{x}_N$ in the previous section.

We first replace $f(\boldsymbol{x}_N) - f(\boldsymbol{x}_*)$ in (P) by $f(\boldsymbol{y}_{N+1}) - f(\boldsymbol{x}_*)$ as follows:

$$\mathcal{B}_{P'}(\boldsymbol{h}, N, d, L, R) := \max_{\substack{f \in \mathcal{F}_L(\mathbb{R}^d), \\ \boldsymbol{x}_0, \dots, \boldsymbol{x}_N, \boldsymbol{y}_{N+1} \in \mathbb{R}^d, \\ \boldsymbol{x}_* \in X_*(f)}} f(\boldsymbol{y}_{N+1}) - f(\boldsymbol{x}_*) \tag{P'}$$

$$\text{s.t.} \begin{cases} ||\boldsymbol{x}_0 - \boldsymbol{x}_*|| \le R, \quad \boldsymbol{y}_{N+1} = \boldsymbol{x}_N - \frac{1}{L}\nabla f(\boldsymbol{x}_N), \\ \boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \frac{1}{L} \sum_{k=0}^{i} h_{i+1,k} \nabla f(\boldsymbol{x}_k), \quad i = 0, \dots, N-1. \end{cases}$$

We could directly repeat relaxations on (P′) as reviewed in Sect. 3.2, but we found it difficult to solve a such relaxed problem of (P′) analytically. Instead, we use the following inequality [8]:

$$f\left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right) \le f(\boldsymbol{x}) - \frac{1}{2L}||\nabla f(\boldsymbol{x})||^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^d. \tag{14}$$

to relax (P′), leading to the following bound:

$$\mathcal{B}_{P1'}(\boldsymbol{h}, N, d, L, R) := \max_{\substack{f \in \mathcal{F}_L(\mathbb{R}^d), \\ \boldsymbol{x}_0, \dots, \boldsymbol{x}_N \in \mathbb{R}^d, \\ \boldsymbol{x}_* \in X_*(f)}} f(\boldsymbol{x}_N) - \frac{1}{2L}||\nabla f(\boldsymbol{x}_N)||^2 - f(\boldsymbol{x}_*) \tag{P1'}$$

$$\text{s.t.} \begin{cases} ||\boldsymbol{x}_0 - \boldsymbol{x}_*|| \le R, \\ \boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \frac{1}{L} \sum_{k=0}^{i} h_{i+1,k} \nabla f(\boldsymbol{x}_k), \quad i = 0, \dots, N-1. \end{cases}$$

This bound has an additional term $-\frac{1}{2L}||\nabla f(\boldsymbol{x}_N)||^2$ compared to (P). We later show that the increase of the worst-case upper bound due to this strict relaxation step using (14) is negligible asymptotically.

Similar to relaxing from (P) to (P1) in Sect. 3.2, we relax (P1′) to the following bound:

$$\mathcal{B}_{P2'}(\boldsymbol{h}, N, d, L, R) := \max_{\substack{\boldsymbol{G} \in \mathbb{R}^{(N+1)d}, \\ \boldsymbol{\delta} \in \mathbb{R}^{N+1}}} LR^2 \left(\delta_N - \frac{1}{2}||\boldsymbol{g}_N||^2\right) \tag{P2'}$$

$$\text{s.t.} \begin{cases} \frac{1}{2}||\boldsymbol{g}_{i-1} - \boldsymbol{g}_i||^2 \le \delta_{i-1} - \delta_i - \left\langle \boldsymbol{g}_i, \sum_{k=0}^{i-1} h_{i,k}\boldsymbol{g}_k \right\rangle, \quad i = 1, \dots, N, \\ \frac{1}{2}||\boldsymbol{g}_i||^2 \le -\delta_i - \left\langle \boldsymbol{g}_i, \sum_{j=1}^{i} \sum_{k=0}^{j-1} h_{j,k}\boldsymbol{g}_k + \boldsymbol{v} \right\rangle, \quad i = 0, \dots, N, \end{cases}$$

for any given unit vector $\boldsymbol{v} \in \mathbb{R}^d$. Then, as in Sect. 3.2 and [1,2], one can show that the dual problem of (P2′) is the following SDP problem

$$\mathcal{B}_{P'}(\boldsymbol{h}, N, d, L, R)$$

$$\le \mathcal{B}_{D'}(\boldsymbol{h}, N, L, R) := \min_{\substack{(\boldsymbol{\lambda}, \boldsymbol{\tau}) \in \Lambda, \\ \gamma \in \mathbb{R}}} \left\{ \frac{1}{2} LR^2\gamma : \begin{pmatrix} S(\boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\tau}) + \frac{1}{2}\boldsymbol{u}_N\boldsymbol{u}_N^\top & \frac{1}{2}\boldsymbol{\tau} \\ \frac{1}{2}\boldsymbol{\tau}^\top & \frac{1}{2}\gamma \end{pmatrix} \succeq 0 \right\}, \tag{D'}$$

by considering that the Lagrangian of (P2$'$) becomes

$$
\mathcal{L}'(\boldsymbol{G}, \boldsymbol{\delta}, \boldsymbol{\lambda}, \boldsymbol{\tau}; \boldsymbol{h}) := -\delta_N + \sum_{i=1}^{N} \lambda_i (\delta_i - \delta_{i-1}) + \sum_{i=0}^{N} \tau_i \delta_i
$$

$$
+ \mathrm{Tr}\left\{ \boldsymbol{G}^{\top}\left(\boldsymbol{S}(\boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\tau}) + \frac{1}{2} \boldsymbol{u}_N \boldsymbol{u}_N^{\top} \right) \boldsymbol{G} + \boldsymbol{v}\boldsymbol{\tau}^{\top}\boldsymbol{G} \right\} \quad (15)
$$

when we replace $\max_{\boldsymbol{G},\boldsymbol{\delta}} LR^2 \left(\delta_N - \frac{1}{2}\|\boldsymbol{g}_N\|^2 \right)$ in (P2$'$) by $\min_{\boldsymbol{G},\boldsymbol{\delta}} \left(-\delta_N + \frac{1}{2}\|\boldsymbol{g}_N\|^2 \right)$ for simplicity as we did for (P1) and (8). The formulation (15) is similar to (8), except the term $\frac{1}{2}\boldsymbol{u}_N \boldsymbol{u}_N^{\top}$. The derivation of (D$'$) and (15) is omitted here, since it is almost identical to the derivation of (D) and (8) in [1,2].

4.2 Convergence Analysis for the Primary Sequence of OGM

To find an upper bound for (D$'$), it suffices to specify a feasible point.

Lemma 4.1 *The following choice of* $(\hat{\boldsymbol{h}}', \hat{\boldsymbol{\lambda}}', \hat{\boldsymbol{\tau}}', \hat{\gamma}')$ *is a feasible point of* (D$'$):

$$
\hat{h}'_{i+1,k} = \begin{cases} \frac{t_i - 1}{t_{i+1}} \hat{h}'_{i,k}, & k = 0, \ldots, i-2, \\ \frac{t_i - 1}{t_{i+1}} (\hat{h}'_{i,i-1} - 1), & k = i-1, \\ 1 + \frac{2t_i - 1}{t_{i+1}}, & k = i, \end{cases} \quad (16)
$$

$$
= \begin{cases} \frac{1}{t_{i+1}} \left(2t_k - \sum_{j=k+1}^{i} \hat{h}'_{j,k} \right), & k = 0, \ldots, i-1, \\ 1 + \frac{2t_i - 1}{t_{i+1}}, & k = i, \end{cases} \quad (17)
$$

$$
\hat{\lambda}'_i = \frac{t_{i-1}^2}{t_N^2}, \quad i = 1, \ldots, N, \quad \hat{\tau}'_i = \frac{t_i}{t_N^2}, \quad i = 0, \ldots, N, \quad \hat{\gamma}' = \frac{1}{2t_N^2}. \quad (18)
$$

Proof The equivalency between (16) and (17) follows from [1, Proposition 3]. Also, it is obvious that $(\hat{\boldsymbol{\lambda}}', \hat{\boldsymbol{\tau}}') \in \Lambda$ using $t_i^2 = \sum_{k=0}^{i} t_k$.

We next rewrite $\boldsymbol{S}(\hat{\boldsymbol{h}}', \hat{\boldsymbol{\lambda}}', \hat{\boldsymbol{\tau}}')$ to show that the choice $(\hat{\boldsymbol{h}}', \hat{\boldsymbol{\lambda}}', \hat{\boldsymbol{\tau}}', \hat{\gamma}')$ satisfies the positive semidefinite condition in (D$'$). For any $\boldsymbol{h}$ and $(\boldsymbol{\lambda}, \boldsymbol{\tau}) \in \Lambda$, the (i, k)th entry of the symmetric matrix $\boldsymbol{S}(\boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\tau})$ in (9) can be written as

$$
S_{i,k}(\boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\tau})
$$
$$
= \begin{cases} \frac{1}{2}\left((\lambda_i + \tau_i) h_{i,k} + \tau_i \sum_{j=k+1}^{i-1} h_{j,k} \right), & i = 2, \ldots, N, \ k = 0, \ldots, i-2, \\ \frac{1}{2}\left((\lambda_i + \tau_i) h_{i,k} - \lambda_i \right), & i = 1, \ldots, N, \ k = i-1, \\ \lambda_{i+1}, & i = 0, \ldots, N-1, \ k = i, \\ \frac{1}{2}, & i = N, \ k = i. \end{cases} \quad (19)
$$

Inserting $\hat{\boldsymbol{h}}', \hat{\boldsymbol{\lambda}}'$ and $\hat{\boldsymbol{\tau}}'$ into (19), we get

$$
\begin{aligned}
&S_{i,k}(\hat{\boldsymbol{h}}', \hat{\boldsymbol{\lambda}}', \hat{\boldsymbol{\tau}}') + \frac{1}{2}[u_N u_N^T]_{i,k} \\
&= \begin{cases}
\frac{1}{2}\left(\frac{t_i^2}{t_N^2}\frac{1}{t_i}\left(2t_k - \sum_{j=k+1}^{i-1}\hat{h}'_{j,k}\right) + \frac{t_i}{t_N^2}\sum_{j=k+1}^{i-1}\hat{h}'_{j,k}\right), \\
\qquad\qquad\qquad\qquad i = 2,\ldots,N,\ k=0,\ldots,i-2, \\
\frac{1}{2}\left(\frac{t_i^2}{t_N^2}\left(1 + \frac{2t_{i-1}-1}{t_i}\right) - \frac{t_{i-1}^2}{t_N^2}\right), \quad i=1,\ldots,N,\ k=i-1, \\
\frac{t_i^2}{t_N^2}, \qquad\qquad\qquad\qquad\quad i=0,\ldots,N-1,\ k=i, \\
1, \qquad\qquad\qquad\qquad\qquad i=N,\ k=i,
\end{cases} \\
&= \frac{t_i t_k}{t_N^2}
\end{aligned}
$$

where the second equality uses $t_i^2 - t_i - t_{i-1}^2 = 0$.

Finally, using $\hat{\gamma}'$, we have

$$
\begin{pmatrix} S(\hat{\boldsymbol{h}}', \hat{\boldsymbol{\lambda}}', \hat{\boldsymbol{\tau}}') + \frac{1}{2}u_N u_N^\top & \frac{1}{2}\hat{\boldsymbol{\tau}}' \\ \frac{1}{2}\hat{\boldsymbol{\tau}}'^\top & \frac{1}{2}\hat{\gamma}' \end{pmatrix} = \begin{pmatrix} \frac{1}{t_N^2}\boldsymbol{t}\,\boldsymbol{t}^\top & \frac{1}{2t_N^2}\boldsymbol{t} \\ \frac{1}{2t_N^2}\boldsymbol{t}^\top & \frac{1}{4t_N^2} \end{pmatrix} = \frac{1}{t_N^2}\begin{pmatrix} \boldsymbol{t} \\ \frac{1}{2} \end{pmatrix}\begin{pmatrix} \boldsymbol{t} \\ \frac{1}{2} \end{pmatrix}^\top \succeq 0,
$$

where $\boldsymbol{t} = (t_0, \ldots, t_N)^\top$. □

Since $\hat{\boldsymbol{h}}$ (10) and $\hat{\boldsymbol{h}}'$ (16) are identical except for the last iteration, the intermediate iterates $\{x_i\}_{i=0}^{N-1}$ of FO with both $\hat{\boldsymbol{h}}$ and $\hat{\boldsymbol{h}}'$ are equivalent. We can also easily notice that the sequence $\{y_i\}_{i=0}^N$ of FO' with both $\hat{\boldsymbol{h}}$ and $\hat{\boldsymbol{h}}'$ are also identical, implying that both the primary sequence $\{y_i\}$ of OGM and FO' with $\hat{\boldsymbol{h}}'$ are equivalent.

Using Lemma 4.1, the following theorem provides an analytical convergence bound for the primary sequence $\{y_i\}$ of OGM.

Theorem 4.1 *Let $f \in \mathcal{F}_L(\mathbb{R}^d)$ and let $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_N \in \mathbb{R}^d$ be generated by OGM1 and OGM2. Then for $1 \le i \le N$, the primary sequence for OGM satisfies:*

$$
f(\boldsymbol{y}_i) - f(\boldsymbol{x}_*) \le \frac{LR^2}{4t_{i-1}^2} \le \frac{LR^2}{(i+1)^2}. \tag{20}
$$

Proof The sequence $\{y_i\}_{i=0}^N$ generated by FO' with $\hat{\boldsymbol{h}}'$ is equivalent to that of OGM1 and OGM2 [1, Propositions 4 and 5].

Using $\hat{\gamma}'$ (18) and $t_i^2 \ge \frac{(i+2)^2}{4}$, we have

$$
f(\boldsymbol{y}_N) - f(\boldsymbol{x}_*) \le \mathcal{B}_{\mathrm{D}'}(\hat{\boldsymbol{h}}', N-1, L, R) = \frac{1}{2}LR^2\hat{\gamma}' = \frac{LR^2}{4t_{N-1}^2} \le \frac{LR^2}{(N+1)^2}, \tag{21}
$$

based on Lemma 4.1. Since the primary sequence $\{\boldsymbol{y}_i\}_{i=0}^N$ of OGM1 and OGM2 does not depend on a given N, we can extend (21) for all $1 \leq i \leq N$. $\qquad\square$

Due to a strict relaxation leading to (P1′), we cannot guarantee that the bound (20) is tight. However, the next proposition shows that bound (20) is asymptotically tight by specifying one particular worst-case function that was conjectured by Taylor et al. [9, Conjecture 4].

Proposition 4.1 *For the following function in* $\mathcal{F}_L(\mathbb{R}^d)$:

$$
f_{1,\mathrm{OGM}'}(\boldsymbol{x}; N) = \begin{cases} \dfrac{LR}{2t_{N-1}^2+1}||\boldsymbol{x}|| - \dfrac{LR^2}{2(2t_{N-1}^2+1)^2}, & \text{if } ||\boldsymbol{x}|| \geq \dfrac{R}{2t_{N-1}^2+1}, \\ \dfrac{L}{2}||\boldsymbol{x}||^2, & \text{otherwise,} \end{cases} \tag{22}
$$

the iterate $\boldsymbol{y}_N$ *generated by OGM1 and OGM2 provides the following lower bound:*

$$
\frac{LR^2}{4t_{N-1}^2+2} = f_{1,\mathrm{OGM}'}(\boldsymbol{y}_N; N) - f_{1,\mathrm{OGM}'}(\boldsymbol{x}_*; N) \leq \max_{\substack{f \in \mathcal{F}_L(\mathbb{R}^d), \\ \boldsymbol{x}_* \in X_*(f)}} f(\boldsymbol{y}_N) - f(\boldsymbol{x}_*).
$$
$$\tag{23}$$

Proof Starting from $\boldsymbol{x}_0 = R\boldsymbol{v}$, where $\boldsymbol{v}$ is a unit vector, and using the following property of the coefficients $\hat{\boldsymbol{h}}'$ [1, Equation (8.2)]:

$$
\sum_{j=1}^{i} \sum_{k=0}^{j-1} \hat{h}'_{j,k} = t_i^2 - 1, \quad i = 1, \ldots, N, \tag{24}
$$

the primary iterates of OGM1 and OGM2 are as follows

$$
\begin{aligned}
\boldsymbol{y}_i &= \boldsymbol{x}_{i-1} - \frac{1}{L}\nabla f_{1,\mathrm{OGM}'}(\boldsymbol{x}_{i-1}; N) \\
&= \boldsymbol{x}_0 - \frac{1}{L}\sum_{j=1}^{i-1}\sum_{k=0}^{j-1} \hat{h}'_{j,k}\nabla f_{1,\mathrm{OGM}'}(\boldsymbol{x}_k; N) - \frac{1}{L}\nabla f_{1,\mathrm{OGM}'}(\boldsymbol{x}_{i-1}; N) \\
&= \left(1 - \frac{t_{i-1}^2}{2t_{N-1}^2+1}\right)R\boldsymbol{v}, \quad i = 1, \ldots, N,
\end{aligned}
$$

where the corresponding sequence $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ stays in the affine region of the function $f_{1,\mathrm{OGM}'}(\boldsymbol{x}; N)$ with the same gradient value:

$$
\nabla f_{1,\mathrm{OGM}'}(\boldsymbol{x}_i; N) = \nabla f_{1,\mathrm{OGM}'}(\boldsymbol{y}_{i+1}; N) = \frac{LR}{2t_{N-1}^2+1}\boldsymbol{v}, \quad i = 0, \ldots, N-1.
$$

Therefore, after N iterations of OGM1 and OGM2, we have

$$f_{1,\text{OGM}'}(\boldsymbol{y}_N; N) - f_{1,\text{OGM}'}(\boldsymbol{x}_*; N) = f_{1,\text{OGM}'}\left(\frac{t_{N-1}^2+1}{2t_{N-1}^2+1}R\boldsymbol{v}; N\right) = \frac{LR^2}{2\left(2t_{N-1}^2+1\right)},$$

exactly matching the lower bound (23). □

The lower bound (23) matches the tight numerical worst-case bound in [9] (see Table 1). While Taylor et al. [9] provide numerical evidence about the tight bound of the primary sequence of OGM, our (23) provides an analytical bound that suffices for asymptotically tight worst-case analysis.

4.3 New Formulations of OGM

Using [1, Propositions 4 and 5], Algorithm FO$'$ with the coefficients $\hat{\boldsymbol{h}}'$ (16) and (17) can be implemented efficiently as follows:

Algorithm OGM1$'$	**Algorithm OGM2$'$**
Input: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\boldsymbol{x}_0 = \boldsymbol{y}_0 \in \mathbb{R}^d$, $t_0 = 1$.	Input: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, $t_0 = 1$.
For $i = 0, \ldots, N-1$	For $i = 0, \ldots, N-1$
$\boldsymbol{y}_{i+1} = \boldsymbol{x}_i - \dfrac{1}{L}\nabla f(\boldsymbol{x}_i)$	$\boldsymbol{y}_{i+1} = \boldsymbol{x}_i - \dfrac{1}{L}\nabla f(\boldsymbol{x}_i)$
$t_{i+1} = \dfrac{1+\sqrt{1+4t_i^2}}{2}$	$\boldsymbol{z}_{i+1} = \boldsymbol{x}_0 - \dfrac{1}{L}\displaystyle\sum_{k=0}^{i} 2t_k \nabla f(\boldsymbol{x}_k)$
$\boldsymbol{x}_{i+1} = \boldsymbol{y}_{i+1} + \dfrac{t_i - 1}{t_{i+1}}(\boldsymbol{y}_{i+1} - \boldsymbol{y}_i)$ $\qquad\qquad + \dfrac{t_i}{t_{i+1}}(\boldsymbol{y}_{i+1} - \boldsymbol{x}_i)$	$t_{i+1} = \dfrac{1+\sqrt{1+4t_i^2}}{2}$ $\boldsymbol{x}_{i+1} = \left(1 - \dfrac{1}{t_{i+1}}\right)\boldsymbol{y}_{i+1} + \dfrac{1}{t_{i+1}}\boldsymbol{z}_{i+1}$

The OGM$'$ is very similar to OGM, because it generates same primary and secondary sequence; only, the last iterate of the secondary sequence differs. Therefore, the bound (20) applies to the primary sequence $\{\boldsymbol{y}_i\}$ of both OGM and OGM$'$, as summarized in the following corollary.

Corollary 4.1 *Let $f \in \mathcal{F}_L(\mathbb{R}^d)$ and let $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_N \in \mathbb{R}^d$ be generated by OGM1$'$ and OGM2$'$. Then for $1 \leq i \leq N$,*

$$f(\boldsymbol{y}_i) - f(\boldsymbol{x}_*) \leq \frac{LR^2}{4t_{i-1}^2} \leq \frac{LR^2}{(i+1)^2}. \tag{25}$$

4.4 Comparing Tight Worst-Case Bounds of FGM, OGM and OGM$'$

While some analytical upper bounds of FGM, OGM and OGM$'$ such as (2), (3) (5), (20) and (25) are available for comparison, some of those are tight only asymptotically

Table 1 Exact numerical cost function bound of the last primary iterate $\boldsymbol{y}_N$ and the last secondary iterate $\boldsymbol{x}_N$ of FGM, OGM and OGM$'$ (The values of the denominator d_N of $f(\cdot) - f(x_*) \leq \frac{LR^2}{d_N}$ are shown)

N	FGM (prim.)	FGM (second.)	OGM (prim.)	OGM (second.)	OGM$'$ (second.)
1	6.00	6.00	6.00	8.00	5.24
2	10.00	11.13	12.47	16.16	9.62
3	15.13	17.35	21.25	26.53	15.12
4	21.35	24.66	32.25	39.09	21.71
5	28.66	33.03	45.42	53.80	29.38
10	81.07	90.69	143.23	159.07	83.54
20	263.65	283.55	494.68	525.09	269.56
40	934.89	975.10	1810.08	1869.22	947.55
80	3490.22	3570.75	6866.93	6983.13	3516.00

or some bounds for such algorithms are even unknown analytically. Therefore, we used the code of Taylor et al. [9] for tight (numerical) comparison of algorithms of interest for some given N. Table 1 provides tight numerical bounds of the primary and secondary sequence of FGM, OGM and OGM$'$. Interestingly, the worst-case performance of the secondary sequence of OGM$'$ is worse than that of FGM sequences, whereas the primary sequence of OGM (and OGM$'$) is roughly twice better.

The following proposition uses a quadratic function to define a lower bound on the worst-case performance of OGM1$'$ and OGM2$'$.

Proposition 4.2 *For the following quadratic function in $\mathcal{F}_L(\mathbb{R}^d)$:*

$$f_2(\boldsymbol{x}) = \frac{L}{2}||\boldsymbol{x}||^2 \tag{26}$$

both OGM1$'$ and OGM2$'$ provide the following lower bound:

$$\frac{LR^2}{2t_i^2} = f_2(\boldsymbol{x}_i) - f_2(\boldsymbol{x}_*) \leq \max_{\substack{f \in \mathcal{F}_L(\mathbb{R}^d), \\ \boldsymbol{x}_* \in X_*(f)}} f(\boldsymbol{x}_i) - f(\boldsymbol{x}_*), \tag{27}$$

Proof We use induction to show that the following iterates:

$$\boldsymbol{x}_i = (-1)^i \frac{1}{t_i} R\boldsymbol{v}, \quad i = 0, \ldots, N, \tag{28}$$

correspond to the iterates of OGM1$'$ and OGM2$'$ applied to $f_2(\boldsymbol{x})$. Starting from $\boldsymbol{x}_0 = R\boldsymbol{v}$, where $\boldsymbol{v}$ is a unit vector, and assuming that (28) holds for $i < N$, we have

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \frac{1}{L} \sum_{k=0}^{i} \hat{h}'_{i+1,k} \nabla f_2(\boldsymbol{x}_k)$$

$$= \left(\boldsymbol{x}_i - \frac{1}{L}\hat{h}'_{i+1,i}\nabla f_2(\boldsymbol{x}_i)\right) - \frac{1}{L}\sum_{k=0}^{i-1}\frac{t_i - 1}{t_{i+1}}\hat{h}'_{i,k}\nabla f_2(\boldsymbol{x}_k) + \frac{1}{L}\frac{t_i - 1}{t_{i+1}}\nabla f_2(\boldsymbol{x}_{i-1})$$

$$= \frac{1 - 2t_i}{t_{i+1}}\boldsymbol{x}_i + \frac{t_i - 1}{t_{i+1}}(\boldsymbol{x}_i - \boldsymbol{x}_{i-1}) + \frac{t_i - 1}{t_{i+1}}\boldsymbol{x}_{i-1} = -\frac{t_i}{t_{i+1}}\boldsymbol{x}_i$$

$$= (-1)^{i+1}\frac{1}{t_{i+1}}R\boldsymbol{v},$$

where the second and third equalities use (1) and (16). Therefore, we have

$$f_2(\boldsymbol{x}_N) - f_2(\boldsymbol{x}_*) = f_2\left((-1)^N\frac{1}{t_N}R\boldsymbol{v}\right) = \frac{LR^2}{2t_N^2},$$

after N iterations of OGM1′ and OGM2′, which is equivalent to the lower bound (27).

□

Since the analytical lower bound (27) matches the numerical tight bound in Table 1, we conjecture that the quadratic function $f_2(\boldsymbol{x})$ is the worst-case function for the secondary sequence of OGM′ and thus (27) is the tight worst-case bound. Whereas FGM has similar worst-case bounds (and behavior as conjectured by Taylor et al. [9, Conjectures 4 and 5]) for both its primary and secondary sequence, the two sequences of OGM′ (or intermediate iterates of OGM) have two different worst-case behaviors, as discussed further in Sect. 5.2.

4.5 Related Work: Nesterov's Accelerated First-Order Method in [10]

Interestingly, an algorithm in [10, Section 4] is similar to OGM2′ and satisfies same convergence bound (20) for the primary sequence $\{\boldsymbol{y}_i\}$, which we call Nes13 in this paper for convenience.[1]

Algorithm Nes13 [10]

Input: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, $t_0 = 1$.

For $i = 0, 1, \ldots$

$$\boldsymbol{y}_{i+1} = \boldsymbol{x}_i - \frac{1}{L}\nabla f(\boldsymbol{x}_i)$$

$$\boldsymbol{z}_{i+1} = \boldsymbol{x}_0 - \frac{1}{L}\sum_{k=0}^{i}2t_k\nabla f(\boldsymbol{y}_{k+1})$$

$$t_{i+1} = \frac{1 + \sqrt{1 + 4t_i^2}}{2}$$

$$\boldsymbol{x}_{i+1} = \left(1 - \frac{1}{t_{i+1}}\right)\boldsymbol{y}_{i+1} + \frac{1}{t_{i+1}}\boldsymbol{z}_{i+1}$$

---

[1] Nes13 was developed originally to deal with nonsmooth composite convex functions with a line-search scheme [10, Section 4], whereas the algorithm shown here is a simplified version of [10, Section 4] for unconstrained smooth convex minimization (M) without a line-search.

The only difference between OGM2$'$ and Nes13 is the gradient used for the update of z_i. While both algorithms achieve same bound (20), Nes13 is less attractive in practice since it requires computing gradients at two different points x_i and y_{i+1} at each ith iteration.

Similar to Proposition 4.1, the following proposition shows that the bound (20) is asymptotically tight for Nes13.

Proposition 4.3 *For the function $f_{1,\mathrm{OGM}'}(x; N)$ (22) in $\mathcal{F}_L(\mathbb{R}^d)$, the iterate y_N generated by Nes13 achieves the lower bound (23).*

Proof See the proof of Proposition 4.1.   □

5 Two Worst-Case Functions for an Optimal Fixed-step GM and OGM

This section discusses two algorithms, an optimal fixed-step GM and OGM, in class FO that have a piecewise affine-quadratic function and a quadratic function as two worst-case functions. Considering that OGM is optimal among first-order methods (for $d \geq N + 1$), it is interesting that OGM has two different types of worst-case functions, because this property resembles the (numerical) analysis of the optimal fixed-step GM in [9] (reviewed below).

5.1 Two Worst-Case Functions for an Optimal Fixed-step GM

The following is GM with a constant step size h.

<div style="border:1px solid">

Algorithm GM

Input: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $x_0 \in \mathbb{R}^d$.

For $i = 0, \ldots, N - 1$

$$x_{i+1} = x_i - \frac{h}{L} \nabla f(x_i)$$

</div>

For GM with $0 < h < 2$, both [2] and [9] conjecture the following tight convergence bound:

$$f(x_N) - f(x_*) \leq \frac{LR^2}{2} \max\left(\frac{1}{2Nh + 1}, (1 - h)^{2N} \right).   \tag{29}$$

The proof of the bound (29) for $0 < h \leq 1$ is given in [2], while the proof for $1 < h < 2$ is still unknown but strong numerical evidence is given in [9]. In other words, at least one of the two functions specified below is conjectured to be a worst-case for GM with a constant step size $0 < h < 2$. Such functions are a piecewise affine-quadratic function

$$f_{1,\mathrm{GM}}(x; h, N) = \begin{cases} \frac{LR}{2Nh+1} \|x\| - \frac{LR^2}{2(2Nh+1)^2}, & \text{if } \|x\| \geq \frac{R}{2Nh+1}, \\ \frac{L}{2} \|x\|^2, & \text{otherwise}, \end{cases}   \tag{30}$$

and a quadratic function $f_2(\boldsymbol{x})$ (26), where $f_{1,\text{GM}}(\boldsymbol{x}; h, N)$ and $f_2(\boldsymbol{x})$ contribute to the factors $\frac{1}{2Nh+1}$ and $(1-h)^{2N}$, respectively, in (29). Here, $f_{1,\text{GM}}(\boldsymbol{x}; h, N)$ is a worst-case function where the GM iterates approach the optimum slowly, whereas $f_2(\boldsymbol{x})$ is a worst-case function where the iterates overshoot the optimum.

Assuming that the above conjecture for a fixed-step GM holds, Taylor et al. [9] searched (numerically) for the optimal fixed-step size $0 < h_{\text{opt}}(N) < 2$ for given N that minimizes the bound (29):

$$h_{\text{opt}}(N) := \underset{0<h<2}{\arg\min} \max \left(\frac{1}{2Nh+1}, (1-h)^{2N} \right). \tag{31}$$

GM with the step $h_{\text{opt}}(N)$ has two worst-case functions $f_{1,\text{GM}}(\boldsymbol{x}; h, N)$ and $f_2(\boldsymbol{x})$, and must compromise between two extreme cases. On the other hand, the case $0 < h < h_{\text{opt}}(N)$ has only $f_{1,\text{GM}}(\boldsymbol{x}; h, N)$ as the worst-case and the case $h_{\text{opt}}(N) < h < 2$ has only $f_2(\boldsymbol{x})$ as the worst-case. We believe this compromise is inherent to optimizing the worst-case performance of FO methods. The next section shows that the optimal OGM also has this desirable property.

For the special case of $N = 1$, the optimal OGM reduces to GM with a fixed-step $h = 1.5$, and this confirms the conjecture in [9] that the step $h_{\text{opt}}(1) = 1.5$ (31) is optimal for a fixed-step GM with $N = 1$. However, proving the optimality of $h_{\text{opt}}(N)$ (31) for the fixed-step GM for $N > 1$ is left as future work.

5.2 Two Worst-Case Functions for the Last Iterate x_N of OGM

[1, Theorem 3] showed that $f_{1,\text{OGM}}(\boldsymbol{x}; N)$ (7) is a worst-case function for the last iterate $\boldsymbol{x}_N$ of OGM. The following theorem shows that a quadratic function $f_2(\boldsymbol{x})$ (26) is also a worst-case function for the last iterate of OGM.

Theorem 5.1 *For the quadratic function $f_2(\boldsymbol{x}) = \frac{L}{2}||\boldsymbol{x}||^2$ (26) in $\mathcal{F}_L(\mathbb{R}^d)$, both OGM1 and OGM2 exactly achieve the convergence bound* (5), *i.e.,*

$$f_2(\boldsymbol{x}_N) - f_2(\boldsymbol{x}_*) = \frac{LR^2}{2\theta_N^2}.$$

Proof We use induction to show that the following iterates:

$$\boldsymbol{x}_i = (-1)^i \frac{1}{\theta_i} R\boldsymbol{v}, \quad i = 0, \dots, N, \tag{32}$$

correspond to the iterates of OGM1 and OGM2 applied to $f_2(\boldsymbol{x})$.

Starting from $\boldsymbol{x}_0 = R\boldsymbol{v}$, where $\boldsymbol{v}$ is a unit vector, and assuming that (32) holds for $i < N$, we have

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \frac{1}{L} \sum_{k=0}^{i} \hat{h}_{i+1,k} \nabla f_2(\boldsymbol{x}_k)$$

$$
\begin{aligned}
&= \left(\boldsymbol{x}_i - \frac{1}{L}\hat{h}_{i+1,i}\nabla f_2(\boldsymbol{x}_i) \right) - \frac{1}{L}\sum_{k=0}^{i-1}\frac{\theta_i - 1}{\theta_{i+1}}\hat{h}_{i,k}\nabla f_2(\boldsymbol{x}_k) + \frac{1}{L}\frac{\theta_i - 1}{\theta_{i+1}}f_2(\boldsymbol{x}_{i-1}) \\
&= \frac{1 - 2\theta_i}{\theta_{i+1}}\boldsymbol{x}_i + \frac{\theta_i - 1}{\theta_{i+1}}(\boldsymbol{x}_i - \boldsymbol{x}_{i-1}) + \frac{\theta_i - 1}{\theta_{i+1}}\boldsymbol{x}_{i-1} = -\frac{\theta_i}{\theta_{i+1}}\boldsymbol{x}_i \\
&= (-1)^{i+1}\frac{1}{\theta_{i+1}}R\boldsymbol{v},
\end{aligned}
$$

where the second and third equalities use (1) and (10). Therefore, we have

$$
f_2(\boldsymbol{x}_N) - f_2(\boldsymbol{x}_*) = f_2\left((-1)^N\frac{1}{\theta_N}R\boldsymbol{v} \right) = \frac{LR^2}{2\theta_N^2}
$$

after N iterations of OGM1 and OGM2, exactly matching the bound (5).   □

Thus the last iterate $\boldsymbol{x}_N$ of OGM has two worst-case functions: $f_{1,\mathrm{OGM}}(\boldsymbol{x}; N)$ and $f_2(\boldsymbol{x})$, similar to an optimal fixed-step GM in Sect. 5.1. In [9, Conjecture 4] and Sect. 4.2, the primary sequence of OGM is conjectured to have $f_{1,\mathrm{OGM}'}(\boldsymbol{x}; N)$ as a worst-case function, whereas the quadratic function $f_2(\boldsymbol{x})$ becomes the best-case as the first primary iterate of OGM reaches the optimum just in one step. On the other hand, Sect. 4.4 conjectured that $f_2(\boldsymbol{x})$ is a worst-case function for the secondary sequence of OGM prior to the last iterate. Apparently the primary and secondary sequences of OGM have two extremely different worst-case analyses, whereas the last iterate $\boldsymbol{x}_N$ of OGM compromises between the two worst-case behaviors, making the worst-case behavior of the optimal OGM interesting.

6 Conclusions

We provided an analytical convergence bound for the primary sequence of OGM1 and OGM2, augmenting the bounds of the last iterate of the secondary sequence of OGM in [1]. The corresponding convergence bound is twice as small as that of Nesterov's FGM, showing that the primary sequence of OGM is faster than FGM. However, interestingly the intermediate iterates of the *secondary* sequence of OGM were found to be slower than FGM in the worst-case.

We proposed two new formulations of OGM, called OGM1′ and OGM2′ that are related closely to Nesterov's accelerated first-order methods in [10] (originally developed for nonsmooth composite convex functions and differing from FGM in [3,7]). For smooth problems, OGM and OGM′ provide faster convergence speed than [10] considering the number of gradient computations required per iteration.

We showed that the last iterate of the secondary sequence of OGM has two types of worst-case functions, a piecewise affine-quadratic function and a quadratic function. In light of the optimality of OGM (for $d \geq N + 1$) in [4], it is interesting that OGM has these two types of worst-case functions. Because the optimal fixed-step GM also appears to have two such worst-case functions, one might conjecture that this behavior is a general characteristic of optimal fixed-step first-order methods.

In addition to the optimality of fixed-step first-order methods for the cost function value, studying the optimality for an alternative criteria such as the gradient ($||\nabla f(\boldsymbol{x}_N)||$) is an interesting research direction. Just as Nesterov's FGM was extended for solving nonsmooth composite convex functions [10,15], it would be interesting to extend OGM to such problems; recently, this was numerically studied by Taylor et al. [13].

References

1. Kim, D., Fessler, J.A.: Optimized first-order methods for smooth convex minimization. Math. Program. **159**(1), 81–107 (2016)
2. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: A novel approach. Math. Program. **145**(1–2), 451–82 (2014)
3. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. Dokl. Akad. Nauk. USSR **269**(3), 543–7 (1983)
4. Drori, Y.: The exact information-based complexity of smooth convex minimization (2016). arXiv:1606.01424
5. Cevher, V., Becker, S., Schmidt, M.: Convex optimization for big data: scalable, randomized, and parallel algorithms for big data analytics. IEEE Sig. Proc. Mag. **31**(5), 32–43 (2014)
6. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Comp. Math. Math. Phys. **4**(5), 1–17 (1964)
7. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
8. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Dordrecht (2004)
9. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Math. Program. (2016). doi:10.1007/s10107-016-1009-3
10. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
11. Drori, Y., Teboulle, M.: An optimal variant of Kelley's cutting-plane method. Math. Program. (2016). doi:10.1007/s10107-016-0985-7
12. Drori, Y.: Contributions to the complexity analysis of optimization algorithms. Ph.D. thesis, Tel-Aviv Univ., Israel (2014)
13. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Exact worst-case performance of first-order algorithms for composite convex optimization (2015). arXiv:1512.07516
14. Lessard, L., Recht, B., Packard, A.: Analysis and design of optimization algorithms via integral quadratic constraints. SIAM J. Optim. **26**(1), 57–95 (2016)
15. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)