Factor- $$\sqrt{2}$$ 2 Acceleration of Accelerated Gradient Methods

Citation: Applied Mathematics & Optimization. 2023 Aug 23;88(3):77

As Published: https://doi.org/10.1007/s00245-023-10047-9

Publisher: Springer US

Persistent URL: https://hdl.handle.net/1721.1/152276

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Factor-$\sqrt{2}$2 Acceleration of Accelerated Gradient Methods

Noname manuscript No.
(will be inserted by the editor)

Factor-$\sqrt{2}$ Acceleration of Accelerated Gradient Methods

Chanwoo Park · Jisun Park · Ernest K. Ryu

Received: date / Accepted: date

Abstract The optimized gradient method (OGM) provides a factor-$\sqrt{2}$ speedup upon Nesterov's celebrated accelerated gradient method in the convex (but non-strongly convex) setup. However, this improved acceleration mechanism has not been well understood; prior analyses of OGM relied on a computer-assisted proof methodology, so the proofs were opaque for humans despite being verifiable and correct. In this work, we present a new analysis of OGM based on a Lyapunov function and linear coupling. These analyses are developed and presented without the assistance of computers and are understandable by humans. Furthermore, we generalize OGM's acceleration mechanism and obtain a factor-$\sqrt{2}$ speedup in other setups: acceleration with a simpler rational stepsize, the strongly convex setup, and the mirror descent setup.

1 Introduction

Nesterov's celebrated accelerated gradient method (AGM) solves the problem of finding the minimum of an L-smooth convex function with an "optimal" accelerated $\mathcal{O}(1/k^2)$ complexity [38]. Surprisingly, AGM turned out to be not exactly optimal, but optimal only up to a constant. The optimized gradient method (OGM) has a factor-2 smaller (better) worst-case guarantee and thereby requires factor-$\sqrt{2}$ fewer iterations to guarantee the same accuracy [22, 26].

Chanwoo Park
Department of Statistics, Seoul National University
E-mail: chanwoo.park@snu.ac.kr

Jisun Park
Department of Mathematical Sciences, Seoul National University
E-mail: colleenp0515@snu.ac.kr

Ernest K. Ryu
Department of Mathematical Sciences, Seoul National University
E-mail: eryu@snu.ac.kr

However, this remarkable discovery has not been well understood. OGM was originally obtained through a computer-assisted methodology based on the performance estimation problem (PEP). The resulting convergence analyses involve arduous but elementary calculations that are verifiable but arguably not understandable by humans.

Contribution. In this work, we present human-understandable analyses of OGM. First, we show that the improved acceleration mechanism of OGM can be understood and analyzed through an unconventional Lyapunov function. We then use this insight to propose a new method that obtains the factor-$\sqrt{2}$ speedup in the strongly convex setup. Finally, we present a human-understandable derivation of OGM based on refining the linear coupling analysis of Allen-Zhu and Orecchia [5], and generalize OGM to the mirror descent setup.

As minor contributions, we analyze the primary and secondary sequences of OGM through a single unified analysis; to the best of our knowledge, prior works provide two separate convergence proofs for x- and y-sequences. Moreover, we present a unified class of accelerated methods containing AGM and OGM through the linear coupling analysis.

1.1 Definitions and prior work

For $L > 0$, a differentiable convex function $f\colon \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to a norm $\|\cdot\|$ if

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \qquad \forall\, x, y \in \mathbb{R}^n,$$

where $\|\cdot\|_*$ denotes the dual norm. A convex function $f\colon \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex if $f(x) - (\mu/2)\|x\|^2$ is convex [39, 47].

Throughout this paper, we consider the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x)$$

and make the following assumptions on $f\colon \mathbb{R}^n \to \mathbb{R}$:

(A1) f is convex, differentiable, and L-smooth with respect to $\|\cdot\|$ and
(A2) f has a minimizer (not necessarily unique).

We write $x_\star$ for a minimizer of f and $f_\star = f(x_\star)$ for the optimal value. To clarify, the proofs of Section 2 do not require the minimizer $x_\star$ to be unique.

Nesterov's AGM. Nesterov's AGM is

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{\theta_k - 1}{\theta_{k+1}}(y_{k+1} - y_k), \end{aligned}$$

where $y_0 = x_0$, $\theta_0 = 1$, and $\theta_{k+1}^2 - \theta_{k+1} = \theta_k^2$ for $k = 0, 1, \ldots$ [38]. We can equivalently write AGM as

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ z_{k+1} &= z_k - \frac{\theta_k}{L}\nabla f(x_k) \\ x_{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1} \end{aligned}$$

with $z_0 = x_0$ [40].

AGM can be generalized to use the relaxed parameter requirement $\theta_{k+1}^2 - \theta_{k+1} \leq \theta_k^2$ on the positive sequence $\{\theta_k\}_{k=0}^{\infty}$. The choice $\theta_k = (k+2)/2$ is a common instance.

In the setup where f is furthermore μ-strongly convex, Nesterov's AGM for the strongly convex setup (SC-AGM) is

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}(y_{k+1} - y_k) \end{aligned}$$

for $k = 0, 1, \ldots$, where $\kappa = L/\mu$ and $y_0 = x_0$ [39].

Optimized gradient method. OGM is

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{\theta_k - 1}{\theta_{k+1}}(y_{k+1} - y_k) + \frac{\theta_k}{\theta_{k+1}}(y_{k+1} - x_k) \end{aligned}$$

for $k = 0, 1, \ldots$, where $y_0 = x_0$ and $\{\theta_k\}_{k=1}^{\infty}$ is the same as that of AGM [22,26]. We refer to $\frac{\theta_k - 1}{\theta_{k+1}}(y_{k+1} - y_k)$ as the *momentum term* and $\frac{\theta_k}{\theta_{k+1}}(y_{k+1} - x_k)$ as the *correction term*. The added correction term is the difference between AGM and OGM. We can equivalently write OGM as

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ z_{k+1} &= z_k - \frac{2\theta_k}{L}\nabla f(x_k) \\ x_{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1}, \end{aligned}$$

where $z_0 = x_0$ [26]. The factor 2 in z_{k+1} is the difference compared to AGM.

The y_k-sequence of OGM exhibits a rate faster than that of AGM by a factor of $\sqrt{2}$. This rate was proved in [27], and we also state it in Corollary 1. To clarify, the guarantee on the function value is smaller (better) by a factor of 2, and, combined with the $\mathcal{O}(1/k^2)$ iteration dependence, this represents

a factor-$\sqrt{2}$ reduction in the number of iterations necessary to reach a given accuracy.

Furthermore, OGM's original presentation [22, 26] involves what we refer to as the *last-step modification* on the secondary sequence

$$\begin{aligned}\tilde{x}_{k+1} &= y_{k+1} + \frac{\theta_k - 1}{\varphi_{k+1}}(y_{k+1} - y_k) + \frac{\theta_k}{\varphi_{k+1}}(y_{k+1} - x_k)\\ &= \left(1 - \frac{1}{\varphi_{k+1}}\right) y_{k+1} + \frac{1}{\varphi_{k+1}} z_{k+1},\end{aligned}$$

where $\varphi_k^2 - \varphi_k - 2\theta_{k-1}^2 = 0$. The $\tilde{x}_k$-sequence of OGM exhibits a rate slightly better than OGM's y_k-sequence and is in fact exactly optimal [19] under the smooth (non-strongly) convex function class. This rate was proved in the original presentation of OGM [22, 26], and we also state it in Corollary 3. In this work, we present the first variant of OGM for the strongly convex setup.

θ_k-sequence asymptotic characterization. Throughout the exposition of this work, we will use the following asymptotic characterization: if $\theta_0 = 1$ and $\theta_{k+1}^2 - \theta_{k+1} = \theta_k^2$ for $k = 0, 1, \ldots$, then

$$\theta_k = \frac{k + \zeta + 1}{2} + \frac{\log k}{4} + o(1) \tag{1}$$

as $k \to \infty$, where $\zeta \approx 0.646$. While we suspect this result may be known, we could not find it in any reference. Therefore, we formally state and prove (1) as Lemma 7 in the appendix.

Computer-assisted derivation and analysis of OGM. OGM was originally obtained through a computer-assisted methodology based on the performance estimation problem (PEP); it was first discovered numerically [22] and then its analytical form and convergence analysis was found [26]. The PEP methodology's key insight is to optimize over the class of fixed-step first-order gradient methods, with the objective being the convergence guarantee. Surprisingly, this problem is semidefinite programming- (SDP-) representable and has a tightness guarantee [54]. OGM was re-discovered by using the PEP to find a greedy first-order method simplified with a "subspace-search elimination procedure" [21].

However, these prior analyses of OGM, generated by computers, are verifiable but arguably not understandable by humans. Moreover, as the analyses rely on finding analytical solutions to the SDPs arising from the PEP, they are inaccessible to those unfamiliar with the methodology.

Lyapunov analysis of AGM. Nesterov's original 1983 paper established the celebrated $\mathcal{O}(1/k^2)$ rate using a Lyapunov analysis [38]. Subsequent works [11, 12, 32, 39–41, 43, 55] analyzed AGM and its variants through the "estimate sequence" technique, which many consider to be less transparent than Lyapunov analyses. In recent years, there has been a renewed interest in studying accelerated methods via Lyapunov analyses [1, 7–9, 13, 16, 50, 52]. In this work, we present the first Lyapunov analysis of OGM.

Linear coupling analysis of AGM. The interpretation of AGM as a *linear coupling* between gradient descent and mirror descent was presented in [5]. Specifically, AGM can be written as

$$\begin{aligned} y_{k+1} &= \arg\min_y \left\{ \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2} \|y - x_k\|^2 \right\} \\ z_{k+1} &= \arg\min_y \{ V_{z_k}(y) + \langle \alpha_{k+1} \nabla f(x_k), y - x_k \rangle \} \\ x_{k+1} &= (1 - \tau_{k+1}) y_{k+1} + \tau_{k+1} z_{k+1}, \end{aligned}$$

where V_z is a Bregman divergence. The y_k-update can be viewed as a gradient descent update and the z_k-update can be viewed as a mirror descent update. Mirror descent [37] was originally presented as a method that maps the current point to a dual space, performs a gradient update, and maps the point back to the primal space. An alternate proximal form of mirror descent (which we use) was presented in [15]. An alternate "dual averaging" interpretation of mirror descent as a method that constructs a lower bound of the function was presented in [42]. The key insight of linear coupling is to carefully interpolate between mirror descent and gradient descent to obtain AGM.

Linear coupling has been used to obtain and analyze many extensions of AGM [2–4, 6], but whether the linear coupling argument itself can be further refined seems not to have been studied. In this work, we show that refining the linear coupling analysis naturally leads to OGM.

Tight inequalities. We informally say an inequality is tight if it cannot be improved without further assumptions and formally if it satisfies the "interpolation conditions" [54]. The recent literature on performance estimation problem focuses on using tight inequalities to obtain proofs that are provably cannot be improved [17, 24, 25, 33, 46, 52, 53].

The tight inequality we use is

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_*^2$$

for all L-smooth convex function f and $x, y \in \mathbb{R}^n$. The linear coupling analysis of AGM uses strictly weaker inequalities, as discussed in Section 4. By refining the analysis by replacing the non-tight inequalities with tight ones, we obtain OGM.

Accelerated methods for smooth strongly convex minimization. For the problem setup of minimizing smooth strongly convex functions, Nesterov's SC-AGM [39] achieves the convergence rate $\mathcal{O}\left(\exp\left(-k/\sqrt{\kappa}\right)\right)$. Recently, the triple momentum method [31] and the information-theoretic exact method [51] were presented with an improved $\mathcal{O}\left(\exp\left(-2k/\sqrt{\kappa}\right)\right)$-rate, and their optimality was established through the matching $\Theta\left(\exp\left(-2k/\sqrt{\kappa}\right)\right)$-lower bound of [20], which improves upon the classical $\Theta\left(\exp\left(-4k/\sqrt{\kappa}\right)\right)$-lower bound of [35, 36]. The SC-OGM method we present in this work has a rate of $\mathcal{O}\left(\exp\left(-\sqrt{2}k/\sqrt{\kappa}\right)\right)$, between

the rates of SC-AGM and TMM. For strongly convex *quadratic* functions, the heavy ball method exhibits the rate $\mathcal{O}\left(\exp\left(-4k/\sqrt{\kappa}\right)\right)$ [39] and OGM-q exhibits the rate $\mathcal{O}\left(\exp\left(-2\sqrt{2}k/\sqrt{\kappa}\right)\right)$ [28]. The heavy ball method's rate matches the classical $\Theta\left(\exp\left(-4k/\sqrt{\kappa}\right)\right)$-lower bound of [35, 36].

2 Lyapunov analysis of OGM

In this section, we present a Lyapunov analysis of OGM. Our key insight is to use

$$\left(f(x_k) - f_\star - \frac{1}{2L}\|\nabla f(x_k)\|^2\right),$$

which is nonnegative due to L-smoothness, instead of $(f(x_k) - f_\star)$ or $(f(y_k) - f_\star)$ in the construction of the Lyapunov function. Throughout this section, $\|\cdot\| = \|\cdot\|_*$ denotes the Euclidean norm.

Based on this insight, we present: (i) a more human-understandable analysis of OGM (ii) a unified analysis of both the primary and secondary sequences of OGM that admits simpler θ_k-choices.

2.1 Nesterov's AGM

Nesterov's AGM has the rate

$$\begin{aligned} f(y_k) - f_\star &\leq \frac{L\|x_0 - x_\star\|^2}{2\theta_{k-1}^2} \\ &= \frac{2L\|x_0 - x_\star\|^2}{(k+\zeta)^2} - \frac{2L\|x_0 - x_\star\|^2 \log k}{(k+\zeta)^3} + o\left(\frac{1}{k^3}\right) \end{aligned}$$

for $k = 0, 1, \ldots$. (We derived the equality in Appendix E.) This rate can be established through the following Lyapunov analysis [38]: for $k = 0, 1, \ldots$, define

$$U_k = \theta_{k-1}^2 (f(y_k) - f_\star) + \frac{L}{2}\|z_k - x_\star\|^2$$

with $\theta_{-1} = 0$ and show $U_k \leq \cdots \leq U_0$. Conclude with

$$\theta_{k-1}^2 (f(y_k) - f_\star) \leq U_k \leq U_0 = \frac{L}{2}\|x_0 - x_\star\|^2.$$

2.2 Primary sequence analysis of OGM

We now analyze OGM's convergence through an analogous Lyapunov analysis.

Theorem 1 *Assume (A1) and (A2). Let the positive sequence $\{\theta_k\}_{k=0}^{\infty}$ satisfy $\theta_0 = 1$ and $0 \leq \theta_{k+1}^2 - \theta_{k+1} \leq \theta_k^2$ for $k = 0, 1, \ldots$. OGM's y_k-sequence exhibits the rate*

$$f(y_k) - f_\star \leq \frac{L \|x_0 - x_\star\|^2}{4\theta_{k-1}^2}$$

for $k = 1, 2, \ldots$.

Proof Set $\theta_{-1} = 0$ and $x_{-1} = x_0$. For $k = -1, 0, 1, \ldots$, define

$$U_k = 2\theta_k^2 \left(f(x_k) - f_\star - \frac{1}{2L} \|\nabla f(x_k)\|^2 \right) + \frac{L}{2} \|z_{k+1} - x_\star\|^2 .$$

We can show that $\{U_k\}_{k=-1}^{\infty}$ is nonincreasing. Using $f(y_k) \leq f(x_{k-1}) - \frac{1}{2L} \|\nabla f(x_{k-1})\|^2$, which follows from L-smoothness, we conclude the rate with

$$\begin{aligned} 2\theta_{k-1}^2 \left(f(y_k) - f_\star \right) &\leq 2\theta_{k-1}^2 \left(f(x_{k-1}) - f_\star - \frac{1}{2L} \|\nabla f(x_{k-1})\|^2 \right) \\ &\leq U_{k-1} \leq U_{-1} = \frac{L}{2} \|z_0 - x_\star\|^2 \end{aligned}$$

for $k = 1, 2, \ldots$. Now we complete the proof by showing that $\{U_k\}_{k=-1}^{\infty}$ is nonincreasing. For $k = -1, 0, 1, \ldots$, we have

$$\begin{aligned} &U_k - U_{k+1} \\ &= 2\theta_k^2 \left(f(x_k) - f_\star - \frac{1}{2L} \|\nabla f(x_k)\|^2 \right) - 2\theta_{k+1}^2 \left(f(x_{k+1}) - f_\star - \frac{1}{2L} \|\nabla f(x_{k+1})\|^2 \right) \\ &\quad + \frac{L}{2} \|z_{k+1} - x_\star\|^2 - \frac{L}{2} \|z_{k+2} - x_\star\|^2 \\ &= 2\theta_k^2 \left(f(x_k) - f_\star - \frac{1}{2L} \|\nabla f(x_k)\|^2 \right) - 2\theta_{k+1}^2 \left(f(x_{k+1}) - f_\star - \frac{1}{2L} \|\nabla f(x_{k+1})\|^2 \right) \\ &\quad - \langle 2\theta_{k+1} \nabla f(x_{k+1}), x_\star - z_{k+1} \rangle - \frac{2}{L} \theta_{k+1}^2 \|\nabla f(x_{k+1})\|^2 \\ &= 2\theta_k^2 \left(f(x_k) - f_\star - \frac{1}{2L} \|\nabla f(x_k)\|^2 \right) - 2\theta_{k+1}^2 \left(f(x_{k+1}) - f_\star + \frac{1}{2L} \|\nabla f(x_{k+1})\|^2 \right) \\ &\quad - \langle 2\theta_{k+1} \nabla f(x_{k+1}), x_\star - z_{k+1} \rangle \\ &\geq 2(\theta_{k+1}^2 - \theta_{k+1}) \left(f(x_k) - f_\star - \frac{1}{2L} \|\nabla f(x_k)\|^2 \right) \\ &\quad - 2\theta_{k+1}^2 \left(f(x_{k+1}) - f_\star + \frac{1}{2L} \|\nabla f(x_{k+1})\|^2 \right) - \langle 2\theta_{k+1} \nabla f(x_{k+1}), x_\star - z_{k+1} \rangle \\ &= 2(\theta_{k+1}^2 - \theta_{k+1}) \left(f(x_k) - f_\star - \frac{1}{2L} \|\nabla f(x_k)\|^2 - f(x_{k+1}) + f_\star - \frac{1}{2L} \|\nabla f(x_{k+1})\|^2 \right) \\ &\quad - 2\theta_{k+1} \left(f(x_{k+1}) - f_\star + \frac{1}{2L} \|\nabla f(x_{k+1})\|^2 \right) - \langle 2\theta_{k+1} \nabla f(x_{k+1}), x_\star - z_{k+1} \rangle \end{aligned}$$

$$
\begin{aligned}
&= 2(\theta_{k+1}^2 - \theta_{k+1})\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2\right) \\
&\quad + 2\theta_{k+1}\left(f_\star - f(x_{k+1}) - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2 + \langle \nabla f(x_{k+1}), x_{k+1} - x_\star\rangle\right) \\
&\quad + 2\theta_{k+1}\langle \nabla f(x_{k+1}), z_{k+1} - x_{k+1}\rangle \\
&\geq 2(\theta_{k+1}^2 - \theta_{k+1})\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2\right) \\
&\quad + 2\theta_{k+1}\langle \nabla f(x_{k+1}), z_{k+1} - x_{k+1}\rangle,
\end{aligned}
$$

where the inequalities follow from the cocoercivity of f.

Consider two separate cases $k = -1$ and $k = 0, 1, \ldots$. In case of $k = -1$, $\theta_{k+1}^2 - \theta_{k+1} = 1 - 1 = 0$ and $z_{k+1} - x_{k+1} = z_0 - x_0 = 0$. The last formula becomes zero, so $U_{-1} - U_0 \geq 0$. For $k = 0, 1, \ldots,$

$$
\begin{aligned}
&2(\theta_{k+1}^2 - \theta_{k+1})\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2\right) \\
&\quad + 2\theta_{k+1}\langle \nabla f(x_{k+1}), z_{k+1} - x_{k+1}\rangle \\
&= 2(\theta_{k+1}^2 - \theta_{k+1})\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2\right) \\
&\quad + 2\theta_{k+1}(\theta_{k+1} - 1)\langle \nabla f(x_{k+1}), x_{k+1} - x_k + \frac{1}{L}\nabla f(x_k)\rangle \\
&= (2\theta_{k+1}^2 - 2\theta_{k+1})\Bigg(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\left\|\nabla f(x_k) - \nabla f(x_{k+1})\right\|^2 \\
&\qquad\qquad + \langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle\Bigg) \geq 0,
\end{aligned}
$$

where the inequalities follow from the cocoercivity of f. □

As with AGM, the optimal $\{\theta_k\}_{k=0}^\infty$ is given by $\theta_{k+1}^2 - \theta_{k+1} = \theta_k^2$, which was used in the original presentation of OGM [22, 26].

Corollary 1 *Under the setup of Theorem 1, the choice $\theta_{k+1}^2 - \theta_{k+1} = \theta_k^2$ leads to the rate*

$$
f(y_k) - f_\star \leq \frac{L\left\|x_0 - x_\star\right\|^2}{4\theta_{k-1}^2} = \frac{L\left\|x_0 - x_\star\right\|^2}{(k+\zeta)^2} - \frac{L\left\|x_0 - x_\star\right\|^2 \log k}{(k+\zeta)^3} + o\left(\frac{1}{k^3}\right)
$$

for $k = 1, 2, \ldots$.

Proof This follows from Theorem 1 and (1). □

The relaxed parameter requirement $0 \leq \theta_{k+1}^2 - \theta_{k+1} \leq \theta_k^2$ of Theorem 1 is reminiscent of the requirement for AGM. We note that [30] had presented a generalized analysis with requirement $\theta_{k+1}^2 \leq \sum_{i=1}^{k+1} \theta_i$ based on the performance estimation problem methodology.

The relaxed parameter requirement allows us to use the simpler rational coefficients $\theta_k = (k+2)/2$. This leads to

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$
$$x_{k+1} = y_{k+1} + \frac{k}{k+3}(y_{k+1} - y_k) + \frac{k+2}{k+3}(y_{k+1} - x_k),$$

which we call *Simple-OGM*.

Corollary 2 *Assume (A1) and (A2). Simple-OGM's y_k-sequence exhibits the rate*

$$f(y_k) - f_\star \leq \frac{L\left\|x_0 - x_\star\right\|^2}{(k+1)^2}$$

for $k = 1, 2, \ldots$.

Proof This follows from Theorem 1. □

2.3 Secondary sequence analysis of OGM

We now analyze the convergence of OGM's secondary sequence with last-step modification through a unified Lyapunov analysis.

Theorem 2 *Assume (A1) and (A2). Let the positive sequence $\{\theta_k\}_{k=0}^\infty$ satisfy $\theta_0 = 1$, and $0 \leq \theta_{k+1}^2 - \theta_{k+1} \leq \theta_k^2$ for $k = 0, 1, \ldots$. Let the positive sequence $\{\varphi_k\}_{k=0}^\infty$ satisfy $0 \leq \varphi_k^2 - \varphi_k \leq 2\theta_{k-1}^2$ for $k = 0, 1, \ldots$, where we define $\theta_{-1} = 0$. OGM's $\tilde{x}_k$-sequence, the secondary sequence with last-step modification, exhibits the rate*

$$f(\tilde{x}_k) - f_\star \leq \frac{L\left\|x_0 - x_\star\right\|^2}{2\varphi_k^2}$$

for $k = 0, 1, \ldots$.

Proof Let $\{U_k\}_{k=-1}^\infty$ be as defined in the proof of the Theorem 1. Define $\{\tilde{U}_k\}_{k=0}^\infty$ as

$$\tilde{U}_k = \varphi_k^2\left(f(\tilde{x}_k) - f_\star\right) + \frac{L}{2}\left\|z_k - \frac{1}{L}\varphi_k\nabla f(\tilde{x}_k) - x_\star\right\|^2.$$

We can show that $\tilde{U}_k \leq U_{k-1}$, we conclude the rate with

$$\varphi_k^2(f(\tilde{x}_k) - f_\star) \leq \tilde{U}_k \leq U_{-1} = \frac{L}{2}\|x_0 - x_\star\|^2$$

for $k = 0, 1, \ldots$. Now we complete the proof by showing that $\tilde{U}_k \leq U_{k-1}$. For $k = 0, 1, \ldots$, we have

$U_{k-1} - \tilde{U}_k$

$$
\begin{aligned}
&= 2\theta_{k-1}^2 \left(f(x_{k-1}) - f_\star - \frac{1}{2L} \|\nabla f(x_{k-1})\|^2 \right) - \varphi_k^2 \left(f(\tilde{x}_k) - f_\star \right) \\
&\quad + \frac{L}{2} \|z_k - x_\star\|^2 - \frac{L}{2} \left\| z_k - \frac{1}{L} \varphi_k \nabla f(\tilde{x}_k) - x_\star \right\|^2 \\
&= 2\theta_{k-1}^2 \left(f(x_{k-1}) - f_\star - \frac{1}{2L} \|\nabla f(x_{k-1})\|^2 \right) - \varphi_k^2 \left(f(\tilde{x}_k) - f_\star \right) \\
&\quad - \langle \varphi_k \nabla f(\tilde{x}_k), x_\star - z_k \rangle - \frac{1}{2L} \varphi_k^2 \|\nabla f(\tilde{x}_k)\|^2 \\
&= 2\theta_{k-1}^2 \left(f(x_{k-1}) - f_\star - \frac{1}{2L} \|\nabla f(x_{k-1})\|^2 \right) \\
&\quad - \varphi_k^2 \left(f(\tilde{x}_k) - f_\star + \frac{1}{2L} \|\nabla f(\tilde{x}_k)\|^2 \right) - \langle \varphi_k \nabla f(\tilde{x}_k), x_\star - z_k \rangle \\
&\geq (\varphi_k^2 - \varphi_k) \left(f(x_{k-1}) - f_\star - \frac{1}{2L} \|\nabla f(x_{k-1})\|^2 \right) \\
&\quad - \varphi_k^2 \left(f(\tilde{x}_k) - f_\star + \frac{1}{2L} \|\nabla f(\tilde{x}_k)\|^2 \right) - \langle \varphi_k \nabla f(\tilde{x}_k), x_\star - z_k \rangle \\
&= (\varphi_k^2 - \varphi_k) \left(f(x_{k-1}) - f_\star - \frac{1}{2L} \|\nabla f(x_{k-1})\|^2 - f(\tilde{x}_k) + f_\star - \frac{1}{2L} \|\nabla f(\tilde{x}_k)\|^2 \right) \\
&\quad + \varphi_k \left(f_\star - f(\tilde{x}_k) - \frac{1}{2L} \|\nabla f(\tilde{x}_k)\|^2 + \langle \nabla f(\tilde{x}_k), \tilde{x}_k - x_\star \rangle \right) \\
&\qquad + \langle \varphi_k \nabla f(\tilde{x}_k), z_k - \tilde{x}_k \rangle \\
&\geq (\varphi_k^2 - \varphi_k) \left(f(x_{k-1}) - f(\tilde{x}_k) - \frac{1}{2L} \|\nabla f(x_{k-1})\|^2 - \frac{1}{2L} \|\nabla f(\tilde{x}_k)\|^2 \right) \\
&\quad + \langle \varphi_k \nabla f(\tilde{x}_k), z_k - \tilde{x}_k \rangle \\
&= (\varphi_k^2 - \varphi_k) \left(f(x_{k-1}) - f(\tilde{x}_k) - \frac{1}{2L} \|\nabla f(x_{k-1})\|^2 - \frac{1}{2L} \|\nabla f(\tilde{x}_k)\|^2 \right) \\
&\quad + \varphi_k(\varphi_k - 1) \langle \nabla f(\tilde{x}_k), \tilde{x}_k - x_{k-1} + \frac{1}{L} \nabla f(x_{k-1}) \rangle \\
&= (\varphi_k^2 - \varphi_k) \left(f(x_{k-1}) - f(\tilde{x}_k) - \frac{1}{2L} \|\nabla f(x_{k-1}) - \nabla f(\tilde{x}_k)\|^2 + \langle \nabla f(\tilde{x}_k), \tilde{x}_k - x_{k-1} \rangle \right) \\
&\geq 0,
\end{aligned}
$$

where the inequalities follow from the cocoercivity of f. □

Corollary 3 *Under the setup of Theorem 2, the choice* $\theta_{k+1}^2 - \theta_{k+1} = \theta_k^2$ *and* $\varphi_k^2 - \varphi_k = 2\theta_{k-1}^2$ *leads to the rate*

$$
f(\tilde{x}_k) - f_\star \leq \frac{L \|x_0 - x_\star\|^2}{2\varphi_k^2} = \frac{L \|x_0 - x_\star\|^2}{(k + \zeta + 1/\sqrt{2})^2} - \frac{L \|x_0 - x_\star\|^2 \log k}{(k + \zeta + 1/\sqrt{2})^3} + o\left(\frac{1}{k^3} \right)
$$

for $k = 0, 1, \ldots$.

Proof This follows from (1), which implies $\varphi_k = \frac{k+\zeta+\frac{1}{\sqrt{2}}}{\sqrt{2}} + \frac{\sqrt{2}\log k}{4} + o(1)$, and Theorem 2. □

Simple-OGM with the last-step modification is

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$
$$x_{k+1} = y_{k+1} + \frac{k}{k+3}(y_{k+1} - y_k) + \frac{k+2}{k+3}(y_{k+1} - x_k)$$
$$\tilde{x}_{k+1} = y_{k+1} + \frac{k}{\sqrt{2}(k+2)+1}(y_{k+1} - y_k) + \frac{k+2}{\sqrt{2}(k+2)+1}(y_{k+1} - x_k),$$

where $x_0 = y_0$.

Corollary 4 *Assume (A1) and (A2). Simple-OGM's $\tilde{x}_k$-sequence, the secondary sequence with last-step modification, exhibits the rate*

$$f(\tilde{x}_k) - f_\star \leq \frac{L\left\|x_0 - x_\star\right\|^2}{(k+1+1/\sqrt{2})^2}$$

for $k = 0, 1, \ldots$.

Proof Use Corollary 3 with $\theta_k = \frac{k+2}{2}$ and $\varphi_k = \frac{k+1+\frac{1}{\sqrt{2}}}{\sqrt{2}}$. □

2.4 Discussion

We clarify that the presented Lyapunov analysis is a novel contribution, while the results themselves are mostly known [26, 27, 30].

We emphasize two key points. First is the somewhat unusual construction of the Lyapunov function. This key insight will be used in the following section to present a novel method for the strongly convex setup.

The second point we emphasize is that we present a unified analysis of the primary and last-step-modified secondary sequences using the Lyapunov functions U_k and $\tilde{U}_k$. Prior works on the two sequences of AGM and OGM rely on two separate analyses [26, 27].

3 Strongly convex OGM

In this section, we present strongly convex OGM (SC-OGM), a novel method that provides a factor-$\sqrt{2}$ improvement over Nesterov's SC-AGM. The method and its analysis are obtained with following the key insight of Section 2: use the OGM-type correction term in the method and use

$$\left(f(x_k) - f_\star - \frac{1}{2L}\|\nabla f(x_k)\|^2\right)$$

in the construction of the Lyapunov function. Throughout this section, $\|\cdot\| = \|\cdot\|_*$ denotes the Euclidean norm.

Based on this insight, we present: (i) a novel method SC-OGM and (ii) a unified analysis of both the primary and secondary sequences of SC-OGM.

3.1 Nesterov's SC-AGM

Further assume f is μ-strongly convex and write $\kappa = L/\mu$. SC-AGM's convergence rate

$$f(y_k) - f_\star \leq \left(1 + \frac{1}{\sqrt{\kappa} - 1}\right)^{-k} \frac{\mu + L}{2} \|x_0 - x_\star\|^2 = \mathcal{O}\left(\exp\left(-\frac{k}{\sqrt{\kappa}}\right)\right)$$

can be established through the following Lyapunov analysis [13]. For $k = 0, 1, \ldots,$ define

$$U_k = \left(1 + \frac{1}{\sqrt{\kappa} - 1}\right)^{k} \left(f(y_k) - f_\star + \frac{\mu}{2} \|z_k - x_\star\|^2\right)$$

with $z_k = (\sqrt{\kappa} + 1)x_k - \sqrt{\kappa} y_k$ and show $U_k \leq \cdots \leq U_0 \leq \frac{\mu+L}{2} \|x_0 - x_\star\|^2$.

3.2 Primary-sequence analysis of SC-OGM

We newly propose SC-OGM:

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{1}{2\gamma + 1}(y_{k+1} - y_k) + \frac{1}{2\gamma + 1}(y_{k+1} - x_k) \end{aligned}$$

for $k = 0, 1, \ldots,$ where $y_0 = x_0$ and $\gamma = \frac{\sqrt{8\kappa+1}+3}{2\kappa-2}$.

Theorem 3 *Assume (A1), (A2), and that f is μ-strongly convex. SC-OGM's y_k-sequence exhibits the rate*

$$f(y_k) - f_\star \leq (1 + \gamma)^{-k+1} \frac{\mu + 2L}{2} \|x_0 - x_\star\|^2 = \mathcal{O}\left(\exp\left(-\frac{\sqrt{2}k}{\sqrt{\kappa}}\right)\right)$$

for $k = 1, 2, \ldots.$

Proof For $k = 0, 1, \ldots,$ define

$$z_k = \frac{2\gamma + 1}{\gamma} x_k - \frac{\gamma + 1}{\gamma} y_k$$

and

$$U_k = (1 + \gamma)^k \left(f(x_k) - f_\star - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{\mu}{2} \|z_{k+1} - x_\star\|^2\right).$$

We can show that $\{U_k\}_{k=0}^{\infty}$ is nonincreasing and $U_0 \leq \frac{\mu+2L}{2}\|x_0 - x_\star\|^2$. Using $f(y_k) \leq f(x_{k-1}) - \frac{1}{2L}\|\nabla f(x_{k-1})\|^2$, which follows from L-smoothness, we conclude the rate with

$$\begin{aligned}(1+\gamma)^{k-1}(f(y_k) - f_\star) &\leq (1+\gamma)^{k-1}\left(f(x_{k-1}) - f_\star - \frac{1}{2L}\|\nabla f(x_{k-1})\|^2\right)\\ &\leq U_{k-1} \leq U_0 \leq \frac{\mu+2L}{2}\|x_0 - x_\star\|^2\end{aligned}$$

for $k = 1, 2, \ldots$. Now we complete the proof by showing $U_0 \leq \frac{\mu+2L}{2}\|x_0 - x_\star\|^2$, showing some relationships between x_k and z_k, and showing that $\{U_k\}_{k=0}^{\infty}$ is nonincreasing.

Firstly, we have

$$\begin{aligned}U_0 &= f(x_0) - f_\star - \frac{1}{2L}\|\nabla f(x_0)\|^2 + \frac{\mu}{2}\|z_1 - x_\star\|^2\\ &= f(x_0) - f_\star - \frac{1}{2L}\|\nabla f(x_0)\|^2 + \frac{\mu}{2}\left\|x_0 - \frac{1}{L}\frac{\gamma+2}{\gamma}\nabla f(x_0) - x_\star\right\|^2\\ &= f(x_0) - f_\star + \frac{1}{2L}\frac{1}{\gamma+1}\|\nabla f(x_0)\|^2 - \frac{\gamma}{1+\gamma}\langle\nabla f(x_0), x_0 - x_\star\rangle + \frac{\mu}{2}\|x_0 - x_\star\|^2\\ &\leq \frac{1}{\gamma+1}(f(x_0) - f_\star) + \frac{1}{2L}\frac{1}{1+\gamma}\|\nabla f(x_0)\|^2 + \frac{\mu}{2}\|x_0 - x_\star\|^2\\ &\leq \frac{2}{1+\gamma}(f(x_0) - f_\star) + \frac{\mu}{2}\|x_0 - x_\star\|^2\\ &\leq \left(L + \frac{\mu}{2}\right)\|x_0 - x_\star\|^2.\end{aligned}$$

Second, Let $X_k = x_k - x_\star$ and $Z_k = z_k - x_\star$, for $k = 0, 1, \ldots$. We will prove

$$(x_{k+1} - x_k) + \frac{1}{L}\nabla f(x_k) + \gamma X_{k+1} = \frac{1}{1+\gamma}(\gamma Z_{k+1} + \gamma^2 X_{k+1}) \tag{2}$$

$$Z_{k+1} = \frac{1}{\gamma+1}Z_k + \frac{\gamma}{\gamma+1}X_k - \frac{1}{L}\frac{\gamma+2}{\gamma}\nabla f(x_k) \tag{3}$$

for $k = 0, 1, \ldots$.

Plug $y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ in the definition of z_{k+1}. (We remind the reader that z_k was defined in the beginning of the proof.) Then we obtain (2).
For (3), from definition of z_k and z_{k+1}

$$\begin{aligned}z_{k+1} &= \frac{2\gamma+1}{\gamma}x_{k+1} - \frac{\gamma+1}{\gamma}x_k + \frac{1}{L}\frac{1+\gamma}{\gamma}\nabla f(x_k)\\ z_k &= \frac{2\gamma+1}{\gamma}x_k - \frac{\gamma+1}{\gamma}x_{k-1} + \frac{1}{L}\frac{1+\gamma}{\gamma}\nabla f(x_{k-1})\end{aligned}$$

and definition of x_k, we have

$$\begin{aligned} x_{k+1} &= \frac{2\gamma+2}{2\gamma+1}y_{k+1} - \frac{1}{2\gamma+1}y_k - \frac{1}{L}\frac{1}{2\gamma+1}\nabla f(x_k) \\ &= \frac{2\gamma+2}{2\gamma+1}x_k - \frac{1}{2\gamma+1}x_{k-1} - \frac{1}{L}\frac{2\gamma+3}{2\gamma+1}\nabla f(x_k) + \frac{1}{L}\frac{1}{2\gamma+1}\nabla f(x_{k-1}). \end{aligned}$$

Therefore,

$$\begin{aligned} z_{k+1} - \frac{1}{\gamma+1}z_k &= \frac{2\gamma+1}{\gamma}x_{k+1} - \frac{\gamma+1}{\gamma}x_k + \frac{1}{L}\frac{1+\gamma}{\gamma}\nabla f(x_k) \\ &\quad - \frac{1}{\gamma+1}\left(\frac{2\gamma+1}{\gamma}x_k - \frac{\gamma+1}{\gamma}x_{k-1} + \frac{1}{L}\frac{1+\gamma}{\gamma}\nabla f(x_{k-1})\right) \\ &= \frac{2\gamma+1}{\gamma}x_{k+1} - \frac{\gamma^2+4\gamma+2}{\gamma(\gamma+1)}x_k + \frac{1}{\gamma}x_{k-1} + \frac{1}{L}\frac{1+\gamma}{\gamma}\nabla f(x_k) \\ &\quad - \frac{1}{L}\frac{1}{\gamma}\nabla f(x_{k-1}) \\ &= \frac{2\gamma+1}{\gamma}\left(\frac{2\gamma+2}{2\gamma+1}x_k - \frac{1}{2\gamma+1}x_{k-1} - \frac{1}{L}\frac{2\gamma+3}{2\gamma+1}\nabla f(x_k)\right. \\ &\quad \left. + \frac{1}{L}\frac{1}{2\gamma+1}\nabla f(x_{k-1})\right) - \frac{\gamma^2+4\gamma+2}{\gamma(\gamma+1)}x_k + \frac{1}{\gamma}x_{k-1} \\ &\quad + \frac{1}{L}\frac{1+\gamma}{\gamma}\nabla f(x_k) - \frac{1}{L}\frac{1}{\gamma}\nabla f(x_{k-1}) \\ &= \frac{\gamma}{\gamma+1}x_k - \frac{1}{L}\frac{\gamma+2}{\gamma}\nabla f(x_k) \end{aligned}$$

so we obtained (3).

Lastly, we will show that $\{U_k\}_{k=0}^{\infty}$ is nonincreasing. It suffices to show that for $k = 0, 1, \ldots,$

$$(1+\gamma)^{-k}(U_k - U_{k+1}) \geq 0$$

which is equivalent to showing

$$\begin{aligned} &\left((f(x_k) - f_\star - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2) - (1+\gamma)(f(x_{k+1}) - f_\star - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2)\right) \\ &\qquad + \frac{\mu}{2}\left(\left\|z_{k+1} - x_\star\right\|^2 - (1+\gamma)\left\|z_{k+2} - x_\star\right\|^2\right) \geq 0. \end{aligned}$$

By L-smoothness of f, we have

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 + \langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle$$

and from strong convexity,

$$f(x_{k+1}) - f_\star \leq \langle \nabla f(x_{k+1}), x_{k+1} - x_\star\rangle - \frac{\mu}{2}\left\|x_{k+1} - x_\star\right\|^2.$$

For $k = 0, 1, \ldots,$ using above two inequalities, (2), and (3),

$$
\begin{aligned}
&\left(f(x_k) - f_\star - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2\right) - (1+\gamma)(f(x_{k+1}) - f_\star - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2) \\
&\quad = (f(x_k) - f(x_{k+1})) - \gamma(f(x_{k+1}) - f_\star) + \frac{1+\gamma}{2L}\left\|\nabla f(x_{k+1})\right\|^2 - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 \\
&\quad \geq \left(\frac{1}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle\right) \\
&\qquad - \gamma\left(\langle \nabla f(x_{k+1}), x_{k+1} - x_\star\rangle - \frac{\mu}{2}\left\|x_{k+1} - x_\star\right\|^2\right) \\
&\qquad\quad + \frac{1+\gamma}{2L}\left\|\nabla f(x_{k+1})\right\|^2 - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 \\
&\quad = \langle \nabla f(x_{k+1}), -\frac{1}{L}\nabla f(x_k) - x_{k+1} + x_k - \gamma(x_{k+1} - x_\star)\rangle \\
&\qquad + \frac{2+\gamma}{2L}\left\|\nabla f(x_{k+1})\right\|^2 + \frac{\mu\gamma}{2}\left\|x_{k+1} - x_\star\right\|^2 \\
&\quad = \langle \nabla f(x_{k+1}), -\frac{1}{1+\gamma}(\gamma Z_{k+1} + \gamma^2 X_{k+1})\rangle \\
&\qquad + \frac{2+\gamma}{2L}\left\|\nabla f(x_{k+1})\right\|^2 + \frac{\mu\gamma}{2}\left\|x_{k+1} - x_\star\right\|^2.
\end{aligned}
$$

In addition,

$$
\begin{aligned}
&\frac{\mu}{2}\left((1+\gamma)\left\|Z_{k+2}\right\|^2 - \left\|Z_{k+1}\right\|^2\right) \\
&\quad = \frac{\mu}{2}\left((1+\gamma)\left\|\frac{1}{1+\gamma}Z_{k+1} + \frac{\gamma}{1+\gamma}X_{k+1} - \frac{1}{L}\frac{2+\gamma}{\gamma}\nabla f(x_{k+1})\right\|^2 - \left\|Z_{k+1}\right\|^2\right) \\
&\quad = \frac{\mu}{2}\Bigg(-\frac{\gamma}{1+\gamma}\left\|Z_{k+1}\right\|^2 + \frac{\gamma^2}{1+\gamma}\left\|X_{k+1}\right\|^2 + (1+\gamma)\frac{1}{L^2}\frac{(2+\gamma)^2}{\gamma^2}\left\|\nabla f(x_{k+1})\right\|^2 \\
&\qquad + 2\frac{\gamma}{1+\gamma}\langle Z_{k+1}, X_{k+1}\rangle - 2\frac{2+\gamma}{L\gamma}\langle \nabla f(x_{k+1}), Z_{k+1}\rangle \\
&\qquad\quad - 2\frac{2+\gamma}{L}\langle \nabla f(x_{k+1}), X_{k+1}\rangle\Bigg).
\end{aligned}
$$

Since

$$
\mu\frac{2+\gamma}{L\gamma^2} = \frac{1}{1+\gamma},
$$

we can telescope concerned $\nabla f(x_{k+1})$'s inner product in $U_k - U_{k+1}$.

For $k = 0, 1, \ldots,$ we have

$$\begin{aligned}
&(1+\gamma)^{-k}(U_k - U_{k+1}) \\
&\quad \geq \frac{2+\gamma}{2L}\left\|\nabla f(x_{k+1})\right\|^2 + \frac{\mu\gamma}{2}\left\|X_{k+1}\right\|^2 \\
&\qquad - \frac{\mu}{2}\Bigg(-\frac{\gamma}{1+\gamma}\left\|Z_{k+1}\right\|^2 + \frac{\gamma^2}{1+\gamma}\left\|X_{k+1}\right\|^2 \\
&\qquad\quad + (1+\gamma)\frac{1}{L^2}\frac{(2+\gamma)^2}{\gamma^2}\left\|\nabla f(x_{k+1})\right\|^2 + 2\frac{\gamma}{1+\gamma}\langle Z_{k+1}, X_{k+1}\rangle\Bigg) \\
&\quad = -\frac{\mu}{2}\left(-\frac{\gamma}{1+\gamma}\left\|X_{k+1}\right\|^2 - \frac{\gamma}{1+\gamma}\left\|Z_{k+1}\right\|^2 + 2\frac{\gamma}{1+\gamma}\langle Z_{k+1}, X_{k+1}\rangle\right) \\
&\quad = \frac{\mu}{2}\frac{\gamma}{1+\gamma}\left\|Z_{k+1} - X_{k+1}\right\|^2 \geq 0.
\end{aligned}$$

□

3.3 Secondary sequence analysis

We now analyze the convergence of SC-OGM's secondary sequence with a unified Lyapunov analysis. We note that SC-OGM does not require the last-step modification, unlike the non-strongly convex counterpart.

Theorem 4 *Assume (A1), (A2), and that f is μ-strongly convex. SC-OGM's x_k-sequence, the secondary sequence without last-step modification, exhibits the rate*

$$f(x_k) - f_\star \leq \frac{(1+\gamma)^{-k+2}}{2\gamma}\left(\frac{\mu+2L}{2}\left\|x_0 - x_\star\right\|^2\right)$$

for $k = 1, 2, \ldots.$

Proof Let $\{z_k\}_{k=0}^\infty$ and $\{U_k\}_{k=0}^\infty$ be defined as in the proof of the Theorem 3. For $k = 0, 1, \ldots,$ define

$$\tilde{U}_k = (1+\gamma)^{k-1}\left(\frac{2\gamma}{1+\gamma}(f(x_k) - f_\star) + \frac{\mu}{2}\left\|z_k - \left(\frac{\gamma+2}{\gamma}\right)\frac{1}{L}\nabla f(x_k) - x_\star\right\|^2\right)$$

We can show that $\tilde{U}_k \leq U_{k-1}$. We conclude the rate with

$$(1+\gamma)^{k-1}\frac{2\gamma}{1+\gamma}\left(f(x_k) - f_\star\right) \leq \tilde{U}_k \leq U_0 \leq \frac{\mu+2L}{2}\left\|x_0 - x_\star\right\|^2$$

for $k = 1, 2, \ldots.$ Now we complete the proof by showing that $\tilde{U}_k \leq U_{k-1}$. Note that $\frac{\gamma+1}{\gamma}\left((x_k - x_{k-1}) + \frac{1}{L}\nabla f(x_{k-1})\right) = (Z_k - X_k)$. Then we have

$$\left(f(x_{k-1}) - f_\star - \frac{1}{2L}\left\|\nabla f(x_{k-1})\right\|^2\right) - \frac{2\gamma}{1+\gamma}\left(f(x_k) - f_\star\right)$$

$$
\begin{aligned}
&+ \frac{L\gamma^2}{2(1+\gamma)(2+\gamma)} \left\|z_k - x_\star\right\|^2 - \frac{L\gamma^2}{2(1+\gamma)(2+\gamma)} \left\|z_k - \left(\frac{\gamma+2}{\gamma}\right)\frac{1}{L}\nabla f(x_k) - x_\star\right\|^2 \\
&= \left(f(x_{k-1}) - f_\star - \frac{1}{2L}\left\|\nabla f(x_{k-1})\right\|^2\right) - \frac{2\gamma}{1+\gamma}\left(f(x_k) - f_\star\right) \\
&\quad + \frac{\gamma}{1+\gamma}\langle Z_k, \nabla f(x_k)\rangle - \frac{1}{2L}\frac{2+\gamma}{1+\gamma}\left\|\nabla f(x_k)\right\|^2 \\
&= \left(f(x_{k-1}) - f_\star - \frac{1}{2L}\left\|\nabla f(x_{k-1})\right\|^2\right) - \frac{2\gamma}{1+\gamma}\left(f(x_k) - f_\star\right) \\
&\quad + \frac{\gamma}{1+\gamma}\left\langle \frac{\gamma+1}{\gamma}\left((x_k - x_{k-1}) + \frac{1}{L}\nabla f(x_{k-1})\right) + X_k, \nabla f(x_k)\right\rangle \\
&\qquad - \frac{1}{2L}\frac{2+\gamma}{1+\gamma}\left\|\nabla f(x_k)\right\|^2 \\
&= \left(f(x_{k-1}) - f_\star - \frac{1}{2L}\left\|\nabla f(x_{k-1})\right\|^2\right) - \frac{2\gamma}{1+\gamma}\left(f(x_k) - f_\star\right) \\
&\quad + \langle x_k - x_{k-1}, \nabla f(x_k)\rangle + \frac{1}{L}\langle \nabla f(x_{k-1}), \nabla f(x_k)\rangle + \frac{\gamma}{1+\gamma}\langle X_k, \nabla f(x_k)\rangle \\
&\qquad - \frac{1}{2L}\frac{2+\gamma}{1+\gamma}\left\|\nabla f(x_k)\right\|^2 \\
&= \left(f(x_{k-1}) - f(x_k) - \frac{1}{2L}\|\nabla f(x_{k-1}) - \nabla f(x_k)\|^2 + \langle \nabla f(x_k), x_k - x_{k-1}\rangle\right) \\
&\quad + \frac{1}{2L}\frac{\gamma}{1+\gamma}\left\|\nabla f(x_k)\right\|^2 + \frac{1}{1+\gamma}\left(f(x_k) - f_\star - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2\right) \\
&\qquad + \frac{\gamma}{1+\gamma}\left(f_\star - f(x_k) - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 + \langle X_k, \nabla f(x_k)\rangle\right) \\
&\geq 0.
\end{aligned}
$$

Since $\frac{L\gamma^2}{2(1+\gamma)(2+\gamma)} = \frac{\mu}{2}$, above inequality indicates that

$$
\begin{aligned}
&\left(f(x_{k-1}) - f_\star - \frac{1}{2L}\left\|\nabla f(x_{k-1})\right\|^2\right) + \frac{\mu}{2}\left\|z_k - x_\star\right\|^2 \\
&\qquad \geq \frac{2\gamma}{1+\gamma}(f(x_k) - f_\star) + \frac{\mu}{2}\left\|z_k - \left(\frac{\gamma+2}{\gamma}\right)\frac{1}{L}\nabla f(x_k) - x_\star\right\|^2.
\end{aligned}
$$

□

3.4 Discussion

The factor-$\sqrt{2}$ improvement of SC-OGM over SC-AGM is consistent with the factor-$\sqrt{2}$ improvement of OGM over AGM. AGM and OGM share the same momentum term while OGM has the additional "correction term". In contrast, the momentum coefficients differ in the strongly convex case: SC-AGM has

$$
\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 1 - \frac{2}{\sqrt{\kappa}} + \mathcal{O}\left(\frac{1}{\kappa}\right)
$$

while SC-OGM has

$$\frac{1}{2\gamma+1} = 1 - \frac{2\sqrt{2}}{\sqrt{\kappa}} + \mathcal{O}\left(\frac{1}{\kappa}\right).$$

Of course, SC-OGM also has the correction term, which is essential in the analysis. We clarify that SC-OGM is not an optimal algorithm for the set of minimizing smooth strongly convex functions as discussed in Section 1.1.

Another interesting line of research is to extend the faster rates to the composite minimization setup, which minimize $f + g$ with a smooth strongly convex f and convex but possibly non-smooth g, as has been pursued in [49] and [10]. Interestingly, the algorithm of [10, Theorem 6] is different from SC-OGM, but achieves the same $\mathcal{O}\left(\exp\left(-\sqrt{2}k/\sqrt{\kappa}\right)\right)$-rate as SC-OGM, while having an extension to the composite minimization setup.

4 Linear coupling analysis

While the Lyapunov analyses of Sections 2 and 3 do provide insight into the acceleration mechanism of OGM, they do not shed light onto the *provenance* of the method. Originally, OGM was generated through a computer-assisted proof methodology as the exactly optimal first-order method, but this approach is arguably opaque to humans.

In this section, we present a human-understandable *derivation* of OGM based on linear coupling. Specifically, we obtain OGM by refining the linear coupling analysis of Allen-Zhu and Orecchia [5] through replacing the use of non-tight inequalities with tight inequalities.

We specifically provide: (i) a natural (and non-computer assisted) derivation of OGM, (ii) a generalization of OGM to the mirror descent setup, and (iii) a unification of AGM and OGM. We moreover provide (iv) a generalization of SC-OGM to the mirror descent setup in the appendix, in Section D.

Assumption and notation. In this section, assume

(A3) $\|\cdot\| = \sqrt{x^T Q x}$ is a quadratic norm, where Q is a symmetric positive definite matrix.

Assumption (A1) is to be interpreted as L-smoothness with respect to norm $\|\cdot\|$. Write $\|\cdot\|_* = x^T Q^{-1} x$ for the dual norm of $\|\cdot\|$. However, $\langle\cdot,\cdot\rangle$ is the standard Euclidean inner product (unrelated to Q). Let $w\colon \mathbb{R}^n \to \mathbb{R}$ be a "distance generating function" that is differentiable and 1-strongly convex with respect to $\|\cdot\|$, and let

$$V_x(y) = w(y) - \langle \nabla w(x), y - x\rangle - w(x) \qquad \forall x, y \in \mathbb{R}^n$$

be the Bregman divergence generated by w.

4.1 Linear coupling analysis of AGM

We briefly outline the linear coupling analysis of AGM presented in [5] and point out where the analysis can be refined.

Consider the problem of minimizing f under assumptions (A1), (A2), and (A3). The linear coupling method is

$$\begin{aligned} y_{k+1} &= x_k - L^{-1}Q^{-1}\nabla f(x_k) \\ z_{k+1} &= \operatorname*{arg\,min}_{y\in\mathbb{R}^n} \{V_{z_k}(y) + \langle \alpha_{k+1}\nabla f(x_k), y - x_k\rangle\} \\ x_{k+1} &= (1-\tau_{k+1})y_{k+1} + \tau_{k+1}z_{k+1} \end{aligned} \tag{LC}$$

for $k = 0, 1, \ldots$, where $x_0 = z_0$ and $\{\alpha_k\}_{k=1}^\infty$ and $\{\tau_k\}_{k=1}^\infty$ are positive sequences to be determined.

We obtain AGM by performing a non-tight analysis of (LC) and letting the analysis inform the choices of $\{\alpha_k\}_{k=1}^\infty$ and $\{\tau_k\}_{k=1}^\infty$. The first step of this analysis is

$$\begin{aligned} \alpha_{k+1}\langle \nabla f(x_k), z_k - x_\star\rangle &\le \frac{\alpha_{k+1}^2}{2}\|\nabla f(x_k)\|_*^2 + V_{z_k}(x_\star) - V_{z_{k+1}}(x_\star) \\ &\le \alpha_{k+1}^2 L(f(x_k) - f(y_{k+1})) + V_{z_k}(x_\star) - V_{z_{k+1}}(x_\star). \end{aligned}$$

The second inequality follows from

$$f(x_k) - f(y_{k+1}) \ge \frac{1}{2L}\|\nabla f(x_k)\|_*^2 + \underline{\frac{1}{2L}\|\nabla f(y_{k+1})\|_*^2},$$

but the underscored term $\frac{1}{2L}\|\nabla f(y_{k+1})\|_*^2$ is not used, i.e., proof utilizes the weaker and non-tight inequality

$$f(x_k) - f(y_{k+1}) \ge \frac{1}{2L}\|\nabla f(x_k)\|_*^2.$$

The second step of this analysis is to choose $\tau_k = \frac{1}{\alpha_{k+1}L}$ to eliminate $f(x_k)$ and to show

$$\alpha_{k+1}^2 L\big(f(y_{k+1}) - f_\star\big) + V_{z_{k+1}}(x_\star) \le \big(\alpha_{k+1}^2 L - \alpha_{k+1}\big)\big(f(y_k) - f_\star\big) + V_{z_k}(x_\star).$$

The inequality follows from

$$f(x_k) - f_\star \le \langle \nabla f(x_k), x_k - x_\star\rangle - \underline{\frac{1}{2L}\|\nabla f(x_k)\|_*^2}$$

and

$$\langle \nabla f(x_k), y_k - x_k\rangle \le f(y_k) - f(x_k) - \underline{\frac{1}{2L}\|\nabla f(y_k) - \nabla f(x_k)\|_*^2},$$

but the underscored terms are not used. Finally, convergence is established through a telescoping sum argument as Appendix C.

4.2 Linear coupling analysis of OGM

We now derive OGM through performing a tight analysis of (LC) and letting the analysis inform the choices of $\{\alpha_k\}_{k=1}^{\infty}$ and $\{\tau_k\}_{k=0}^{\infty}$.

In the first step of our linear coupling analysis, we follow the same arguments but do not take the step utilizing the non-tight inequality.

Lemma 1 *Assume (A1) and (A2). The iterates* (LC) *satisfy*

$$\alpha_{k+1}\langle \nabla f(x_k), z_k - x_\star\rangle \leq \frac{\alpha_{k+1}^2}{2}\|\nabla f(x_k)\|_*^2 + V_{z_k}(x_\star) - V_{z_{k+1}}(x_\star)$$

for $k = 0, 1, \dots$.

Proof This is exactly the first part of Lemma 4.2 of [5]. □

In the second step of our linear coupling analysis, we choose $\tau_k = \frac{2}{\alpha_{k+1}L}$ to allow for a telescoping sum argument and show the following lemma.

Lemma 2 *Assume (A1), (A2) and (A3). Let* $0 < \tau_k = \frac{2}{\alpha_{k+1}L} \leq 1$ *for* $k = 0, 1, ..,$ $\alpha_1 = \frac{2}{L}$, *and* $x_{-1} = x_0$. *Set* $h(x) = f(x) - f_\star - \frac{1}{2L}\|\nabla f(x)\|_*^2$. *The iterates* (LC) *satisfy*

$$\frac{\alpha_{k+1}^2 L}{2}h(x_k) + V_{z_{k+1}}(x_\star) \leq \frac{\alpha_{k+1}^2 L - 2\alpha_{k+1}}{2}h(x_{k-1}) + V_{z_k}(x_\star)$$

for $k = 0, 1, \dots$.

Proof For $k = 1, 2, \dots$, we have

$$\begin{aligned}
&\alpha_{k+1}\left(f(x_k) - f_\star\right)) \\
&\leq \alpha_{k+1}\langle \nabla f(x_k), x_k - x_\star\rangle - \frac{\alpha_{k+1}}{2L}\|\nabla f(x_k)\|_*^2 && (4)\\
&= \alpha_{k+1}\langle \nabla f(x_k), x_k - z_k\rangle + \alpha_{k+1}\langle \nabla f(x_k), z_k - x_\star\rangle - \frac{\alpha_{k+1}}{2L}\|\nabla f(x_k)\|_*^2 \\
&= \frac{1-\tau_k}{\tau_k}\alpha_{k+1}\langle \nabla f(x_k), y_k - x_k\rangle + \alpha_{k+1}\langle \nabla f(x_k), z_k - x_\star\rangle - \frac{\alpha_{k+1}}{2L}\|\nabla f(x_k)\|_*^2 \\
&= \frac{1-\tau_k}{\tau_k}\alpha_{k+1}\langle \nabla f(x_k), x_{k-1} - x_k - \frac{1}{L}Q^{-1}\nabla f(x_{k-1})\rangle \\
&\quad + \alpha_{k+1}\langle \nabla f(x_k), z_k - x_\star\rangle - \frac{\alpha_{k+1}}{2L}\|\nabla f(x_k)\|_*^2 && (5)\\
&\leq \frac{1-\tau_k}{\tau_k}\alpha_{k+1}\left(f(x_{k-1}) - f(x_k) - \frac{1}{2L}\|\nabla f(x_{k-1})\|_*^2 - \frac{1}{2L}\|\nabla f(x_k)\|_*^2\right) && (6)\\
&\qquad + \alpha_{k+1}\langle \nabla f(x_k), z_k - x_\star\rangle - \frac{\alpha_{k+1}}{2L}\|\nabla f(x_k)\|_*^2 \\
&\leq \frac{1-\tau_k}{\tau_k}\alpha_{k+1}\left(f(x_{k-1}) - f(x_k) - \frac{1}{2L}\|\nabla f(x_{k-1})\|_*^2 - \frac{1}{2L}\|\nabla f(x_k)\|_*^2\right) && (7)\\
&\qquad + \frac{\alpha_{k+1}^2}{2}\|\nabla f(x_k)\|_*^2 + V_{z_k}(x_\star) - V_{z_{k+1}}(x_\star) - \frac{\alpha_{k+1}}{2L}\|\nabla f(x_k)\|_*^2.
\end{aligned}$$

(4) and (6) follow from Lemma 11, (5) follows from the definition of linear coupling, and (7) follows from Lemma 1.

The case of $k = 0$ follows from $\alpha_1 = \frac{2}{L}$ and $f_\star - f(x_0) - \langle \nabla f(x_0), x_\star - x_0 \rangle - \frac{1}{2L} \|\nabla f(x_0)\|_*^2 \geq 0$ with Lemma 1. □

Theorem 5 *Assume (A1), (A2), and (A3). Let the positive sequence* $\{\alpha_k\}_{k=1}^\infty$ *satisfy* $0 \leq \alpha_{k+1}^2 L - 2\alpha_{k+1} \leq \alpha_k^2 L$ *for* $k = 1, 2 \ldots$ *and* $\alpha_1 = \frac{2}{L}$. *Let* $\tau_k = \frac{2}{\alpha_{k+1} L}$ *for* $k = 1, 2, \ldots$. *The* y_k*-sequence of* (LC) *exhibits the rate*

$$f(y_k) - f_\star \leq \frac{2V_{x_0}(x_\star)}{L\alpha_k^2}$$

for $k = 1, 2, \ldots$.

Proof Sum the inequality of Lemma 2 from 0 to $(k-1)$. Then use $V_{z_k}(x_\star) \geq 0$ and $f(y_k) \leq f(x_{k-1}) - \frac{1}{2L} \|\nabla f(x_{k-1})\|_*^2$ to conclude the rate. □

The $\{\theta_k\}_{k=0}^\infty$ of the original OGM formulation is related to $\{\alpha_k\}_{k=1}^\infty$ through $\alpha_{k+1} = 2\theta_k/L$ for $k = 0, 1, \ldots$. The seemingly different parameter choices $\tau_k = \frac{1}{\alpha_{k+1} L}$ for AGM and $\tau_k = \frac{2}{\alpha_{k+1} L}$ for OGM actually turn out to be the same as $\{\alpha_k\}_{k=1}^\infty$ for AGM and OGM differ by a factor of 2.

The parameters $\{\alpha_k\}_{k=1}^\infty$ and $\{\tau_k\}_{k=1}^\infty$ are chosen to make the telescoping sum argument work and to make it work tightly, as described in Section C. Specifically, one starts with the form

$$M_k \left(f(x_k) - f_\star - \frac{1}{2L} \|\nabla f(x_k)\|_*^2 \right) + V_{z_{k+1}}(x_\star) \\ \leq N_{k-1} \left(f(x_{k-1}) - f_\star - \frac{1}{2L} \|\nabla f(x_{k-1})\|_*^2 \right) + V_{z_k}(x_\star),$$

where the scalar coefficients M_k, N_{k-1} are determined by (7). Comparing the coefficients of $\|\nabla f(x_k)\|_*^2$, we have

$$-\frac{1}{2L} \left(\alpha_{k+1} + \frac{1 - \tau_k}{\tau_k} \alpha_{k+1} \right) = -\frac{\alpha_{k+1}^2}{2} + \frac{1}{2L} \left(\alpha_{k+1} + \frac{1 - \tau_k}{\tau_k} \alpha_{k+1} \right).$$

Solving this equation leads to the choice $\tau_k = \frac{2}{L\alpha_{k+1}}$. The requirement $\alpha_{k+1}^2 L - 2\alpha_{k+1} \leq \alpha_k^2 L$ is needed for the telescoping sum argument to work, and the choice $\alpha_{k+1}^2 L - 2\alpha_{k+1} = \alpha_k^2 L$ makes the argument tight.

4.3 Secondary sequence analysis

In the linear coupling context, the last-step modification can be expressed as

$$\tilde{x}_k = (1 - \tilde{\tau}_k) y_k + \tilde{\tau}_k z_k \tag{8}$$

for $k = 0, 1, \ldots$, where $\{\tilde{\tau}_k\}_{k=0}^\infty$ is a positive sequence to be determined.

Lemma 3 *Assume (A1), (A2) and (A3). Let* $0 < \tilde{\tau}_k = \frac{1}{\tilde{\alpha}_{k+1}L} \leq 1$ *for* $k = 0, 1, \ldots$, $\tilde{\alpha}_1 = \frac{1}{L}$, *and* $x_{-1} = x_0$. *Then the* $\tilde{x}_k$*-sequence of* (8)*, the secondary sequence with last-step modification of* (LC)*, satisfies*

$$\tilde{\alpha}_{k+1}^2 L \left(f(\tilde{x}_k) - f_\star\right) + V_{z_{k+1}}(x_\star) \leq \left(\tilde{\alpha}_{k+1}^2 L - \tilde{\alpha}_{k+1}\right) h(x_{k-1}) + V_{z_k}(x_\star)$$

for $k = 0, 1, \ldots$.

Proof Proof is identical to that of Lemma 2 with substituted τ_k by $\tilde{\tau}_k$. □

Theorem 6 *In the setup of Theorem 5, let* $0 \leq \tilde{\alpha}_{k+1}^2 L - \tilde{\alpha}_{k+1} \leq \frac{1}{2}\alpha_k^2 L$ *and* $\tilde{\alpha}_1 = \frac{1}{L}$. *Then the* $\tilde{x}_k$*-sequence, the secondary sequence with last-step modification, of the linear coupling method* (LC) *exhibits the rate*

$$f(\tilde{x}_k) - f_\star \leq \frac{V_{x_0}(x_\star)}{L\tilde{\alpha}_{k+1}^2}$$

for $k = 0, 1, \ldots$

Proof Sum the inequality of Lemma 2 from 0 to $(k-2)$ and the inequality of Lemma 3 with $k-1$. Then use $V_{z_k}(x_\star) \geq 0$ to conclude the rate. □

4.4 Comparison of the linear coupling analyses of AGM and OGM

The linear coupling analysis of Allen-Zhu and Orecchia [5], which derives AGM, relies on the following two key lemmas.

Lemma 4 [5, Lemma 4.2] *In the linear coupling setup,*

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(x_k), z_k - x_\star\rangle &\leq \frac{\alpha_{k+1}^2}{2}\left\|\nabla f(x_k)\right\|_*^2 + V_{z_k}(x_\star) - V_{z_{k+1}}(x_\star)\\ &\leq \alpha_{k+1}^2 L\left(f(x_k) - f(y_{k+1})\right) + V_{z_k}(x_\star) - V_{z_{k+1}}(x_\star)\end{aligned}$$

for $k = 0, 1, \ldots$.

Lemma 5 [5, Lemma 4.3] (Coupling Lemma) *In the linear coupling setup,*

$$\alpha_{k+1}^2 L\left(f(y_{k+1}) - f_\star\right) + V_{z_{k+1}}(x_\star) \leq \left(\alpha_{k+1}^2 L - \alpha_{k+1}\right)\left(f(y_k) - f_\star\right) + V_{z_k}(x_\star).$$

for $k = 0, 1, \ldots$.

As discussed, the proof of [5, Lemma 4.2] uses of the non-tight inequality

$$f(x_k) - f(y_{k+1}) \geq \frac{1}{2L}\left\|\nabla f(x_k)\right\|_*^2,$$

and the proof of [5, Lemma 4.3] follows steps similar to that of Lemma 2, but uses the non-tight inequalities

$$f(x_k) - f_\star \leq \langle \nabla f(x_k), x_{k+1} - x_\star\rangle$$

and

$$\langle \nabla f(x_k), y_k - x_k \rangle \leq f(y_k) - f(x_k).$$

In both linear coupling analyses, for OGM and AGM, the telescoping sum argument is made tight by choosing $\{\alpha_k\}_{k=1}^{\infty}$ and $\{\tau_k\}_{k=1}^{\infty}$ appropriately. However, the analysis of Allen-Zhu and Orecchia [5] uses non-tight inequalities before the telescoping sum argument, while our analysis uses tight inequalities in all steps.

4.5 Unification of AGM and OGM

If we choose $w(y) = \frac{1}{2t}\|y\|^2$, so that $V_x(y) = \frac{1}{2t}\|x-y\|^2$, and $0 < t \leq 1$, so that w is 1-strongly convex, and substitute $\alpha_{k+1} = 2\theta_k/L$, (LC) becomes

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ z_{k+1} &= z_k - \frac{2t\theta_k}{L}\nabla f(x_k) \\ x_{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1} \end{aligned}$$

for $k = 0, 1, \ldots$. We also express this method with the momentum and correction terms and without the z^k-iterates in Lemma 6. This method unifies AGM and OGM through the constant t; AGM and OGM respectively correspond to $t = (1/2)$ and $t = 1$.

Corollary 5 *Assume (A1), (A2) and (A3). Let $0 < t \leq 1$. Then*

$$f(y_k) - f_\star \leq \frac{L\|x_0 - x_\star\|^2}{4t\theta_{k-1}^2}$$

for $k = 1, 2, \ldots$

Proof This follows from Theorem 5 with $\alpha_{k+1} = \frac{2\theta_k}{L}$. □

The rates of Corollary 5 at $t = \frac{1}{2}$ and $t = 1$ exactly match the previously discussed rates of AGM and OGM.

Lemma 6 *The unified form is equivalent to*

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{\theta_k - 1}{\theta_{k+1}}(y_{k+1} - y_k) + (2t-1)\frac{\theta_k}{\theta_{k+1}}(y_{k+1} - x_k). \end{aligned}$$

Proof To prove the equivalency, we show that the above sequence leads to

$$x_{k+1} = \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1}.$$

That is,

$$\begin{aligned}
x_{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{\theta_k}{\theta_{k+1}} y_{k+1} - \frac{\theta_k - 1}{\theta_{k+1}} y_k - (2t-1)\frac{\theta_k}{\theta_{k+1}}\frac{1}{L}\nabla f(x_k) \\
&= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{\theta_k}{\theta_{k+1}}\left(x_k - \frac{1}{L}\nabla f(x_k)\right) \\
&\quad - \frac{\theta_k - 1}{\theta_{k+1}} y_k - (2t-1)\frac{\theta_k}{\theta_{k+1}}\frac{1}{L}\nabla f(x_k) \\
&= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{\theta_k}{\theta_{k+1}} x_k - \frac{\theta_k - 1}{\theta_{k+1}} y_k - 2t\frac{\theta_k}{\theta_{k+1}}\frac{1}{L}\nabla f(x_k) \\
&= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} - \frac{\theta_k - 1}{\theta_{k+1}} y_k - 2t\frac{\theta_k}{\theta_{k+1}}\frac{1}{L}\nabla f(x_k) \\
&\quad + \frac{\theta_k}{\theta_{k+1}}\left(y_k + \frac{\theta_{k-1} - 1}{\theta_k}(y_k - y_{k-1}) - (2t-1)\frac{\theta_{k-1}}{\theta_k}\frac{1}{L}\nabla f(x_{k-1})\right) \\
&= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \left(\frac{\theta_k}{\theta_{k+1}} + \frac{\theta_{k-1} - 1}{\theta_{k+1}} - \frac{\theta_k - 1}{\theta_{k+1}}\right) y_k - \frac{\theta_{k-1} - 1}{\theta_{k+1}} y_{k-1} \\
&\quad - (2t-1)\frac{\theta_{k-1}}{\theta_{k+1}}\frac{1}{L}\nabla f(x_{k-1}) - 2t\frac{\theta_k}{\theta_{k+1}}\frac{1}{L}\nabla f(x_k) \\
&= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{\theta_{k-1}}{\theta_{k+1}} y_k - \frac{\theta_{k-1} - 1}{\theta_{k+1}} y_{k-1} \\
&\quad - (2t-1)\frac{\theta_{k-1}}{\theta_{k+1}}\frac{1}{L}\nabla f(x_{k-1}) - 2t\frac{\theta_k}{\theta_{k+1}}\frac{1}{L}\nabla f(x_k) \\
&= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{\theta_{k-1}}{\theta_{k+1}}\left(x_{k-1} - \frac{1}{L}\nabla f(x_{k-1})\right) - \frac{\theta_{k-1} - 1}{\theta_{k+1}} y_{k-1} \\
&\quad - (2t-1)\frac{\theta_{k-1}}{\theta_{k+1}}\frac{1}{L}\nabla f(x_{k-1}) - 2t\frac{\theta_k}{\theta_{k+1}}\frac{1}{L}\nabla f(x_k) \\
&= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{\theta_{k-1}}{\theta_{k+1}} x_{k-1} - \frac{\theta_{k-1} - 1}{\theta_{k+1}} y_{k-1} \\
&\quad - 2t\frac{\theta_k}{\theta_{k+1}}\frac{1}{L}\nabla f(x_k) - 2t\frac{\theta_{k-1}}{\theta_{k+1}}\frac{1}{L}\nabla f(x_{k-1}) \\
&\quad \vdots \\
&= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{\theta_0}{\theta_{k+1}} x_0 - \frac{\theta_0 - 1}{\theta_{k+1}} y_0 - \frac{1}{\theta_{k+1}} \sum_{i=0}^{k} 2t\theta_i \frac{1}{L}\nabla f(x_i) \\
&= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1}.
\end{aligned}$$

□

4.6 Discussion

By identifying OGM as an instance of linear coupling, we generalized OGM to the setup with quadratic norms and mirror descent steps while maintaining the factor-$\sqrt{2}$ improvement. However, we do point out that the generalization is narrower than that of [5], which allows non-quadratic norms and constrained y_k-and z_k-updates. The analysis on strongly convex case follows from a similar line of reasoning, and is presented in Appendix, Section D.

In addition to the human-understandable derivation of OGM, this section provides two non-obvious observations, which we point out again. The first is that AGM and OGM can be unified into a single parameterized family of accelerated gradient methods, all achieving the $\mathcal{O}(1/k^2)$ rate. Another is that the linear coupling analysis of Allen-Zhu and Orecchia [5] was suboptimal in the same way that AGM is suboptimal and can be improved.

5 Conclusion

In this work, we presented human-understandable analyses of OGM. The first key insight is to use a Lyapunov function with $f(x_k) - f_\star - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2$, a somewhat unusual term in Lyapunov analyses. The second key insight is to obtain OGM by refining the linear coupling analysis of Allen-Zhu and Orecchia [5] through replacing non-tight inequalities with tight ones. With these insights, we extended the factor-$\sqrt{2}$ acceleration to other setups.

In our view, the most significant contribution of this work is the improved understanding of OGM's acceleration mechanism. While Nesterov's acceleration mechanism has been utilized as a component in a wide range of setups, OGM's acceleration mechanism has not yet seen any external use. Through the understanding provided by the analysis of this work, we hope OGM's acceleration becomes more widely utilized to gain a (perhaps factor-$\sqrt{2}$) speedup compared to what can be achieved with AGM's acceleration. For example, whether accelerated coordinate gradient methods [6, 44] or non-convex stochastic optimization [23] can be improved with OGM's acceleration mechanism would be an interesting question to address in future work. Improving the FISTA [16] and the more general mirror descent setup [14,34] are also interesting directions, although there are known limitations [18, 29].

Finally, studying how OGM's acceleration interacts with other techniques used to analyze AGM, such as the continuous-time analysis [50], high-resolution ODEs [48], and variational perspective [55] is also an interesting direction.

Acknowledgements

JP and EKR were supported by the Samsung Science and Technology Foundation (Project Number SSTF-BA2101-02) and the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) [NRF-2022R1C1C1010010]. We thank Gyumin Roh for reviewing the manuscript and

providing valuable feedback. We thank Bryan Van Scoy and Suvrit Sra for the discussions regarding the triple momentum method and estimate sequences, respectively.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Ahn, K., Sra, S.: From Nesterov's estimate sequence to Riemannian acceleration. COLT (2020)
2. Allen-Zhu, Z.: Katyusha: The first direct acceleration of stochastic gradient methods. STOC (2017)
3. Allen-Zhu, Z., Hazan, E.: Variance reduction for faster non-convex optimization. ICML (2016)
4. Allen-Zhu, Z., Lee, Y.T., Orecchia, L.: Using optimization to obtain a width-independent, parallel, simpler, and faster positive SDP solver. SODA (2016)
5. Allen-Zhu, Z., Orecchia, L.: Linear coupling: An ultimate unification of gradient and mirror descent. ITCS (2017)
6. Allen-Zhu, Z., Qu, Z., Richtárik, P., Yuan, Y.: Even faster accelerated coordinate descent using non-uniform sampling. ICML (2016)
7. Aujol, J., Dossal, C.: Optimal rate of convergence of an ODE associated to the fast gradient descent schemes for $b > 0$. HAL Archives Ouvertes (2017)
8. Aujol, J.F., Dossal, C., Fort, G., Moulines, É.: Rates of convergence of perturbed FISTA-based algorithms. HAL Archives Ouvertes (2019)
9. Aujol, J.F., Dossal, C., Rondepierre, A.: Optimal convergence rates for Nesterov acceleration. SIAM Journal on Optimization **29**(4), 3131–3153 (2019)
10. Aujol, J.F., Dossal, C., Rondepierre, A.: Convergence rates of the heavy-ball method for quasi-strongly convex optimization (2021)
11. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization. SIAM Journal on Optimization **16**(3), 697–725 (2006)
12. Baes, M.: Estimate sequence methods: extensions and approximations. Tech. rep., Institute for Operations Research, ETH, Zürich, Switzerland (2009)
13. Bansal, N., Gupta, A.: Potential-function proofs for gradient methods. Theory of Computing **15**(4), 1–32 (2019)
14. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. Mathematics of Operations Research **42**(2), 330–348 (2017)
15. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters **31**(3), 167–175 (2003)
16. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009)
17. De Klerk, E., Glineur, F., Taylor, A.B.: Worst-case convergence analysis of inexact gradient and newton methods through semidefinite programming performance estimation. SIAM Journal on Optimization **30**(3), 2053–2082 (2020)
18. Dragomir, R.A., Taylor, A.B., d'Aspremont, A., Bolte, J.: Optimal complexity and certification of Bregman first-order methods. Mathematical Programming (2021)
19. Drori, Y.: The exact information-based complexity of smooth convex minimization. Journal of Complexity **39**, 1–16 (2017)
20. Drori, Y., Taylor, A.: On the oracle complexity of smooth strongly convex minimization. Journal of Complexity **68**, 101590 (2022)
21. Drori, Y., Taylor, A.B.: Efficient first-order methods for convex minimization: a constructive approach. Mathematical Programming **184**(1), 183–220 (2020)
22. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: a novel approach. Mathematical Programming **145**(1-2), 451–482 (2014)
23. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Mathematical Programming **156**(1-2), 59–99 (2016)
24. Gu, G., Yang, J.: Tight sublinear convergence rate of the proximal point algorithm for maximal monotone inclusion problems. SIAM Journal on Optimization **30**(3), 1905–1921 (2020)

25. Kim, D.: Accelerated proximal point method for maximally monotone operators. Mathematical Programming (2021)
26. Kim, D., Fessler, J.A.: Optimized first-order methods for smooth convex minimization. Mathematical Programming **159**(1-2), 81–107 (2016)
27. Kim, D., Fessler, J.A.: On the convergence analysis of the optimized gradient method. Journal of Optimization Theory and Applications **172**(1), 187–205 (2017)
28. Kim, D., Fessler, J.A.: Adaptive restart of the optimized gradient method for convex optimization. Journal of Optimization Theory and Applications **178**(1), 240–263 (2018)
29. Kim, D., Fessler, J.A.: Another look at the fast iterative shrinkage/thresholding algorithm (FISTA). SIAM Journal on Optimization **28**(1), 223–250 (2018)
30. Kim, D., Fessler, J.A.: Generalizing the optimized gradient method for smooth convex minimization. SIAM Journal on Optimization **28**(2), 1920–1950 (2018)
31. Lessard, L., Recht, B., Packard, A.: Analysis and design of optimization algorithms via integral quadratic constraints. SIAM Journal on Optimization **26**(1), 57–95 (2016)
32. Li, B., Coutiño, M., Giannakis, G.B.: Revisit of estimate sequence for accelerated gradient methods. ICASSP (2020)
33. Lieder, F.: On the convergence rate of the halpern-iteration. Optimization Letters pp. 1–14 (2020)
34. Lu, H., Freund, R.M., Nesterov, Y.: Relatively smooth convex optimization by first-order methods, and applications. SIAM Journal on Optimization **28**(1), 333–354 (2018)
35. Nemirovsky, A.S.: On optimality of Krylov's information when solving linear operator equations. Journal of Complexity **7**(2), 121–130 (1991)
36. Nemirovsky, A.S.: Information-based complexity of linear operator equations. Journal of Complexity **8**(2), 153–175 (1992)
37. Nemirovsky, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. (1983)
38. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. Proceedings of the USSR Academy of Sciences **269**, 543–547 (1983)
39. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course (2004)
40. Nesterov, Y.: Smooth minimization of non-smooth functions. Mathematical Programming **103**(1), 127–152 (2005)
41. Nesterov, Y.: Accelerating the cubic regularization of Newton's method on convex problems. Mathematical Programming **112**(1), 159–181 (2008)
42. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Mathematical Programming **120**(1), 221–259 (2009)
43. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization **22**(2), 341–362 (2012)
44. Nesterov, Y., Stich, S.U.: Efficiency of the accelerated coordinate descent method on structured optimization problems. SIAM Journal on Optimization **27**(1), 110–123 (2017)
45. Rockafellar, R.T.: Convex Analysis (1970)
46. Ryu, E.K., Taylor, A.B., Bergeling, C., Giselsson, P.: Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. SIAM Journal on Optimization **30**(3), 2251–2271 (2020)
47. Ryu, E.K., Yin, W.: Large-Scale Convex Optimization via Monotone Operators. Draft (2021)
48. Shi, B., Du, S.S., Su, W., Jordan, M.I.: Acceleration via symplectic discretization of high-resolution differential equations. NeurIPS (2019)
49. Siegel, J.W.: Accelerated first-order methods: Differential equations and lyapunov functions. arXiv preprint arXiv:1903.05671 (2019)
50. Su, W., Boyd, S., Candes, E.: A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. NeurIPS (2014)
51. Taylor, A., Drori, Y.: An optimal gradient method for smooth strongly convex minimization. Mathematical Programming (2022)
52. Taylor, A.B., Bach, F.: Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. COLT (2019)
53. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Exact worst-case performance of first-order methods for composite convex optimization. SIAM Journal on Optimization **27**(3), 1283–1313 (2017)

54. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Mathematical Programming **161**(1-2), 307–345 (2017)
55. Wibisono, A., Wilson, A.C., Jordan, M.I.: A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences **113**(47), E7351–E7358 (2016)

A Method reference

For reference, we restate all aforementioned methods. In all methods, we assume that f is L-smooth function, $\{\theta_k\}_{k=0}^{\infty}$ and $\{\varphi_k\}_{k=0}^{\infty}$ are the sequences of positive scalars, and $x_0 = y_0 = z_0$.

OGM. One form of OGM is

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{\theta_k - 1}{\theta_{k+1}}(y_{k+1} - y_k) + \frac{\theta_k}{\theta_{k+1}}(y_{k+1} - x_k) \end{aligned}$$

and an equivalent form with z-iterates is

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ z_{k+1} &= z_k - \frac{2\theta_k}{L}\nabla f(x_k) \\ x_{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1} \end{aligned}$$

for $k = 0, 1, \ldots$. The *last-step modification* on the secondary sequence can be written as

$$\begin{aligned} \tilde{x}_{k+1} &= y_{k+1} + \frac{\theta_k - 1}{\varphi_{k+1}}(y_{k+1} - y_k) + \frac{\theta_k}{\varphi_{k+1}}(y_{k+1} - x_k) \\ &= \left(1 - \frac{1}{\varphi_{k+1}}\right) y_{k+1} + \frac{1}{\varphi_{k+1}} z_{k+1} \end{aligned}$$

where $k = 0, 1, \ldots$.

OGM-simple. OGM-simple is a simpler variant of **OGM** with $\theta_k = \frac{k+2}{2}$ and $\varphi_k = \frac{k+1+\frac{1}{\sqrt{2}}}{\sqrt{2}}$. One form of OGM-simple is

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{k}{k+3}(y_{k+1} - y_k) + \frac{k+2}{k+3}(y_{k+1} - x_k) \end{aligned}$$

and an equivalent form with z-iterates is

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ z_{k+1} &= z_k - \frac{k+2}{L}\nabla f(x_k) \\ x_{k+1} &= \left(1 - \frac{2}{k+3}\right) y_{k+1} + \frac{2}{k+3} z_{k+1} \end{aligned}$$

for $k = 0, 1, \ldots$. The *last-step modification* on secondary sequence is written as

$$\tilde{x}_{k+1} = y_{k+1} + \frac{k}{\sqrt{2}(k+2)+1}(y_{k+1} - y_k) + \frac{k+2}{\sqrt{2}(k+2)+1}(y_{k+1} - x_k)$$

where $k = 0, 1, \ldots$.

SC-OGM. Here, we assume that f is a μ-strongly convex function, condition number of f is $\kappa = L/\mu$, and $\gamma = \frac{\sqrt{8\kappa+1}+3}{2\kappa-2}$. SC-OGM is written as

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{1}{2\gamma+1}(y_{k+1} - y_k) + \frac{1}{2\gamma+1}(y_{k+1} - x_k) \end{aligned}$$

for $k = 0, 1, \ldots$.

LC-OGM. LC-OGM (Linear Coupling OGM) is defined as

$$\begin{aligned} y_{k+1} &= x_k - L^{-1}Q^{-1}\nabla f(x_k) \\ z_{k+1} &= \operatorname*{arg\,min}_{y\in\mathbb{R}^n}\{V_{z_k}(y) + \langle \alpha_{k+1}\nabla f(x_k), y - x_k\rangle\} \\ x_{k+1} &= (1-\tau_{k+1})y_{k+1} + \tau_{k+1}z_{k+1} \end{aligned}$$

for $k = 0, 1, \ldots$, where $V_z(y)$ is a Bregman divergence, $\{\alpha_k\}_{k=1}^\infty$ and $\{\tau_k\}_{k=1}^\infty$ are nonnegative sequences defined as $\alpha_1 = \frac{2}{L}$, $0 \le \alpha_{k+1}^2 L - 2\alpha_{k+1} \le \alpha_k^2 L$, $\tau_k = \frac{2}{\alpha_{k+1}L}$, and Q is a positive definite matrix defining $\|x\|^2 = x^T Q x$.

For *last step modification*, we define positive sequences $\{\tilde{\alpha}_k\}_{k=1}^\infty$ and $\{\tilde{\tau}_k\}_{k=1}^\infty$ as $\alpha_1 = \frac{1}{L}$, $0 \le \tilde{\alpha}_{k+1}^2 L - \tilde{\alpha_{k+1}} \le \frac{1}{2}\alpha_k^2 L$, and $\tilde{\tau}_k = \frac{1}{\tilde{\alpha}_{k+1}L}$, and also define

$$\tilde{x}_k = (1-\tilde{\tau}_k)y_k + \tilde{\tau}_k z_k$$

for $k = 1, 2, \ldots$.

Unification of AGM and OGM. Using LC-OGM, we can unify AGM and OGM as

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ z_{k+1} &= z_k - \frac{2t\theta_k}{L}\nabla f(x_k) \\ x_{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1}. \end{aligned}$$

for $k = 0, 1, \ldots$. This is equivalent to

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{\theta_k - 1}{\theta_{k+1}}(y_{k+1} - y_k) + (2t-1)\frac{\theta_k}{\theta_{k+1}}(y_{k+1} - x_k). \end{aligned}$$

LC-SC-OGM. LC-SC-OGM (Linear Coupling Strongly Convex OGM) is

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}Q^{-1}\nabla f(x_k) \\ z_{k+1} &= \frac{1}{1+\gamma}\left(z_k + \gamma x_k - \frac{\gamma}{\mu}Q^{-1}\nabla f(x_k)\right) \\ x_{k+1} &= \tau z_{k+1} + (1-\tau)y_{k+1}, \end{aligned}$$

for $k = 0, 1, \ldots$, where Q is a positive definite matrix.

B Co-coercivity inequality in general norm

Lemma 7 *Let f be a closed convex proper function. Then,*

$$0 \leq f(x) + f^*(u) - \langle x, u\rangle$$

and

$$\inf_x \{f(x) + f^*(u) - \langle x, u\rangle\} = 0$$
$$\inf_u \{f(x) + f^*(u) - \langle x, u\rangle\} = 0.$$

Proof By the definition of the conjugate function,

$$-f^*(u) = \inf_x \{f(x) - \langle x, u\rangle\}$$

and

$$\inf_x \{f(x) + f^*(u) - \langle x, u\rangle\} = 0.$$

Therefore,

$$0 \leq f(x) + f^*(u) - \langle x, u\rangle \quad \forall x.$$

The statement with u follows from the same argument and the fact that $f^{**} = f$. □

Lemma 8 *Consider a norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$. Then,*

$$0 \leq \frac{1}{2}\|x\|^2 + \frac{1}{2}\|u\|_*^2 - \langle x, u\rangle$$

and

$$\inf_{x\in\mathbb{R}^n}\left\{\frac{1}{2}\|x\|^2 + \frac{1}{2}\|u\|_*^2 - \langle x, u\rangle\right\} = 0$$
$$\inf_{u\in\mathbb{R}^n}\left\{\frac{1}{2}\|x\|^2 + \frac{1}{2}\|u\|_*^2 - \langle x, u\rangle\right\} = 0.$$

Proof This follows from Lemma 7 with $f(x) = \frac{1}{2}\|x\|^2$ and $\left(\frac{1}{2}\|\cdot\|^2\right)^* = \frac{1}{2}\|\cdot\|_*^2$. □

Lemma 9 *Let*

$$\texttt{Grad}(x) = \arg\min_{y\in\mathbb{R}^n}\left\{\frac{L}{2}\|y - x\|^2 + \langle\nabla f(x), y - x\rangle\right\}.$$

Then,

$$\langle\nabla f(x), \texttt{Grad}(x) - x\rangle + \frac{L}{2}\|\texttt{Grad}(x) - x\|^2 = -\frac{1}{2L}\|\nabla f(x)\|_*^2.$$

Proof Let $z = L(\texttt{Grad}(x) - x)$. By the definition of $\texttt{Grad}(x)$ and Lemma 8, we have

$$\begin{aligned}\frac{1}{2L}\|\nabla f(x)\|_*^2 + \frac{L}{2}\|\texttt{Grad}(x) - x\|^2 + \langle\nabla f(x), \texttt{Grad}(x) - x\rangle&\\ = \inf_{z\in\mathbb{R}^n}\frac{1}{2L}\|\nabla f(x)\|_*^2 + \frac{1}{2L}\|z\|^2 + \frac{1}{L}\langle\nabla f(x), z\rangle&\\ = 0.&\end{aligned}$$

□

Lemma 10 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function such that*

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|$$

for all $x, y \in \mathbb{R}^n$. Then

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

Proof Since a differentiable convex function is continuously differentiable [45, Theorem 25.5],

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \langle \nabla f(x + t(y-x)), y - x \rangle dt \\ &= \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle dt + \langle \nabla f(x), y - x \rangle \\ &\leq \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\|_* \|y - x\| \, dt + \langle \nabla f(x), y - x \rangle \\ &\leq \int_0^1 tL \|y - x\|^2 \, dt + \langle \nabla f(x), y - x \rangle = \frac{L}{2} \|y - x\|^2 + \langle \nabla f(x), y - x \rangle. \end{aligned}$$

□

Lemma 11 (Co-coercivity inequality with general norm) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function such that*

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\|$$

for all $x, y \in \mathbb{R}^n$. Then

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_*^2 .$$

Proof Set $\phi(y) = f(y) - \langle \nabla f(x), y - x \rangle$. Then $x \in \arg\min \phi$. So by Lemma 9,

$$\begin{aligned} \phi(x) &\leq \phi(\texttt{Grad}(y)) \\ &\leq \phi(y) + \langle \nabla \phi(y), \texttt{Grad}(y) - y \rangle + \frac{L}{2} \|\texttt{Grad}(y) - y\|^2 \\ &= \phi(y) - \frac{1}{2L} \|\nabla \phi(y)\|_*^2 . \end{aligned}$$

Substituting f back in ϕ yields the co-coercivity inequality. □

C Telescoping sum argument

Suppose we established the inequality

$$a_i F_i + b_i G_i \leq c_i F_{i-1} + d_i G_{i-1} - E_i$$

for $i = 1, 2, \ldots$, where E_i, F_i, G_i are nonnegative quantities and a_i, b_i, c_i, and d_i are nonnegative scalars. Assume $c_i \leq a_{i-1}$ and $d_i \leq b_{i-1}$. By summing the inequalities for $i = 1, 2, \ldots, k$, we obtain

$$\begin{aligned} a_k F_k &\leq -b_k G_k - \sum_{i=2}^{k} (a_{i-1} - c_i) F_{i-1} - \sum_{i=2}^{k} (b_{i-1} - d_i) G_{i-1} - \sum_{i=2}^{k} E_i + c_1 F_0 + d_1 G_0 \\ &\leq c_1 F_0 + d_1 G_0. \end{aligned}$$

However, note that the

$$-b_k G_k - \sum_{i=2}^{k} (a_{i-1} - c_i) F_{i-1} - \sum_{i=2}^{k} (b_{i-1} - d_i) G_{i-1} - \sum_{i=1}^{k} E_i$$

terms are wasted in the analysis. If one has the freedom to do so, it may be good to choose parameters so that

$$a_{i-1} = c_i, \; b_{i-1} = d_i$$

and $E_i = 0$ for $i = 1, 2, \ldots$. Not having wasted terms may be an indication that the analysis is tight.

D SC-OGM via linear coupling

In this section, we analyze SC-OGM through the linear coupling analysis. We consider the linear coupling form

$$\begin{aligned}
y_{k+1} &= x_k - \frac{1}{L}Q^{-1}\nabla f(x_k)\\
z_{k+1} &= \frac{1}{1+\gamma}\left(z_k + \gamma x_k - \frac{\gamma}{\mu}Q^{-1}\nabla f(x_k)\right)\\
x_{k+1} &= \tau z_{k+1} + (1-\tau)y_{k+1},
\end{aligned}$$

where τ is a coupling coefficient to be determined. As an aside, we can view z_{k+1} as a mirror descent update of the form

$$z_{k+1} = \arg\min_z \left\{\frac{1}{2}\|z - z_k\|^2 + \frac{\gamma}{2}\|z - x_k\|^2 + \frac{\gamma}{\mu}\langle \nabla f(x_k), z\rangle\right\},$$

which is similar to what was considered in [6].

Lemma 12 *Assume (A1), (A2) and (A3). Then,*

$$\begin{aligned}
&\frac{\gamma}{\mu}\langle \nabla f(x_k), z_{k+1} - x_\star\rangle - \frac{\gamma}{2}\|x_k - x_\star\|^2\\
&\qquad\le -\frac{\gamma^2}{2(1+\gamma)\mu^2}\|\nabla f(x_k)\|_*^2 + \frac{1}{2}\|z_k - x_\star\|^2 - \frac{1+\gamma}{2}\|z_{k+1} - x_\star\|^2
\end{aligned}$$

for $k = 0, 1, \ldots$.

Proof This proof follows steps similar to that of [6, Lemma 5.4].

From the definition of z_{k+1}, we say

$$\begin{aligned}
0 =& \left\langle \frac{\partial}{\partial z}\left\{\frac{1}{2}\|z - z_k\|^2 + \frac{\gamma}{2}\|z - x_k\|^2 + \frac{\gamma}{\mu}\langle \nabla f(x_k), z\rangle\right\}\Big|_{z_{k+1}}, z_{k+1} - x_\star\right\rangle\\
=& \langle Q(z_{k+1} - z_k), z_{k+1} - x_\star\rangle + \frac{\gamma}{\mu}\langle \nabla f(x_k), z_{k+1} - x_\star\rangle + \gamma\langle Q(z_{k+1} - x_k), z_{k+1} - x_\star\rangle
\end{aligned}$$

By three point equation,

$$\begin{aligned}
&\frac{\gamma}{\mu}\langle \nabla f(x_k), z_{k+1} - x_\star\rangle + \gamma\left(\frac{1}{2}\|x_k - z_{k+1}\|^2 - \frac{1}{2}\|x_k - x_\star\|^2\right)\\
&\qquad = -\frac{1}{2}\|z_k - z_{k+1}\|^2 + \frac{1}{2}\|z_k - x_\star\|^2 - \frac{1+\gamma}{2}\|z_{k+1} - x_\star\|^2.
\end{aligned}$$

Plugging the definition of z_{k+1},

$$\begin{aligned}
&\frac{\gamma}{2}\|x_k - z_{k+1}\|^2 + \frac{1}{2}\|z_k - z_{k+1}\|^2\\
&= \frac{\gamma}{2}\left\|\frac{1}{1+\gamma}(x_k - z_k) + \frac{\gamma}{(1+\gamma)\mu}Q^{-1}\nabla f(x_k)\right\|^2 + \frac{1}{2}\left\|-\frac{\gamma}{1+\gamma}(x_k - z_k) + \frac{\gamma}{(1+\gamma)\mu}Q^{-1}\nabla f(x_k)\right\|^2\\
&\ge \frac{\gamma^2}{2(1+\gamma)\mu^2}\|\nabla f(x_k)\|_*^2.
\end{aligned}$$

Combining results above, we get

$$\begin{aligned}
&\frac{\gamma}{\mu}\langle \nabla f(x_k), z_{k+1} - x_\star\rangle - \frac{\gamma}{2}\|x_k - x_\star\|^2\\
&\qquad\le -\frac{\gamma^2}{2(1+\gamma)\mu^2}\|\nabla f(x_k)\|_*^2 + \frac{1}{2}\|z_k - x_\star\|^2 - \frac{1+\gamma}{2}\|z_{k+1} - x_\star\|^2.
\end{aligned}$$

□

Lemma 13 (Coupling lemma in SC-OGM) *Assume (A1), (A2) and (A3). Then*

$$(1+\gamma)\left(f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_*^2 + \frac{\mu}{2}\|z_k - x_\star\|^2\right) \leq \left(f(x_{k-1}) - \frac{1}{2L}\|\nabla f(x_{k-1})\|_*^2 + \frac{\mu}{2}\|z_{k-1} - x_\star\|^2\right)$$

holds for $k = 1, 2, \ldots$

Proof We have

$$\begin{aligned}
&\gamma\left(f(x_k) - f(x_\star)\right)\\
&\leq \gamma\langle \nabla f(x_k), x_k - x_\star\rangle - \frac{\mu\gamma}{2}\|x_k - x_\star\|^2\\
&= \gamma\langle \nabla f(x_k), x_k - z_k\rangle + \gamma\langle \nabla f(x_k), z_k - x_\star\rangle - \frac{\mu\gamma}{2}\|x_k - x_\star\|^2\\
&= \frac{1-\tau}{\tau}\gamma\langle \nabla f(x_k), y_k - x_k\rangle + \gamma\langle \nabla f(x_k), z_k - x_\star\rangle - \frac{\mu\gamma}{2}\|x_k - x_\star\|^2\\
&= \frac{1-\tau}{\tau}\gamma\langle \nabla f(x_k), x_{k-1} - x_k - \frac{1}{L}Q^{-1}\nabla f(x_{k-1})\rangle + \gamma\langle \nabla f(x_k), z_k - x_\star\rangle - \frac{\mu\gamma}{2}\|x_k - x_\star\|^2\\
&\leq \left(\frac{1-\tau}{\tau}\gamma - 1\right)\langle \nabla f(x_k), x_{k-1} - x_k - \frac{1}{L}Q^{-1}\nabla f(x_{k-1})\rangle\\
&\quad + \left(f(x_{k-1}) - f(x_k) - \frac{1}{2L}\|\nabla f(x_{k-1})\|_*^2 - \frac{1}{2L}\|\nabla f(x_k)\|_*^2\right)\\
&\qquad + \gamma\langle \nabla f(x_k), z_k - z_{k+1}\rangle + \gamma\langle \nabla f(x_k), z_{k+1} - x_\star\rangle - \frac{\mu\gamma}{2}\|x_k - x_\star\|^2\\
&\leq \left(\frac{1-\tau}{\tau}\gamma - 1\right)\langle \nabla f(x_k), y_k - x_k\rangle + \left(f(x_{k-1}) - f(x_k) - \frac{1}{2L}\|\nabla f(x_{k-1})\|_*^2 - \frac{1}{2L}\|\nabla f(x_k)\|_*^2\right)\\
&\quad + \gamma\langle \nabla f(x_k), z_k - z_{k+1}\rangle - \frac{\gamma^2}{2(1+\gamma)\mu}\|\nabla f(x_k)\|_*^2 + \frac{\mu}{2}\|z_k - x_\star\|^2 - \frac{(1+\gamma)\mu}{2}\|z_{k+1} - x_\star\|^2,
\end{aligned}$$

where the last inequality is an application of Lemma 12. Note that

$$\begin{aligned}
z_k - z_{k+1} &= z_k - \frac{1}{1+\gamma}\left(z_k + \gamma x_k - \frac{\gamma}{\mu}Q^{-1}\nabla f(x_k)\right)\\
&= \frac{\gamma}{1+\gamma}(z_k - x_k) + \frac{\gamma}{(1+\gamma)\mu}Q^{-1}\nabla f(x_k)\\
&= \frac{\gamma}{1+\gamma}\frac{1-\tau}{\tau}(x_k - y_k) + \frac{\gamma}{(1+\gamma)\mu}Q^{-1}\nabla f(x_k).
\end{aligned}$$

To eliminate the $\langle \nabla f(x_k), \cdot\rangle$ term, we choose τ to satisfy

$$\frac{1-\tau}{\tau}\gamma - 1 = \frac{\gamma}{1+\gamma}\frac{1-\tau}{\tau}. \tag{9}$$

Plugging this in, the inequality above is

$$\begin{aligned}
&\gamma\left(f(x_k) - f(x_\star)\right)\\
&\leq \left(f(x_{k-1}) - f(x_k) - \frac{1}{2L}\|\nabla f(x_{k-1})\|_*^2 - \frac{1}{2L}\|\nabla f(x_k)\|_*^2\right)\\
&\quad + \frac{\gamma^2}{2(1+\gamma)\mu}\|\nabla f(x_k)\|_*^2 + \frac{\mu}{2}\|z_k - x_\star\|^2 - \frac{(1+\gamma)\mu}{2}\|z_{k+1} - x_\star\|^2.
\end{aligned}$$

In order to make the telescoping form such as

$$M_k\Big(f(x_k) - B_k\left\|\nabla f(x_k)\right\|_*^2 + C_k\left\|z_{k+1} - x_\star\right\|^2\Big) \\ \leq N_{k-1}\left(f(x_{k-1}) - B_{k-1}\left\|\nabla f(x_{k-1})\right\|_*^2 + C_{k-1}\left\|z_k - x_\star\right\|^2\right),$$

we chose $B_k = \frac{1}{2L}$ and $C_k = \frac{\mu}{2}$, which leads to the choice of γ satisfying

$$\frac{2+\gamma}{2L} = \frac{\gamma^2}{2(1+\gamma)\mu}. \tag{10}$$

We get the desired result by plugging (9) and (10) in the above inequality. □

E Asymptotic characterization of θ_k

Theorem 7 *Let the positive sequence $\{\theta_k\}_{k=0}^\infty$ satisfy $\theta_0 = 1$ and $\theta_{k+1}^2 - \theta_{k+1} - \theta_k^2 = 0$ for $k = 0, 1, \ldots$. Then,*

$$\theta_k = \frac{k+\zeta+1}{2} + \frac{\log k}{4} + o(1).$$

Proof Let $\theta_k = \frac{k+2}{2} + c_k \log k$. The proof consists of the following 3 steps:

1. If $c_k < \frac{1}{4}$, then $c_{k+1} < \frac{1}{4}$.
2. $c_k \to \frac{1}{4}$ as $k \to \infty$.
3. If $\theta_k = \frac{k+2}{2} + \frac{\log k}{4} + e_k$, then e_k is convergent.

First step. If $c_k < \frac{1}{4}$, then $c_{k+1} < \frac{1}{4}$.

For our convenience, let $c_0 = 0$ with $c_0 \log 0 = 0$. Plugging this in $\theta_{k+1}^2 - \theta_{k+1} - \theta_k^2 = 0$, we have

$$\left(\frac{k+2}{2} + c_{k+1}\log(k+1)\right)^2 = \left(\frac{k+2}{2} + c_k \log k\right)^2 + \frac{1}{4},$$

so

$$\left(c_{k+1}\log(k+1) - c_k\log k\right)\left(k+2+c_{k+1}\log(k+1) + c_k\log k\right) = \frac{1}{4}.$$

Assume $c_{k+1} \geq 1/4$. Then

$$\begin{aligned}\frac{1}{4} &= \left(c_{k+1}\log(k+1) - c_k\log k\right)\left(k+2+c_{k+1}\log(k+1) + c_k\log k\right) \\ &\geq \frac{1}{4}\log\left(1+\frac{1}{k}\right)(k+2) \\ &> \frac{1}{4},\end{aligned}$$

which proves the first claim.

Second step. $c_k \to \frac{1}{4}$ as $k \to \infty$.

Put $d_k = \frac{1}{4} - c_k$, then $0 < d_k \leq \frac{1}{4}$.

$$\begin{aligned}\frac{1}{4} &= \left(\frac{1}{4}\log\left(1+\frac{1}{k}\right) - d_{k+1}\log(k+1) + d_k\log k\right)\left(k+2+\frac{1}{4}\log k(k+1) - d_{k+1}\log(k+1) - d_k\log k\right) \\ &\leq \left(\frac{1}{4}\log\left(1+\frac{1}{k}\right) - d_{k+1}\log(k+1) + d_k\log k\right)\left(k+2+\frac{1}{2}\log(k+1)\right)\end{aligned}$$

Therefore

$$d_{k+1}\log(k+1) - d_k \log k \le \frac{1}{4}\log\left(1+\frac{1}{k}\right) - \frac{1}{4}\frac{1}{k+2+\frac{1}{2}\log(k+1)}.$$

By talyor expansion,

$$d_{k+1}\log(k+1) - d_k \log k \le \frac{1}{4}\left(\frac{3+2\log k}{2k^2} + \mathcal{O}\left(\frac{1}{k^2}\right)\right).$$

So, By summing all the above inequality from 1 to k,

$$d_{k+1}\log(k+1) \le C$$

so $d_{k+1} < \frac{C}{\log(k+1)}$. In conclusion, as $k \to \infty$, $d_k \to 0$.

Third step. If $\theta_k = \frac{k+2}{2} + \frac{\log k}{4} + e_k$, then, e_k converges.

From the previous claim, we can say that for some sufficiently large k, $|e_k| < \frac{1}{6}\log k$.

$$\left(\frac{k+2}{2} + \frac{1}{4}\log(k+1) + e_{k+1}\right)^2 = \left(\frac{k+2}{2} + \frac{1}{4}\log k + e_k\right)^2 + \frac{1}{4}$$

Then,

$$\begin{aligned}\frac{1}{4} &= \left(\frac{1}{4}\log\left(1+\frac{1}{k}\right) + e_{k+1} - e_k\right)\left(k+2+\frac{1}{4}\log k(k+1) + e_{k+1} + e_k\right)\\ &\le \left(\frac{1}{4}\log\left(1+\frac{1}{k}\right) + e_{k+1} - e_k\right)\left(k+2+\frac{5}{6}\log(k+1)\right).\end{aligned}$$

So,

$$e_{k+1} - e_k \ge \frac{1}{4\left(k+2+\frac{5}{6}\log(k+1)\right)} - \frac{1}{4}\log\left(1+\frac{1}{k}\right) = -\frac{\frac{5}{6}\log k + \frac{3}{2}}{k^2} + \mathcal{O}\left(\frac{1}{k^2}\right).$$

Summing this for $k = 1, \dots, k$, we get that $e_{k+1} > D$ for some constant D. Moreover,

$$\begin{aligned}\frac{1}{4} &= \left(\frac{1}{4}\log\left(1+\frac{1}{k}\right) + e_{k+1} - e_k\right)\left(k+2+\frac{1}{4}\log k(k+1) + e_{k+1} + e_k\right)\\ &\ge \left(\frac{1}{4}\log\left(1+\frac{1}{k}\right) + e_{k+1} - e_k\right)(k+2) > \frac{1}{4} + (k+2)(e_{k+1} - e_k),\end{aligned}$$

which indicates that $e_{k+1} < e_k$. Since $\{e_k\}_{k=0}^{\infty}$ is a decreasing sequence with a lower bound, it converges. □

Proof of equality in Section 2.1 We have

$$\begin{aligned}\frac{L\|x_0 - x_\star\|^2}{2\theta_{k-1}^2} &= \frac{L\|x_0 - x_\star\|^2}{2\left(\frac{k+\zeta}{2} + \frac{\log(k-1)}{4} + o(1)\right)^2}\\ &= \frac{2L\|x_0 - x_\star\|^2}{(k+\zeta)^2\left(1 + \frac{\log(k-1)}{2(k+\zeta)} + o(1/k)\right)^2}\\ &= \frac{2L\|x_0 - x_\star\|^2}{(k+\zeta)^2}\left(1 - 2\frac{\log(k-1)}{2(k+\zeta)} + o(1/k)\right)\\ &= \frac{2L\|x_0 - x_\star\|^2}{(k+\zeta)^2} - \frac{2L\|x_0 - x_\star\|^2\log k}{(k+\zeta)^3} + o\left(\frac{1}{k^3}\right),\end{aligned}$$

which verifies the equality in Section 2.1.