CrossMark

Adaptive Restart of the Optimized Gradient Method for Convex Optimization

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

Abstract First-order methods with momentum, such as Nesterov's fast gradient method, are very useful for convex optimization problems, but can exhibit undesirable oscillations yielding slow convergence rates for some applications. An adaptive restarting scheme can improve the convergence rate of the fast gradient method, when the parameter of a strongly convex cost function is unknown or when the iterates of the algorithm enter a locally strongly convex region. Recently, we introduced the optimized gradient method, a first-order algorithm that has an inexpensive per-iteration computational cost similar to that of the fast gradient method, yet has a worst-case cost function rate that is twice faster than that of the fast gradient method and that is optimal for large-dimensional smooth convex problems. Building upon the success of accelerating the fast gradient method using adaptive restart, this paper investigates similar heuristic acceleration of the optimized gradient method. We first derive a new first-order method that resembles the optimized gradient method for strongly convex quadratic problems with known function parameters, yielding a linear convergence rate that is faster than that of the analogous version of the fast gradient method. We then provide a heuristic analysis and numerical experiments that illustrate that adaptive restart can accelerate the convergence of the optimized gradient method. Numerical results also illustrate that adaptive restart is helpful for a proximal version of the optimized gradient method for nonsmooth composite convex functions.

✉ Donghwan Kim
  kimdongh@umich.edu

  Jeffrey A. Fessler
  fessler@umich.edu

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

1 Introduction

The computational expense of first-order methods depends only mildly on the problem
dimension, so they are attractive for solving large-dimensional optimization prob-
lems [1]. In particular, Nesterov's fast gradient method (FGM) [2–4] is used widely,
because it has a worst-case cost function rate that is optimal up to constant for large-
dimensional smooth convex problems [3]. In addition, for smooth and strongly convex
problems where the strong convexity parameter is known, a version of FGM has a linear
convergence rate [3] that improves upon that of a standard gradient method. However,
without knowledge of the function parameters, conventional FGM does not guarantee
a linear convergence rate.

When the strong convexity parameter is unknown, a simple adaptive restarting
scheme [5] for FGM heuristically improves its convergence rate. (See also [6,7] for
theory and [1,8,9] for applications.) In addition, adaptive restart is useful even when the
function is only locally strongly convex near the minimizer [5]. First-order methods are
known to be suitable when only moderate solution accuracy is required, and adaptive
restart can help first-order methods achieve medium to high accuracy.

Recently, we proposed the optimized gradient method (OGM) [10] (built upon [11])
that has efficient per-iteration computation similar to FGM yet that exactly achieves
the optimal worst-case rate for decreasing a large-dimensional smooth convex function
among all first-order methods [12]. (See [13–15] for further analysis and extensions
of OGM.) This paper examines a general class of accelerated first-order methods that
includes a gradient method (GM), FGM, and OGM for strongly convex *quadratic*
functions, and develops an OGM variant, named OGM-q. This method provides a
linear convergence rate that is faster than that of the analogous version of FGM. The
analysis reveals that, like FGM [5], OGM may exhibit undesirable oscillating behavior
in some cases. Building on the quadratic analysis and the adaptive restart scheme of
FGM in [5], we propose an adaptive restart scheme that heuristically accelerates the
convergence rate of OGM when the function is strongly convex or even when it is
only locally strongly convex. This restart scheme circumvents the oscillating behavior.
Numerical results illustrate that the proposed OGM with restart performs better than
FGM with restart in [5].

Section 2 reviews first-order methods for convex problems such as GM, FGM,
and OGM. Section 3 analyzes a general class of accelerated first-order methods that
includes GM, FGM, and OGM for strongly convex quadratic problems, and proposes
a new OGM variant with a fast linear convergence rate. Section 4 suggests an adaptive
restart scheme for OGM using the quadratic analysis in Sect. 3. Section 5 illustrates the
proposed adaptive version of OGM that we use for numerical experiments on various
convex problems in Sect. 6, including nonsmooth composite convex functions, and
Sect. 7 concludes.

2 Problem and Methods

2.1 Smooth and Strongly Convex Problem

We first consider the smooth and strongly convex minimization problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \; f(\boldsymbol{x}) \tag{1}$$

that satisfies the following smooth and strongly convex conditions:

- $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and has Lipschitz continuous gradient with Lipschitz constant $L > 0$, i.e.,

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

- f is strongly convex with strong convexity parameter $\mu > 0$, i.e.,

$$f(\boldsymbol{x}) \ge f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \; \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\mu}{2}||\boldsymbol{x} - \boldsymbol{y}||^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d. \tag{3}$$

We let $\mathscr{F}_{\mu,L}(\mathbb{R}^d)$ denote the class of functions f that satisfy the above two conditions hereafter, and let $\boldsymbol{x}_*$ denote the unique minimizer of f. We let $q := \mu/L$ denote the reciprocal of the condition number of a function $f \in \mathscr{F}_{\mu,L}(\mathbb{R}^d)$. We also let $\mathscr{F}_{0,L}(\mathbb{R}^d)$ denote the class of smooth convex functions f that satisfy the above two conditions with $\mu = 0$, and let $\boldsymbol{x}_*$ denote a minimizer of f.

Some algorithms discussed in this paper require knowledge of both μ and L, but in many cases estimating μ is challenging compared to computing L.[1] Therefore, this paper focuses on the case where the parameter μ is unavailable, while L is available. Even without knowing μ, the adaptive restart approach in [5] and the proposed adaptive restart approach in this paper both exhibit linear convergence rates in strongly convex cases.

2.2 Review of Accelerated First-Order Methods

This paper focuses on accelerated first-order methods (AFM) of the form shown in Algorithm 1 for solving (1). The fast gradient method (FGM) [2–4] (with $\gamma_k = 0$ in Algorithm 1) accelerates the gradient method (GM) (with $\beta_k = \gamma_k = 0$) using the *momentum* term $\beta_k(\boldsymbol{y}_{k+1} - \boldsymbol{y}_k)$ with negligible additional computation. The optimized gradient method (OGM) [10,14] uses an *over-relaxation* term $\gamma_k(\boldsymbol{y}_{k+1} - \boldsymbol{x}_k) = -\gamma_k \alpha \nabla f(\boldsymbol{x}_k)$ for further acceleration.

Tables 1 and 2 summarize the standard choices of coefficients $(\alpha, \beta_k, \gamma_k)$ for GM, FGM, OGM in [2–4,10,14] and their worst-case rates for smooth convex functions

---

[1] For some applications even estimating L is expensive, and one must employ a backtracking scheme [3,4] or similar approaches. We assume L is known throughout this paper. An estimate of μ could be found by a backtracking scheme as described in [16, Sec. 5.3].

Algorithm 1 Accelerated First-order Methods (AFM)

1: **Input:** $f \in \mathscr{F}_{0,L}(\mathbb{R}^d)$ or $\mathscr{F}_{\mu,L}(\mathbb{R}^d)$, $x_0 = y_0 \in \mathbb{R}^d$.
2: **for** $k \geq 0$ **do**
3:   $y_{k+1} = x_k - \alpha \nabla f(x_k)$
4:   $x_{k+1} = y_{k+1} + \beta_k(y_{k+1} - y_k) + \gamma_k(y_{k+1} - x_k)$

Table 1 Accelerated first-order methods for smooth convex problems

Method	α	β_k	γ_k	Worst-case rate
GM	$\frac{1}{L}$	0	0	$f(y_k) - f(x_*) \leq \frac{L\|x_0 - x_*\|^2}{4k+2}$ [11]
FGM [2]	$\frac{1}{L}$	$\frac{t_k-1}{t_{k+1}}$	0	$f(y_k) - f(x_*) \leq \frac{L\|x_0 - x_*\|^2}{2t_{k-1}^2} \leq \frac{2L\|x_0 - x_*\|^2}{(k+1)^2}$ [4]
				$f(x_k) - f(x_*) \leq \frac{L\|x_0 - x_*\|^2}{2t_k^2} \leq \frac{2L\|x_0 - x_*\|^2}{(k+2)^2}$ [10]
OGM′ [14]	$\frac{1}{L}$	$\frac{t_k-1}{t_{k+1}}$	$\frac{t_k}{t_{k+1}}$	$f(y_k) - f(x_*) \leq \frac{L\|x_0 - x_*\|^2}{4t_{k-1}^2} \leq \frac{L\|x_0 - x_*\|^2}{(k+1)^2}$ [14]
OGM [10]	$\frac{1}{L}$	$\frac{\theta_k-1}{\theta_{k+1}}$	$\frac{\theta_k}{\theta_{k+1}}$	$f(x_N) - f(x_*) \leq \frac{L\|x_0 - x_*\|^2}{2\theta_N^2} \leq \frac{L\|x_0 - x_*\|^2}{(N+1)^2}$ [10]

Parameters

$t_0 = 1$, $\quad t_k = \frac{1}{2}\left(1 + \sqrt{1 + 4t_{k-1}^2}\right)$, $\quad k = 1, \ldots,$

$\theta_0 = 1$, $\quad \theta_k = \begin{cases} \frac{1}{2}\left(1 + \sqrt{1 + 4\theta_{k-1}^2}\right), & k = 1, \ldots, N-1, \\ \frac{1}{2}\left(1 + \sqrt{1 + 8\theta_{k-1}^2}\right), & k = N. \end{cases}$

The worst-case rate of OGM depends on the preselected total number of iterations N, unlike other rates

---

Table 2 Accelerated first-order methods (with $\gamma_k = 0$) for smooth and strongly convex problems

Method	α	β_k	Worst-case rate
GM	$\frac{1}{L}$	0	$f(y_k) - f(x_*) \leq \left(1 - \frac{2\mu}{1+q}\right)^k \frac{L\|x_0 - x_*\|^2}{2}$ [3]
GM-q	$\frac{2}{\mu+L}$	0	$f(y_k) - f(x_*) \leq \left(\frac{1-q}{1+q}\right)^{2k} \frac{L\|x_0 - x_*\|^2}{2}$ [3]
FGM-q [3]	$\frac{1}{L}$	$\frac{1-\sqrt{q}}{1+\sqrt{q}}$	$f(y_k) - f(x_*) \leq (1 - \sqrt{q})^k \frac{(1+q)L\|x_0 - x_*\|^2}{2}$ [3]

The worst-case rates also apply to $\frac{\mu}{2}\|y_k - x_*\|^2$ due to strong convexity (3)

$\mathscr{F}_{0,L}(\mathbb{R}^d)$ and smooth and strongly convex functions $\mathscr{F}_{\mu,L}(\mathbb{R}^d)$, respectively. (Other choices can be found in [3,13,17].) For convenience hereafter, we use the names GM, GM-q, FGM, FGM-q, OGM, and OGM′ to distinguish different choices of standard AFM coefficients in Tables 1 and 2.

The worst-case OGM rate [10] in Table 1 is about twice faster than the FGM rate [4] and is optimal for first-order methods for the function class $\mathscr{F}_{0,L}(\mathbb{R}^d)$ under the large-scale condition $d \geq N + 1$ [12]. However, it is yet unknown which first-order methods provide an optimal worst-case linear convergence rate for the function class $\mathscr{F}_{\mu,L}(\mathbb{R}^d)$;

this topic is left as an interesting future work.[2] Toward this direction, Sect. 3 studies AFM for strongly convex *quadratic* problems, leading to a new method named OGM-q with a linear convergence rate that is faster than that of FGM-q. Section 4 uses this quadratic analysis to analyze an adaptive restart scheme for OGM.

3 Analysis of AFM for Quadratic Functions

This section analyzes the behavior of AFM for minimizing a strongly convex quadratic function. The quadratic analysis of AFM in this section is similar in spirit to the analyses of a heavy-ball method [19, Sec. 3.2] and AFM with $\gamma_k = 0$ [20, Appx. A] [5, Sec. 4].

In addition, Sect. 3.3 optimizes the coefficients of AFM for such quadratic functions, yielding a linear convergence rate that is faster than that of FGM-q. The resulting method, named OGM-q, requires the knowledge of q, and Sect. 3.4 shows that using OGM (and OGM′) in Table 1 instead (without the knowledge of q) will cause the OGM iterates to oscillate when the momentum is larger than a critical value. This analysis stems from the dynamical system analysis of AFM with $\alpha = 1/L$ and $\gamma_k = 0$ in [5, Sec. 4].

3.1 Quadratic Analysis of AFM

This section considers minimizing a strongly convex quadratic function:

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{p}^\top \boldsymbol{x} \in \mathscr{F}_{\mu,L}(\mathbb{R}^d) \tag{4}$$

where $\boldsymbol{Q} \in \mathbb{R}^{d \times d}$ is a symmetric positive definite matrix, and $\boldsymbol{p} \in \mathbb{R}^d$ is a vector. Here, $\nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{p}$ is the gradient, and $\boldsymbol{x}_* = \boldsymbol{Q}^{-1}\boldsymbol{p}$ is the optimum. The smallest and the largest eigenvalues of $\boldsymbol{Q}$ correspond to the parameters μ and L of the function, respectively. For simplicity in the quadratic analysis, we consider the version of AFM that has constant coefficients (α, β, γ).

Defining the vectors $\boldsymbol{\xi}_k := (\boldsymbol{x}_k^\top, \boldsymbol{x}_{k-1}^\top)^\top \in \mathbb{R}^{2d}$ and $\boldsymbol{\xi}_* := (\boldsymbol{x}_*^\top, \boldsymbol{x}_*^\top)^\top \in \mathbb{R}^{2d}$, and extending the analysis for AFM with $\gamma = 0$ in [20, Appx. A], AFM has the following equivalent form for $k \geq 1$:

$$\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_* = \boldsymbol{T}(\alpha, \beta, \gamma)\,(\boldsymbol{\xi}_k - \boldsymbol{\xi}_*), \tag{5}$$

where the system matrix $\boldsymbol{T}(\alpha, \beta, \gamma)$ of AFM is defined as

$$\boldsymbol{T}(\alpha, \beta, \gamma) := \begin{bmatrix} (1+\beta)(\boldsymbol{I} - \alpha\boldsymbol{Q}) - \gamma\alpha\boldsymbol{Q} & -\beta(\boldsymbol{I} - \alpha\boldsymbol{Q}) \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{2d \times 2d} \tag{6}$$

---

[2] Recently, [18] developed a new first-order method for known q that is not in AFM class but achieves a linear worst-case rate $(1 - \sqrt{q})^2$ for the decrease of a strongly convex function that is faster than the linear rate $(1 - \sqrt{q})$ of FGM-q in Table 2.

for an identity matrix $I \in \mathbb{R}^{d \times d}$. The sequence $\{\tilde{\boldsymbol{\xi}}_k := (\boldsymbol{y}_k^\top, \boldsymbol{y}_{k-1}^\top)^\top\}_{k \geq 1}$ also satisfies recursion (5), implying that (5) characterizes the behavior of both the primary sequence $\{\boldsymbol{y}_k\}$ and the secondary sequence $\{\boldsymbol{x}_k\}$ of AFM with constant coefficients.

The spectral radius $\rho(\boldsymbol{T}(\cdot))$ of matrix $\boldsymbol{T}(\cdot)$ determines the convergence rate of the algorithm. Specifically, for any $\epsilon > 0$, there exists $K \geq 0$ such that $[\rho(\boldsymbol{T})]^k \leq ||\boldsymbol{T}^k|| \leq (\rho(\boldsymbol{T}) + \epsilon)^k$ for all $k \geq K$, establishing the following worst-case rate:

$$||\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_*||^2 \leq (\rho(\boldsymbol{T}(\alpha, \beta, \gamma)) + \epsilon)^{2k} ||\boldsymbol{\xi}_1 - \boldsymbol{\xi}_*||^2. \tag{7}$$

We next analyze $\rho(\boldsymbol{T}(\alpha, \beta, \gamma))$.

Considering the eigen decomposition of $\boldsymbol{Q}$ in $\boldsymbol{T}(\cdot)$ as in [20, Appx. A], the spectral radius of $\boldsymbol{T}(\cdot)$ is:

$$\rho(\boldsymbol{T}(\alpha, \beta, \gamma)) = \max_{\mu \leq \lambda \leq L} \rho(\boldsymbol{T}_\lambda(\alpha, \beta, \gamma)), \tag{8}$$

where for any eigenvalue λ of matrix $\boldsymbol{Q}$ we define a matrix $\boldsymbol{T}_\lambda(\alpha, \beta, \gamma) \in \mathbb{R}^{2 \times 2}$ by substituting λ and 1 for $\boldsymbol{Q}$ and $\boldsymbol{I}$ in $\boldsymbol{T}(\alpha, \beta, \gamma)$, respectively. Similar to the analysis of AFM with $\gamma = 0$ in [20, Appx. A], the spectral radius of $\boldsymbol{T}_\lambda(\alpha, \beta, \gamma)$ is:

$$\rho(\boldsymbol{T}_\lambda(\alpha, \beta, \gamma)) = \max\{|r_1(\alpha, \beta, \gamma, \lambda)|, |r_2(\alpha, \beta, \gamma, \lambda)|\} \tag{9}$$
$$= \begin{cases} \frac{1}{2} \left(|(1 + \beta)(1 - \alpha\lambda) - \gamma\alpha\lambda| + \sqrt{\Delta(\alpha, \beta, \gamma, \lambda)}\right), & \Delta(\alpha, \beta, \gamma, \lambda) \geq 0, \\ \sqrt{\beta(1 - \alpha\lambda)}, & \text{otherwise,} \end{cases}$$

where $r_1(\alpha, \beta, \gamma, \lambda)$ and $r_2(\alpha, \beta, \gamma, \lambda)$ denote the roots of the characteristic polynomial of $\boldsymbol{T}_\lambda(\cdot)$:

$$r^2 - ((1 + \beta)(1 - \alpha\lambda) - \gamma\alpha\lambda)r + \beta(1 - \alpha\lambda), \tag{10}$$

and $\Delta(\alpha, \beta, \gamma, \lambda) := ((1 + \beta)(1 - \alpha\lambda) - \gamma\alpha\lambda)^2 - 4\beta(1 - \alpha\lambda)$ denotes the corresponding discriminant. For fixed (α, β, γ), the spectral radius $\rho(\boldsymbol{T}_\lambda(\alpha, \beta, \gamma))$ in (9) is a continuous and quasi-convex[3] function of λ; thus, its maximum over λ occurs at one of its boundary points $\lambda = \mu$ or $\lambda = L$.

The next section reviews the optimization of AFM coefficients to provide the fastest convergence rate, i.e., the smallest spectral radius $\rho(\boldsymbol{T}(\cdot))$ in (8), under certain constraints on (α, β, γ).

---

[3] It is straightforward to show that $\rho(\boldsymbol{T}_\lambda(\alpha, \beta, \gamma))$ in (9) is quasi-convex over λ, i.e., $\rho(\boldsymbol{T}_{\kappa\lambda_1 + (1-\kappa)\lambda_2}(\cdot)) \leq \max\{\rho(\boldsymbol{T}_{\lambda_1}(\cdot)), \rho(\boldsymbol{T}_{\lambda_2}(\cdot))\}$ for all $\lambda_1, \lambda_2 \in \mathbb{R}$ and $\kappa \in [0, 1]$. First, $\sqrt{\beta(1 - \alpha\lambda)}$ is quasi-convex over λ (for $\Delta(\alpha, \beta, \gamma, \lambda) < 0$). Second, the eigenvalue λ satisfying $\Delta(\alpha, \beta, \gamma, \lambda) \geq 0$ is in the region where the function $\frac{1}{2}\left(|(1 + \beta)(1 - \alpha\lambda) - \gamma\alpha\lambda| + \sqrt{\Delta(\alpha, \beta, \gamma, \lambda)}\right)$ either monotonically increases or decreases, which overall makes the continuous function $\rho(\boldsymbol{T}_\lambda(\alpha, \beta, \gamma))$ quasi-convex over λ. This proof can be simply applied to other variables, i.e., $\rho(\boldsymbol{T}_\lambda(\alpha, \beta, \gamma))$ is quasi-convex over either α, β, or γ.

3.2 Review of Optimizing AFM Coefficients Under Certain Constraints on (α, β, γ)

The AFM coefficients that provide the fastest convergence for minimizing a strongly convex quadratic function would solve

$$\arg\min_{\alpha,\beta,\gamma} \rho(\boldsymbol{T}(\alpha, \beta, \gamma)) = \arg\min_{\alpha,\beta,\gamma} \max\{\rho(\boldsymbol{T}_\mu(\alpha, \beta, \gamma)), \rho(\boldsymbol{T}_L(\alpha, \beta, \gamma))\}. \quad (11)$$

Note that a heavy-ball method [19] (that is not in AFM class) with similarly optimized coefficients has a linear worst-case rate with $\rho(\cdot) = \frac{1-\sqrt{q}}{1+\sqrt{q}}$ that is optimal (up to constant) for strongly convex quadratic problems [3]. Thus, optimizing (11) would be of little practical benefit for quadratic problems. Nevertheless, such optimization is new to AFM for $\gamma > 0$ (with the additional constraint $\alpha = 1/L$ introduced below) and is useful in our later analysis for the adaptive restart in Sect. 4. A heavy-ball method with the coefficients optimized for strongly convex quadratic problems does not converge for some strongly convex nonquadratic problems [20], and other choices of coefficients do not yield worst-case rates that are comparable to those of some accelerated choices of AFM [11,20], so we focus on AFM hereafter.

Coefficient optimization (11) for AFM was studied previously with various constraint. For example, optimizing (11) over α with the constraint $\beta = \gamma = 0$ yields GM-q. Similarly, FGM-q results from optimizing (11) over β for the constraint[4] $\alpha = 1/L$ and $\gamma = 0$. In [20, Prop. 1], AFM with coefficients $(\alpha, \beta, \gamma) = \left(\frac{4}{\mu+3L}, \frac{\sqrt{3+q}-2\sqrt{q}}{\sqrt{3+q}+2\sqrt{q}}, 0\right)$, named FGM$'$-$q$ in Table 3, was derived by optimizing (11) over (α, β) with the constraint $\gamma = 0$.

Although a general unconstrained solution to (11) would be an interesting future direction, here we focus on optimizing (11) over (β, γ) with the constraint $\alpha = 1/L$. This choice simplifies problem (11) and is useful for analyzing an adaptive restart scheme for OGM in Sect. 4.

3.3 Optimizing the Coefficients (β, γ) of AFM When $\alpha = 1/L$

When $\alpha = 1/L$ and $\lambda = L$, characteristic polynomial (10) becomes $r^2 + \gamma r = 0$. The roots are $r = 0$ and $r = -\gamma$, so $\rho(\boldsymbol{T}_L(1/L, \beta, \gamma)) = |\gamma|$. In addition, because $\rho(\boldsymbol{T}_\mu(1/L, \beta, \gamma))$ is continuous and quasi-convex over β (see footnote 3), it can be easily shown that the smaller value of β satisfying the following equation:

$$\Delta(1/L, \beta, \gamma, \mu) = ((1+\beta)(1-q) - \gamma q)^2 - 4\beta(1-q) \quad (12)$$
$$= (1-q)^2\beta^2 - 2(1-q)(1+q+q\gamma)\beta + (1-q)(1-q-2q\gamma) + q^2\gamma^2 = 0$$

---

[4] For FGM-q the value of $\rho(\boldsymbol{T}_L(1/L, \beta, 0))$ is 0, and the function $\rho(\boldsymbol{T}_\mu(1/L, \beta, 0))$ is continuous and quasi-convex over β (see footnote 3). The minimum of $\rho(\boldsymbol{T}_\mu(1/L, \beta, 0))$ occurs at the point $\beta = \frac{1-\sqrt{q}}{1+\sqrt{q}}$ in Table 2 satisfying $\Delta(1/L, \beta, 0, \mu) = 0$, verifying the statement that FGM-q results from optimizing (11) over β given $\alpha = 1/L$ and $\gamma = 0$.

minimizes $\rho(\boldsymbol{T}_\mu(1/L, \beta, \gamma))$ for any given γ (satisfying $\gamma \geq -1$). The optimal β for a given γ (when $\alpha = 1/L$) is

$$\beta^\star(\gamma) := \left(1 - \sqrt{q(1+\gamma)}\right)^2 / (1-q), \tag{13}$$

which reduces to $\beta = \beta^\star(0) = \frac{1-\sqrt{q}}{1+\sqrt{q}}$ for FGM-q (with $\gamma = 0$). Substituting (13) into (9) yields $\rho(\boldsymbol{T}_\mu(1/L, \beta^\star(\gamma), \gamma)) = |1 - \sqrt{q(1+\gamma)}|$, leading to the following simplification of (11) with $\alpha = 1/L$ and $\beta = \beta^\star(\gamma)$ from (13):

$$\gamma^\star := \arg\min_{\gamma} \ \max\left\{|1 - \sqrt{q(1+\gamma)}|, \ |\gamma|\right\}. \tag{14}$$

The minimizer of (14) satisfies $1 - \sqrt{q(1+\gamma)} = \pm\gamma$, and with simple algebra, we get the following solutions to (11) with the constraint $\alpha = 1/L$ [and (14)]:

$$\beta^\star := \beta^\star(\gamma^\star) = \frac{(\gamma^\star)^2}{1-q} = \frac{(2 + q - \sqrt{q^2 + 8q})^2}{4(1-q)}, \quad \gamma^\star = \frac{2 + q - \sqrt{q^2 + 8q}}{2}, \tag{15}$$

for which the spectral radius is $\rho^\star := \rho(\boldsymbol{T}(1/L, \beta^\star, \gamma^\star)) = 1 - \sqrt{q(1+\gamma^\star)} = \gamma^\star$. We denote Algorithm 1 with coefficients $\alpha = 1/L$ and $(\beta^\star, \gamma^\star)$ in (15) as OGM-q.

Table 3 compares the spectral radius of the OGM-q to GM-q, FGM-q, and FGM$'$-q [20, Prop. 1]. Simple algebra shows that the spectral radius of OGM-q is smaller than those of FGM-q and FGM$'$-q, i.e., $\frac{2+q-\sqrt{q^2+8q}}{2} \leq 1 - \frac{2\sqrt{q}}{\sqrt{3+q}} \leq 1 - \sqrt{q}$. Therefore, OGM-$q$ achieves a worst-case convergence rate of $||\boldsymbol{\xi}_k - \boldsymbol{\xi}_*||^2$ that is faster than that of FGM variants (but that is slower than a heavy-ball method [19]) for a strongly convex quadratic function.

To further understand the behavior of AFM for each eigen mode, Fig. 1 plots $\rho(\boldsymbol{T}_\lambda(1/L, \beta, \gamma))$ over $\mu \leq \lambda \leq L$ for $\mu = 0.1$ and $L = 1$ ($q = 0.1$) as an

Table 3 Optimally tuned coefficients (α, β, γ) of GM-q, FGM-q, FGM$'$-q, and OGM-q, and their spectral radius $\rho(\boldsymbol{T}(\alpha, \beta, \gamma))$ (8)

Method	α	β	γ	$\rho(T(\alpha,\beta,\gamma))$
GM-q	$\frac{2}{\mu+L}$	0	0	$\frac{1-q}{1+q}$
FGM-q [3]	$\frac{1}{L}$	$\frac{1-\sqrt{q}}{1+\sqrt{q}}$	0	$1-\sqrt{q}$
FGM$'$-q [20]	$\frac{4}{\mu+3L}$	$\frac{\sqrt{3+q}-2\sqrt{q}}{\sqrt{3+q}+2\sqrt{q}}$	0	$1-\frac{2\sqrt{q}}{\sqrt{3+q}}$
OGM-q	$\frac{1}{L}$	$\frac{(2+q-\sqrt{q^2+8q})^2}{4(1-q)}$	$\frac{2+q-\sqrt{q^2+8q}}{2}$	$\frac{2+q-\sqrt{q^2+8q}}{2}$

These optimal coefficients result from solving (11) with the shaded coefficients fixed
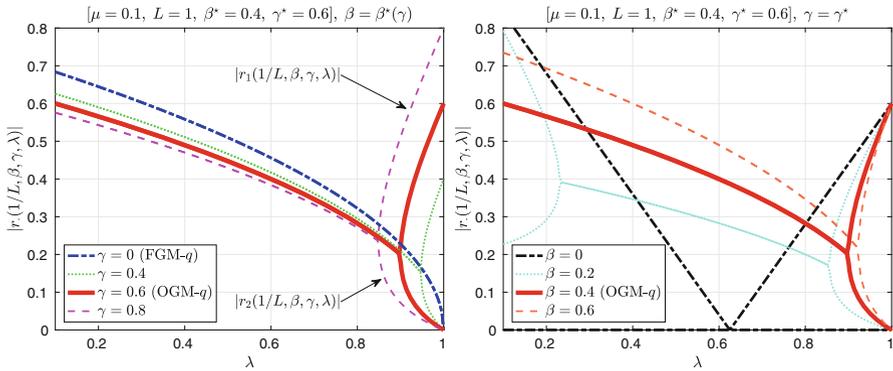
Fig. 1 Plots of $|r_1(1/L, \beta, \gamma, \lambda)|$ and $|r_2(1/L, \beta, \gamma, \lambda)|$ over $\mu \leq \lambda \leq L$ for various (left) γ values for given $\beta = \beta^\star(\gamma)$, and (right) β values for given $\gamma = \gamma^\star$, for a strongly convex quadratic problem with $\mu = 0.1$ and $L = 1$ ($q = 0.1$), where $(\beta^\star, \gamma^\star) = (0.4, 0.6)$. The maximum of $|r_1(1/L, \beta, \gamma, \lambda)|$ and $|r_2(1/L, \beta, \gamma, \lambda)|$, i.e., the upper curve in the plot, corresponds to the value of $\rho(\boldsymbol{T}_\lambda(1/L, \beta, \gamma))$ in (9), and the maximum value of $\rho(\boldsymbol{T}_\lambda(1/L, \beta, \gamma))$ over λ corresponds to a spectral radius $\rho(\boldsymbol{T}(1/L, \beta, \gamma))$ in (8)

example, where $(\beta^\star, \gamma^\star) = (0.4, 0.6)$. The left plot of Fig. 1 first compares the $\rho(\boldsymbol{T}_\lambda(1/L, \beta, \gamma))$ values of OGM-q to those of other choices of $\gamma = 0, 0.4, 0.8$ with $\beta = \beta^\star(\gamma)$ in (13). The OGM-q (see upper red curve in Fig. 1) has the largest value ($\rho^\star = \gamma^\star = 0.6$) of $\rho(\boldsymbol{T}_\lambda(1/L, \beta, \gamma))$ at both the smallest and the largest eigenvalues (μ and L, respectively), unlike other choices of γ (with $\beta^\star(\gamma)$) where either $\rho(\boldsymbol{T}_\mu(1/L, \beta, \gamma))$ or $\rho(\boldsymbol{T}_L(1/L, \beta, \gamma))$ is the largest. The other choices thus have a spectral radius $\rho(\boldsymbol{T}(1/L, \beta, \gamma))$ larger than that of the OGM-q.

The right plot of Fig. 1 illustrates $\rho(\boldsymbol{T}_\lambda(1/L, \beta, \gamma))$ values for different choices of $\beta = 0, 0.2, 0.4, 0.6$ for given $\gamma = \gamma^\star$, showing that suboptimal β value will slow down convergence, compared to the optimal $\beta^\star = 0.4$. AFM with $(\alpha, \beta, \gamma) = (1/L, 0, \gamma^\star)$ in Fig. 1 is equivalent to AFM with $\left(\frac{1}{L}(1 + \gamma^\star), 0, 0\right)$, and this implies that AFM with $\beta = \gamma = 0$ (e.g., GM) may have some modes for mid-valued λ values that will converge faster than the accelerated methods, whereas its overall convergence rate (i.e., the spectral radius value) is worse. Apparently no one method can have superior convergence rates for all modes.

Similarly, although OGM-q has the smallest possible spectral radius $\rho(\boldsymbol{T}(\cdot))$ among known AFM, the upper blue and red curves in the left plot of Fig. 1, corresponding to FGM-q and OGM-q, respectively, illustrate that OGM-q will have modes for large eigenvalues that converge slower than with FGM-q. This behavior may be undesirable when such modes of large eigenvalues dominate the overall convergence behavior.

The next section reveals that the convergence of the primary sequence $\{\boldsymbol{y}_k\}$ of AFM with $\alpha = 1/L$ is not governed by such modes of large eigenvalues unlike its secondary sequence $\{\boldsymbol{x}_k\}$. In addition, Fig. 1 reveals change points across λ meaning that there are different regimes; the next section elaborates on this behavior, building upon the dynamical system analysis of AFM with $\alpha = 1/L$ and $\gamma = 0$ in [5, Sec. 4].

3.4 Convergence Properties of AFM When $\alpha = 1/L$

[5, Sec. 4] analyzed a constant-step AFM with $\alpha = 1/L$ and $\gamma = 0$ as a linear dynamical system for minimizing strongly convex quadratic function (4) and showed that there are three regimes of behavior for the system: low-momentum, optimal-momentum, and high-momentum regimes. This section similarly analyzes AFM with $\alpha = 1/L$ and $\gamma > 0$ to better understand its convergence behavior when solving strongly convex quadratic problem (4), complementing the previous section's spectral radius analysis of AFM.

We use the eigen decomposition of $\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ with $\boldsymbol{\Lambda} := \mathrm{diag}\{\lambda_i\}$, where the eigenvalues $\{\lambda_i\}$ are in an ascending order, i.e., $\mu = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d = L$. And for simplicity, we let $\boldsymbol{p} = \boldsymbol{0}$ without loss of generality, leading to $\boldsymbol{x}_* = \boldsymbol{0}$. By defining $\boldsymbol{w}_k := (w_{k,1}, \ldots, w_{k,d})^\top = \boldsymbol{V}^\top \boldsymbol{y}_k \in \mathbb{R}^d$ and $\boldsymbol{v}_k := (v_{k,1}, \ldots, v_{k,d})^\top = \boldsymbol{V}^\top \boldsymbol{x}_k \in \mathbb{R}^d$ as the mode coefficients of the primary and secondary sequences, respectively, and using (5), we have the following d independently evolving identical recurrence relations for the evolution of $w_{\cdot,i}$ and $v_{\cdot,i}$ of the constant-step AFM with $\alpha = 1/L$, respectively:

$$w_{k+2,i} = ((1+\beta)(1-\lambda_i/L) - \gamma\lambda_i/L) w_{k+1,i} - \beta(1-\lambda_i/L) w_{k,i}, \qquad (16)$$
$$v_{k+2,i} = ((1+\beta)(1-\lambda_i/L) - \gamma\lambda_i/L) v_{k+1,i} - \beta(1-\lambda_i/L) v_{k,i},$$

for $i = 1, \ldots, d$, although the initial conditions differ as follows:

$$w_{1,i} = (1-\lambda_i/L)w_{0,i}, \quad v_{1,i} = ((1+\beta+\gamma)(1-\lambda_i/L) - (\beta+\gamma))v_{0,i} \qquad (17)$$

with $w_{0,i} = v_{0,i}$. The convergence behavior of the ith mode of the dynamical system of both $w_{\cdot,i}$ and $v_{\cdot,i}$ in (16) is determined by characteristic polynomial (10) with $\alpha = 1/L$ and $\lambda = \lambda_i$. Unlike the previous sections that studied only the worst-case convergence performance using the largest absolute value of the roots of polynomial (10), we next discuss the convergence behavior of AFM more comprehensively using (10) with $\alpha = 1/L$ and $\lambda = \lambda_i$ for the two cases 1) $\lambda_i = L$ and 2) $\lambda_i < L$.

1. $\lambda_i = L$: Characteristic polynomial (10) of the mode of $\lambda_i = L$ reduces to $r^2 + \gamma r = 0$ with two roots 0 and $-\gamma$ regardless of the choice of β. Thus we have monotone convergence for this (dth) mode of the dynamical system [21, Sec. 17.1]:

$$w_{k,d} = 0^k + c_d(-\gamma)^k, \quad v_{k,d} = 0^k + \hat{c}_d(-\gamma)^k, \qquad (18)$$

where c_d and $\hat{c}_d$ are constants depending on initial conditions (17). Substituting $w_{1,d} = 0$ and $v_{1,d} = -(\beta+\gamma)v_{0,d}$ (17) into (16) yields

$$c_d = 0, \quad \hat{c}_d = v_{0,d}(1+\beta/\gamma), \qquad (19)$$

illustrating that the primary sequence $\{w_{k,d}\}$ reaches its optimum after one iteration, whereas the secondary sequence $\{v_{k,d}\}$ has slow monotone convergence of the distance to the optimum, while exhibiting undesirable oscillation due to the term $(-\gamma)^k$, corresponding to overshooting over the optimum.

2. $\lambda_i < L$: In (15) we found the optimal overall $\beta^\star$ for AFM when $\alpha = 1/L$. One can alternatively explore what the best value of β would be for any given mode of the system for comparison. Polynomial (10) has repeated roots for the following β, corresponding to the smaller zero of the discriminant $\Delta(1/L, \beta, \gamma, \lambda_i)$ for given γ and λ_i:

$$\beta_i^\star(\gamma) := \left(1 - \sqrt{(1+\gamma)\lambda_i/L}\right)^2 / (1 - \lambda_i/L). \tag{20}$$

This choice satisfies $\beta^\star = \beta^\star(\gamma^\star) = \beta_1^\star(\gamma^\star)$ (15), because λ_1 is the smallest eigenvalue. Next we examine the convergence behavior of AFM with $\alpha = 1/L$ and $\gamma > 0$ in the following three regimes, similar to AFM with $\alpha = 1/L$ and $\gamma = 0$ in [5, Sec. 4.3]:[5]
  – $\beta < \beta_i^\star(\gamma)$: low momentum, over-damped,
  – $\beta = \beta_i^\star(\gamma)$: optimal momentum, critically damped,
  – $\beta > \beta_i^\star(\gamma)$: high momentum, under-damped.

If $\beta \le \beta_i^\star(\gamma)$, polynomial (10) has two real roots, $r_{1,i}$ and $r_{2,i}$, where we omit $(1/L, \beta, \gamma, \lambda_i)$ in $r_{.,i} = r_.(1/L, \beta, \gamma, \lambda_i)$ for simplicity. Then, the system evolves as [21, Sec. 17.1]:

$$w_{k,i} = c_{1,i} r_{1,i}^k + c_{2,i} r_{2,i}^k, \quad v_{k,i} = \hat{c}_{1,i} r_{1,i}^k + \hat{c}_{2,i} r_{2,i}^k, \tag{21}$$

where constants $c_{1,i}$, $c_{2,i}$, $\hat{c}_{1,i}$, and $\hat{c}_{2,i}$ depend on initial conditions (17). In particular, when $\beta = \beta_i^\star(\gamma)$ (20), we have the repeated root:

$$r_i^\star(\gamma) := 1 - \sqrt{(1+\gamma)\lambda_i/L}, \tag{22}$$

corresponding to critical damping, yielding the fastest monotone convergence among (21) for any β s.t. $\beta \le \beta_i^\star(\gamma)$. This property is due to the quasi-convexity of $\rho(T_{\lambda_i}(1/L, \beta, \gamma))$ over β. If $\beta < \beta_i^\star(\gamma)$, the system is over-damped, which corresponds to the low-momentum regime, where the system is dominated by the larger root that is greater than $r_i^\star(\gamma)$ (22), and thus has slow monotone convergence. However, depending on initial conditions (17), the system may only be dominated by the smaller root, as noticed for the case $\lambda_i = L$ in (18) and (19). Also note that the mode of $\lambda_i = L$ is always in the low-momentum regime regardless of the value of β.

If $\beta > \beta_i^\star(\gamma)$, the system is under-damped, which corresponds to the high-momentum regime. This means that the system evolves as [21, Sec. 17.1]:

$$w_{k,i} = c_i \left(\sqrt{\beta(1 - \lambda_i/L)}\right)^k \cos(k\psi_i(\beta, \gamma) - \delta_i), \tag{23}$$

---

[5] For simplicity in the momentum analysis, we considered values β within [0 1], containing the standard β_k values in Tables 1 and 2. This restriction excludes the effect of the β that corresponds to the larger zero of the discriminant $\Delta(1/L, \beta, \gamma, \lambda_i)$ for given γ and λ_i and that is larger than 1. Any β greater than 1 has $\rho(T_{\lambda_i}(1/L, \beta, \gamma))$ values (in (9) with $\alpha = 1/L$) that are larger than those for $\beta \in [\beta_i^\star(\gamma) \; 1]$ due to the quasi-convexity of $\rho(T_{\lambda_i}(1/L, \beta, \gamma))$ over β.

$$v_{k,i} = \hat{c}_i \left(\sqrt{\beta(1 - \lambda_i/L)} \right)^k \cos(k\psi_i(\beta, \gamma) - \hat{\delta}_i),$$

where the frequency of the oscillation is given by

$$\psi_i(\beta, \gamma) := \cos^{-1} \left(((1 + \beta)(1 - \lambda_i/L) - \gamma\lambda_i/L)/\left(2\sqrt{\beta(1 - \lambda_i/L)}\right) \right), \quad (24)$$

and $c_i, \delta_i, \hat{c}_i$, and $\hat{\delta}_i$ denote constants that depend on initial conditions (17); in particular for $\beta \approx 1$, we have $\delta_i \approx 0$ and $\hat{\delta}_i \approx 0$ so we will ignore them.

Based on the above momentum analysis, we categorize the behavior of the ith mode of AFM for each λ_i in Fig. 1. Regimes with two curves and one curve (over λ) in Fig. 1 correspond to the low- and high-momentum regimes, respectively. In particular, for $\beta = \beta^\star(\gamma)$ in the left plot of Fig. 1, most λ_i values (satisfying $\beta > \beta_i^\star(\gamma)$) experience high momentum [and the optimal momentum for λ_i satisfying $\beta^\star(\gamma) = \beta_i^\star(\gamma)$, e.g., $\lambda_i = \mu$], whereas modes where $\lambda_i \approx L$ experience low momentum. The fast convergence of the primary sequence $\{w_{k,d}\}$ in (18) and (19) generalizes to the case $\lambda_i \approx L$, corresponding to the lower curves in Fig. 1. In addition, for $\beta = 0, 0.2$ that are smaller than $\beta^\star(\gamma)$ in the right plot of Fig. 1, both $\lambda \approx \mu$ and $\lambda \approx L$ experience low momentum so increasing β improves the convergence rate.

Based on the quadratic analysis in this section, we would like to use appropriately large β and γ coefficients, namely $(\beta^\star, \gamma^\star)$, to have fast monotone convergence (for the dominating modes). However, such values require knowing the function parameter $q = \mu/L$ that is usually unavailable in practice. Using OGM (and OGM′) in Table 1 without knowing q will likely lead to oscillation due to the high momentum (or underdamping) for strongly convex functions. The next section describes restarting schemes inspired by [5] that we suggest to use with OGM to avoid such oscillation and thus heuristically accelerate the rate of OGM for a strongly convex quadratic function and even for a convex function that is locally strongly convex.

4 Restarting Schemes

Restarting an algorithm (i.e., starting the algorithm again by using the current iterate as the new starting point) after a certain number of iterations or when some restarting condition is satisfied has been found useful, e.g., for the conjugate gradient method [22,23], called "fixed restart" and "adaptive restart," respectively. The fixed restart approach was also studied for accelerated gradient schemes such as FGM in [16,24]. Recently adaptive restart of FGM was shown to provide dramatic practical acceleration without requiring knowledge of function parameters [5–7]. Building upon those ideas, this section reviews and applies restarting approaches for OGM. A quadratic analysis in [5] justified using a restarting condition for FGM; this section extends that analysis to OGM by studying an observable quantity of oscillation that serves as an indicator for restarting the momentum of OGM.

4.1 Fixed Restart

Restarting an algorithm every certain number of iterations can yield a linear rate for decreasing a function in $\mathscr{F}_{\mu,L}(\mathbb{R}^d)$ [16, Sec. 5.1] [24, Sec. 11.4]. Suppose one restarts OGM every k (inner) iterations by initializing the $(j+1)$th outer iteration using $\boldsymbol{x}_{j+1,0} = \boldsymbol{x}_{j,k}$, where $\boldsymbol{x}_{j,i}$ denotes an iterate at the jth outer iteration and ith inner iteration. Combining the OGM rate in Table 1 and strong convexity inequality (3) yields the following rate for each outer iteration of OGM with fixed restart:

$$f(\boldsymbol{x}_{j,k}) - f(\boldsymbol{x}_*) \leq \frac{L||\boldsymbol{x}_{j,0} - \boldsymbol{x}_*||^2}{k^2} \leq \frac{2L}{\mu k^2}(f(\boldsymbol{x}_{j,0}) - f(\boldsymbol{x}_*)). \qquad (25)$$

This rate is faster than the $4L/\mu k^2$ rate of one outer iteration of FGM with fixed restart (using the FGM rate in Table 1). For a given $N = jk$ total number of steps, a simple calculation shows that the optimal restarting interval k minimizing the rate $\left(2L/(\mu k^2)\right)^j$ after N steps [owing from (25)] is $k_{\text{fixed}} := e\sqrt{2/q}$ that does not depend on N, where e is Euler's number.

There are two drawbacks of the fixed restart approach [5, Sec. 3.1]. First, computing the optimal interval k_{fixed} requires knowledge of q that is usually unavailable in practice.[6] Second, using a global parameter q may be too conservative when the iterates enter locally strongly convex region. (As noted in footnote 1, a backtracking scheme for estimating μ in [16, Sec. 5.3] could be used here to somewhat circumvent the above two drawbacks as in [26, Sec. 3], which can be viewed as an adaptive approach that is different from the ones in Sect. 4.2.) Therefore, adaptive restarting in [5] is more useful in practice, which we review next and then apply to OGM.

4.2 Adaptive Restart

To circumvent the drawbacks of fixed restart, [5] proposes the following two adaptive restart schemes for FGM:

– Function scheme for restarting (FR): Restart whenever

$$f(\boldsymbol{y}_{k+1}) > f(\boldsymbol{y}_k), \qquad (26)$$

– Gradient scheme for restarting (GR): Restart whenever

$$\langle -\nabla f(\boldsymbol{x}_k), \ \boldsymbol{y}_{k+1} - \boldsymbol{y}_k \rangle < 0. \qquad (27)$$

---

[6] The choice of the restarting interval k can be relaxed, and any k that is greater than $\sqrt{2/q}$ guarantees a linear rate in (25) for OGM with fixed restart (and similarly for the analogous version of FGM). However, such choice of k still requires knowledge of q. This drawback has been recently relieved for FGM with fixed restart in [25, Thm.1] [26, Thm. 1], exhibiting a linear rate with any restarting interval k. Note that [25,26] use a local quadratic growth condition, i.e., $f(\boldsymbol{x}) \geq f(\boldsymbol{x}_*) + \frac{\mu}{2}||\boldsymbol{x} - \boldsymbol{x}_*||^2$ for all $\boldsymbol{x}$ with $\mu > 0$. This condition is implied by strong convexity condition (3), and one can notice that the second inequality of (25) is also implied by the local quadratic growth condition, without requiring stronger condition (3).

These schemes heuristically improve convergence rates of FGM, and both performed similarly well [5]. Although the function scheme guarantees monotonic decreasing function values, the gradient scheme has two advantages over the function scheme [5]; the gradient scheme involves only arithmetic operations with already computed quantities, and it is numerically more stable.

These two schemes encourage an algorithm to restart whenever the iterates take a "bad" direction, i.e., when the function value increases or the negative gradient and the momentum have an obtuse angle, respectively. However, a convergence proof that justifies their empirical acceleration is yet unknown, so [5] analyzes such restarting schemes for strongly convex quadratic functions. An alternative scheme in [7, Sec. 5] that restarts whenever the magnitude of the momentum decreases, i.e., $||\boldsymbol{y}_{k+1} - \boldsymbol{y}_k|| <$ $||\boldsymbol{y}_k - \boldsymbol{y}_{k-1}||$, has a theoretical convergence analysis for a *continuous* version of the method for the function class $\mathscr{F}_{\mu,L}(\mathbb{R}^d)$. In addition, another function restart condition in [25, Prop. 4] has some theoretical justification. However, empirically both function and gradient schemes (26) and (27) performed better in [7,25]. Thus, this paper focuses on adapting practical restart schemes (26) and (27) to OGM and extending the analysis in [5] to OGM. First we introduce a new additional adaptive scheme designed specifically for AFM with $\alpha = 1/L$ and $\gamma > 0$ (e.g., OGM).

4.3 Adaptive Decrease of γ for AFM with $\alpha = 1/L$ and $\gamma > 0$

Section 3.4 describes that the secondary sequence $\{\boldsymbol{x}_k\}$ of AFM with $\alpha = 1/L$ and $\gamma > 0$ (e.g., OGM) might experience overshoot and thus slow convergence, unlike its primary sequence $\{\boldsymbol{y}_k\}$, when the iterates enter a region where the mode of the largest eigenvalue dominates. (Section 6.1.2 illustrates such an example.) From (18), the overshoot of $\boldsymbol{x}_k$ has magnitude proportional to $|\gamma|$, yet a suitably large γ, such as $\gamma^\star$ (14), is essential for overall acceleration.

To avoid (or reduce) such overshooting, we suggest the following adaptive scheme:

– Gradient scheme for decreasing γ (GDγ): Decrease γ whenever

$$\langle \nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_{k-1}) \rangle < 0. \tag{28}$$

Because the primary sequence $\{\boldsymbol{y}_k\}$ of AFM with $\alpha = 1/L$ is unlikely to overshoot, one could choose to simply use the primary sequence $\{\boldsymbol{y}_k\}$ as algorithm output instead of the secondary sequence $\{\boldsymbol{x}_k\}$. However, if one needs to use the secondary sequence of AFM with $\alpha = 1/L$ and $\gamma > 0$ (e.g., Sect. 5.2), adaptive scheme (28) can help.

4.4 Observable AFM Quantities When $\alpha = 1/L$

This section revisits Sect. 3.4 that suggested that observing the evolution of the mode coefficients $\{w_{k,i}\}$ and $\{v_{k,i}\}$ can help identify the momentum regime. However, in practice that evolution is unobservable because the optimum $\boldsymbol{x}_*$ is unknown, whereas Sect. 3.4 assumes $\boldsymbol{x}_* = \boldsymbol{0}$. Instead we can observe the evolution of the function values,

which are related to the mode coefficients as follows:

$$f(\mathbf{y}_k) = \frac{1}{2} \sum_{i=1}^{d} \lambda_i w_{k,i}^2, \quad f(\mathbf{x}_k) = \frac{1}{2} \sum_{i=1}^{d} \lambda_i v_{k,i}^2, \tag{29}$$

and also the inner products of the gradient and momentum, i.e.,

$$\langle -\nabla f(\mathbf{x}_k), \ \mathbf{y}_{k+1} - \mathbf{y}_k \rangle = - \sum_{i=1}^{d} \lambda_i v_{k,i} (w_{k+1,i} - w_{k,i}), \tag{30}$$

$$\langle \nabla f(\mathbf{x}_k), \ \nabla f(\mathbf{x}_{k-1}) \rangle = \sum_{i=1}^{d} \lambda_i^2 v_{k,i} v_{k-1,i}. \tag{31}$$

These quantities appear in the conditions for adaptive schemes (26), (27), and (28).

One would like to increase β and γ as much as possible for acceleration up to $\beta^\star$ and $\gamma^\star$ (15). However, without knowing q (and $\beta^\star, \gamma^\star$), using large β and γ could end up placing the majority of the modes in the high-momentum regime, eventually leading to slow convergence with oscillation as described in Sect. 3.4. To avoid such oscillation, we hope to detect it using (29) and (30) and restart the algorithm. We also hope to detect overshoot (18) of the modes of the large eigenvalues (in the low-momentum regime) using (31) so that we can then decrease γ and avoid such overshoot.

The rest of this section focuses on the case where $\beta > \beta_1(\gamma)$ for given γ, when the most of the modes are in the high-momentum regime. Because the maximum of $\rho(\mathbf{T}_\lambda(1/L, \beta, \gamma))$ occurs at the points $\lambda = \mu$ or $\lambda = L$, we expect that (29), (30), and (31) will be quickly dominated by the mode of the smallest or the largest eigenvalues. Specifically, plugging $w_{k,i}$ and $v_{k,i}$ in (18), (19), and (23) to (29), (30), and (31) for only the (dominating) mode of the smallest and the largest eigenvalues ($\lambda_1 = \mu$ and $\lambda_d = L$, respectively) leads to the following approximations:

$$f(\mathbf{y}_k) \approx \frac{1}{2} \mu c_1^2 \beta^k (1 - \mu/L)^k \cos^2(k\psi_1), \tag{32}$$

$$f(\mathbf{x}_k) \approx \frac{1}{2} \mu \hat{c}_1^2 \beta^k (1 - \mu/L)^k \cos^2(k\psi_1) + \frac{1}{2} L \hat{c}_d^2 \gamma^{2k}$$

$$\langle -\nabla f(\mathbf{x}_k), \ \mathbf{y}_{k+1} - \mathbf{y}_k \rangle \approx - \mu c_1 \hat{c}_1 \beta^k (1 - \mu/L)^k \cos(k\psi_1)$$
$$\times \left(\sqrt{\beta(1 - \mu/L)} \cos((k+1)\psi_1) - \cos(k\psi_1) \right),$$

$$\langle \nabla f(\mathbf{x}_k), \ \nabla f(\mathbf{x}_{k-1}) \rangle \approx \mu^2 \hat{c}_1^2 \beta^{k-\frac{1}{2}} (1 - \mu/L)^{k-\frac{1}{2}} \cos(k\psi_1) \cos((k-1)\psi_1)$$
$$- L^2 \hat{c}_d^2 \gamma^{2k-1},$$

where $\psi_1 = \psi_1(\beta, \gamma)$ in (24). Furthermore, it is likely that these expressions will be dominated by the mode of either the smallest or largest eigenvalues, so we next analyze each case separately.

4.4.1 Case 1: The Mode of the Smallest Eigenvalue Dominates

When the mode of the smallest eigenvalue dominates, we further approximate (32) as

$$
\begin{aligned}
f(\boldsymbol{y}_k) &\approx \frac{1}{2}\mu c_1^2 \,\beta^k \,(1 - \mu/L)^k \,\cos^2(k\psi_1),\\
f(\boldsymbol{x}_k) &\approx \frac{1}{2}\mu \hat{c}_1^2 \,\beta^k \,(1 - \mu/L)^k \,\cos^2(k\psi_1),\\
\langle -\nabla f(\boldsymbol{x}_k), \, &\boldsymbol{y}_{k+1} - \boldsymbol{y}_k \rangle \\
&\approx -\mu c_1 \hat{c}_1 \,\beta^k \,(1 - \mu/L)^k \,\cos(k\psi_1) \,(\cos((k + 1)\psi_1) - \cos(k\psi_1)) \\
&= 2\mu c_1 \hat{c}_1 \,\beta^k \,(1 - \mu/L)^k \,\cos(k\psi_1) \sin((k + 1/2)\psi_1) \,\sin(\psi_1/2) \\
&\approx 2\mu c_1 \hat{c}_1 \,\sin(\psi_1/2) \,\beta^k \,(1 - \mu/L)^k \,\sin(2k\psi_1),
\end{aligned}
\tag{33}
$$

using simple trigonometric identities and the approximations $\sqrt{\beta(1 - \mu/L)} \approx 1$ and $\sin(k\psi_1) \approx \sin((k + 1/2)\psi_1)$ for small μ (leading to small ψ_1 in (24)). Values (33) exhibit oscillations at a frequency proportional to $\psi_1(\beta, \gamma)$ in (24). This oscillation can be detected by conditions (26) and (27) and is useful in detecting the high-momentum regime where a restart can help improve the convergence rate.

4.4.2 Case 2: The Mode of the Largest Eigenvalue Dominates

Unlike the primary sequence $\{\boldsymbol{y}_k\}$ of AFM with $\alpha = 1/L$ (e.g., OGM), convergence of its secondary sequence $\{\boldsymbol{x}_k\}$ may be dominated by the mode of the largest eigenvalue in (18) and (19). By further approximating (32) for the case when the mode of the largest eigenvalue dominates, the function value $f(\boldsymbol{x}_k) \approx \frac{1}{2}L\hat{c}_d^2 \gamma^{2k}$ decreases slowly but monotonically, whereas $f(\boldsymbol{y}_k) \approx f(\boldsymbol{x}_*) = 0$ and $\langle -\nabla f(\boldsymbol{x}_k), \, \boldsymbol{y}_{k+1} - \boldsymbol{y}_k \rangle \approx 0$. Therefore, neither restart condition (26) nor (27) can detect such nonoscillatory observable values, even though the secondary mode $\{w_{k,d}\}$ of the largest eigenvalue is oscillating (corresponding to overshooting over the optimum). However, the inner product of two sequential gradients:

$$
\langle \nabla f(\boldsymbol{x}_k), \, \nabla f(\boldsymbol{x}_{k-1}) \rangle \approx -L^2 \hat{c}_d^2 \gamma^{2k-1}
\tag{34}
$$

can detect the overshoot of the secondary sequence $\{\boldsymbol{x}_k\}$, suggesting that the algorithm should adapt by decreasing γ when condition (28) holds. Decreasing γ too much may slow down the overall convergence rate when the mode of the smallest eigenvalue is not negligible. Thus, we use (28) only when using the secondary sequence $\{\boldsymbol{x}_k\}$ as algorithm output (e.g., Sect. 5.2).

5 Proposed Adaptive Schemes for OGM

5.1 Adaptive Scheme of OGM for Smooth and Strongly Convex Problems

Algorithm 2 illustrates a new adaptive version of OGM$'$ (rather than OGM)[7] that is used in our numerical experiments in Sect. 6. When a restart condition is satisfied in Algorithm 2, we reset $t_k = 1$ to discard the previous momentum that has a bad direction. When the decreasing γ condition is satisfied in Algorithm 2, we decrease σ to suppress undesirable overshoot of the secondary sequence $\{x_k\}$. Although the analysis in Sect. 3 considered only strongly convex quadratic functions, the numerical experiments in Sect. 6 illustrate that the adaptive scheme is also useful more generally for smooth convex functions in $\mathscr{F}_{0,L}(\mathbb{R}^d)$, as described in [5, Sec. 4.6].

Algorithm 2 OGM$'$ with restarting momentum and decreasing γ

1: **Input:** $f \in \mathscr{F}_{\mu,L}(\mathbb{R}^d)$ or $\mathscr{F}_{0,L}(\mathbb{R}^d)$, $x_{-1} = x_0 = y_0 \in \mathbb{R}^d$, $t_0 = \sigma = 1$, $\bar{\sigma} \in [0, 1]$.
2: **for** $k \geq 0$ **do**
3:     $y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$
4:     **if** $f(y_{k+1}) > f(y_k)$ (or $\langle -\nabla f(x_k),\ y_{k+1} - y_k \rangle < 0$) **then**   ▷ Restart condition
5:         $t_k = 1, \sigma \leftarrow 1$
6:     **else if** $\langle \nabla f(x_k),\ \nabla f(x_{k-1}) \rangle < 0$ **then**   ▷ Decreasing γ condition
7:         $\sigma \leftarrow \bar{\sigma}\sigma$
8:     $t_{k+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right)$
9:     $x_{k+1} = y_{k+1} + \frac{t_k - 1}{t_{k+1}}(y_{k+1} - y_k) + \sigma\frac{t_k}{t_{k+1}}(y_{k+1} - x_k)$

5.2 Adaptive Scheme of a Proximal Version of OGM for Nonsmooth Composite Convex Problems

Modern applications often involve nonsmooth composite convex problems:

$$\min_{x \in \mathbb{R}^d} \{F(x) := f(x) + \phi(x)\}, \tag{35}$$

where $f \in \mathscr{F}_{0,L}(\mathbb{R}^d)$ is a smooth convex function (typically not strongly convex) and $\phi \in \mathscr{F}_{0,\infty}(\mathbb{R}^d)$ is a convex function that is possibly nonsmooth and "proximal-friendly" [27], such as the ℓ_1 regularizer $\phi(x) = ||x||_1$. Our numerical experiments in

---

[7] OGM requires choosing the number of iterations N in advance for computing θ_N in Table 1, which seems incompatible with adaptive restarting schemes. In contrast, the parameters t_k in Table 1 and Algorithm 2 are independent of N. The fact that θ_N is larger than t_N at the last (Nth) iteration helps to dampen (by reducing the values of β and γ) the final update to guarantee a faster (optimal) worst-case rate for the last secondary iterate x_N. This property was studied in [14]. We could perform one last update using θ_N after a restart condition is satisfied, but this step appears unnecessary because restarting already has the effect of dampening (reducing β and γ). Thus, Algorithm 2 uses OGM$'$ instead that uses t_k and that has a worst-case rate that is similar to that of OGM.

Sect. 6 show that a new adaptive version of a proximal variant of OGM can be useful for solving such problems.

To solve (35), [4] developed a fast proximal gradient method, popularized under the name fast iterative shrinkage-thresholding algorithm (FISTA). FISTA has the same rate as FGM in Table 1 for solving (35), by simply replacing the line 3 of Algorithm 1 with FGM coefficients by $\boldsymbol{y}_{k+1} = \text{prox}_{\alpha\phi}(\boldsymbol{x}_k - \alpha\nabla f(\boldsymbol{x}_k))$, where the proximity operator is defined as $\text{prox}_h(\boldsymbol{z}) := \arg\min \boldsymbol{x} \in \mathbb{R}^d\{\frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 + h(\boldsymbol{x})\}$. Variants of FISTA with adaptive restart are studied in [5, Sec. 5.2].

Inspired by the fact that OGM has a worst-case rate faster than FGM, [15] studied a proximal variant[8] of OGM (POGM)[9] by using variations of any (or all) of three adaptive schemes (26), (27), (28), as illustrated in Algorithm 3. Regarding a function restart condition for POGM, we use $F(\boldsymbol{x}_{k+1}) > F(\boldsymbol{x}_k)$, instead of $F(\boldsymbol{y}_{k+1}) > F(\boldsymbol{y}_k)$, because $F(\boldsymbol{y}_k)$ can be unbounded (e.g., $\boldsymbol{y}_k$ can be unfeasible for constrained problems). For gradient conditions of POGM, we consider the composite gradient mapping $G(\boldsymbol{x}_k) \in \nabla f(\boldsymbol{x}_k) + \partial\phi(\boldsymbol{x}_{k+1})$ in Algorithm 3 that differs from the standard composite gradient mapping in [16]. We then use the gradient conditions

$$\langle -G(\boldsymbol{x}_k), \ \boldsymbol{y}_{k+1} - \boldsymbol{y}_k \rangle < 0, \quad \langle G(\boldsymbol{x}_k), \ G(\boldsymbol{x}_{k-1}) \rangle < 0 \qquad (36)$$

for restarting POGM or decreasing γ of POGM, respectively. Here POGM must output the secondary sequence $\{\boldsymbol{x}_k\}$ because the function value $F(\boldsymbol{y}_k)$ of the primary sequence may be unbounded. This situation was the motivation for (28) [and the second inequality of (36)] and Sect. 4.3. When $\phi(\boldsymbol{x}) = 0$, Algorithm 3 reduces to an algorithm that is similar to Algorithm 2, where only the location of the restart and decreasing γ conditions differs.

6 Numerical Results

This section shows the results of applying OGM′ and POGM′ with adaptive schemes in Algorithms 2 and 3 to various numerical examples including both strongly convex quadratic problems and nonstrongly convex problems.[10] (For simplicity, we omit the prime symbol of OGM′ and POGM′ with adaptive restart hereafter.) The results illustrate that OGM (or POGM) with adaptive schemes converges faster than FGM

---

[8] Applying the proximity operator to the primary sequence $\{\boldsymbol{y}_k\}$ of OGM, similar to the extension of FGM to FISTA, leads to a poor worst-case rate [15]. Therefore, [15] applied the proximity operator to the secondary sequence of OGM and showed numerically that this version has a worst-case rate about twice faster than that of FISTA.

[9] Like OGM, POGM in [15, Sec. 4.3] requires choosing the number of iterations N in advance for computing θ_N, and this is incompatible with adaptive restarting schemes. Therefore, analogous to using OGM′ instead of OGM for an adaptive scheme in Algorithm 2 (see footnote 7), Algorithm 3 uses a proximal version of OGM′ (rather than the POGM in [15]) with restart. An extension of OGM′ (without restart) to a proximal version with a fast worst-case rate is unknown yet

[10] Software for the algorithms and for producing the figures in Sect. 6 is available at https://gitlab.eecs.umich.edu/michigan-fast-optimization/ogm-adaptive-restart.

Algorithm 3 POGM$'$ with restarting momentum and decreasing γ

1: **Input:** $f \in \mathscr{F}_{0,L}(\mathbb{R}^d)$, $\phi \in \mathscr{F}_{0,\infty}(\mathbb{R}^d)$, $\boldsymbol{x}_{-1} = \boldsymbol{x}_0 = \boldsymbol{y}_0 = \boldsymbol{u}_0 = \boldsymbol{z}_0 \in \mathbb{R}^d$,
2:         $t_0 = \zeta_0 = \sigma = 1$, $\bar{\sigma} \in [0, \ 1]$.
3: **for** $k \geq 0$ **do**
4:      $\boldsymbol{u}_{k+1} = \boldsymbol{x}_k - \frac{1}{L} \nabla f(\boldsymbol{x}_k)$
5:      $t_{k+1} = \frac{1}{2} \left(1 + \sqrt{1 + 4t_k^2} \right)$
6:      $\boldsymbol{z}_{k+1} = \boldsymbol{u}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{u}_{k+1} - \boldsymbol{u}_k) + \sigma \frac{t_k}{t_{k+1}}(\boldsymbol{u}_{k+1} - \boldsymbol{x}_k) - \frac{t_k - 1}{t_{k+1}} \frac{1}{L\zeta_k}(\boldsymbol{x}_k - \boldsymbol{z}_k)$
7:      $\zeta_{k+1} = \frac{1}{L} \left(1 + \frac{t_k - 1}{t_{k+1}} + \sigma \frac{t_k}{t_{k+1}} \right)$
8:      $\boldsymbol{x}_{k+1} = \text{prox}_{\zeta_{k+1}\phi}(\boldsymbol{z}_{k+1})$
9:      $G(\boldsymbol{x}_k) = \nabla f(\boldsymbol{x}_k) - \frac{1}{\zeta_{k+1}}(\boldsymbol{x}_{k+1} - \boldsymbol{z}_{k+1})$
10:      $\boldsymbol{y}_{k+1} = \boldsymbol{x}_k - \frac{1}{L} G(\boldsymbol{x}_k)$
11:      **if** $F(\boldsymbol{x}_{k+1}) > F(\boldsymbol{x}_k)$ (or $\langle -G(\boldsymbol{x}_k), \ \boldsymbol{y}_{k+1} - \boldsymbol{y}_k \rangle < 0$) **then**   ▷ Restart condition
12:          $t_{k+1} = 1, \sigma \leftarrow 1$
13:      **else if** $\langle G(\boldsymbol{x}_k), \ G(\boldsymbol{x}_{k-1}) \rangle < 0$ **then**   ▷ Decreasing γ condition
14:          $\sigma \leftarrow \bar{\sigma}\sigma$

(or FISTA) with adaptive restart. The plots show the decrease of $F(\boldsymbol{y}_k)$ of the primary sequence for FGM (FISTA) and OGM unless specified. For POGM, we use the secondary sequence $\{\boldsymbol{x}_k\}$ as an output and plot $F(\boldsymbol{x}_k)$, since $F(\boldsymbol{y}_k)$ can be unbounded.

6.1 Strongly Convex Quadratic Examples

This section considers two types of strongly convex quadratic examples, where the mode of either the smallest eigenvalue or the largest eigenvalue dominates, providing examples of the analysis in Sects. 4.4.1 and 4.4.2, respectively.

6.1.1 Case 1: The Mode of the Smallest Eigenvalue Dominates

Figure 2 compares GM, FGM, and OGM, with or without the knowledge of q, for minimizing strongly convex quadratic function (4) in $d = 500$ dimensions with $q = 10^{-4}$, where we generated A (for $Q = A^\top A$) and p randomly. As expected, knowing q accelerates convergence.

Figure 2 also illustrates that adaptive restart helps FGM and OGM to nearly achieve the fast linear convergence rate of their nonadaptive versions that know q. As expected, OGM variants converge faster than FGM variants for all cases. In Fig. 2, "FR" and "GR" stand for function restart (26) and gradient restart (27), respectively, and both behave nearly the same.

6.1.2 Case 2: The Mode of the Largest Eigenvalue Dominates

Consider the strongly convex quadratic function with $Q = \begin{bmatrix} q & 0 \\ 0 & 1 \end{bmatrix}$, $q = 0.01$, $\boldsymbol{p} = \boldsymbol{0}$ and $\boldsymbol{x}_* = \boldsymbol{0}$. When starting the algorithm from the initial point $\boldsymbol{x}_0 = (0.2, \ 1)$, the
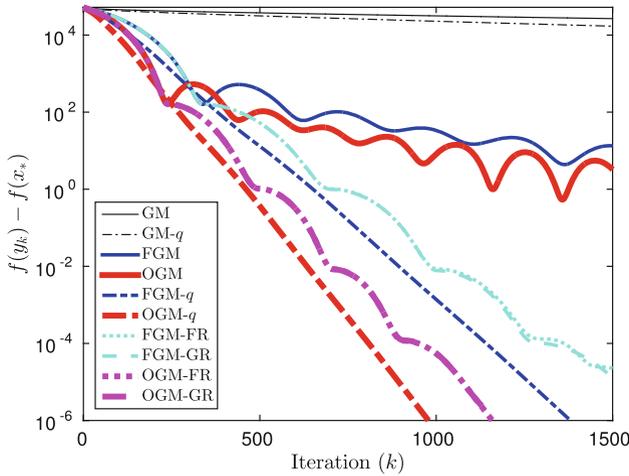
Fig. 2 Minimizing a strongly convex quadratic function— Case 1: The mode of the smallest eigenvalue dominates. (FGM-FR and FGM-GR are almost indistinguishable, as are OGM-FR and OGM-GR.)

secondary sequence $\{\boldsymbol{x}_k\}$ of OGM-GR[11] (or equivalently OGM-GR-GDγ ($\bar{\sigma} = 1.0$)) is dominated by the mode of largest eigenvalue in Fig. 3, illustrating the analysis of Sect. 4.4.2. Figure 3 illustrates that the primary sequence of OGM-GR converges faster than that of FGM-GR, whereas the secondary sequence of OGM-GR initially converges even slower than GM. To deal with such slow convergence coming from the overshooting behavior of the mode of the largest eigenvalue of the secondary sequence of OGM, we employ the decreasing γ scheme in (28). Figure 3 shows that using $\bar{\sigma} < 1$ in Algorithm 2 leads to overall faster convergence of the secondary sequence $\{\boldsymbol{x}_k\}$ than the standard OGM-GR where $\bar{\sigma} = 1$. We leave optimizing the choice of $\bar{\sigma}$ or studying other strategies for decreasing γ as future work.

6.2 Nonstrongly Convex Examples

This section applies adaptive OGM (or POGM) to two nonstrongly convex numerical examples in [5,7].[12] The numerical results show that adaptive OGM (or POGM) converges faster than FGM (or FISTA) with adaptive restart.

---

[11] Figure 3 only compares the results of the gradient restart (GR) scheme for simplicity, where the function restart (FR) behaves similarly.

[12] Additional numerical result can be found in [28].
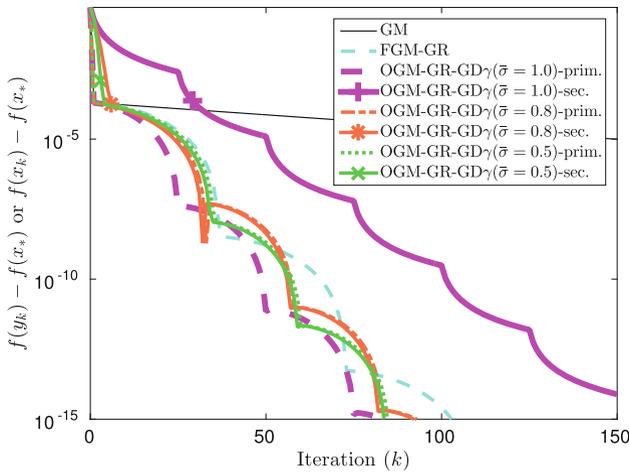
Fig. 3 Minimizing a strongly convex quadratic function—Case 2: The mode of the largest eigenvalue dominates for the secondary sequence $\{x_k\}$ of OGM. Using GDγ (28) with $\bar{\sigma} < 1$ accelerates convergence of the secondary sequence of OGM-GR, where both the primary and secondary sequences behave similarly after first few iterations, unlike $\bar{\sigma} = 1$

6.2.1 Log-Sum-Exp

The following function from [5] is smooth but nonstrongly convex:

$$ f(\boldsymbol{x}) = \eta \log \left(\sum_{i=1}^{m} \exp \left(\frac{1}{\eta} (\boldsymbol{a}_i^\top \boldsymbol{x} - b_i) \right) \right). $$

It approaches $\max_{i=1,\ldots,m} (\boldsymbol{a}_i^\top \boldsymbol{x} - b_i)$ as $\eta \to 0$. Here, η controls the function smoothness $L = \frac{1}{\eta} \lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})$ where $\boldsymbol{A} = [\boldsymbol{a}_1 \cdots \boldsymbol{a}_m]^\top \in \mathbb{R}^{m \times d}$. The region around the optimum is approximately quadratic since the function is smooth, and thus the adaptive restart can be useful without knowing the local condition number.

For $(m, d) = (100, 20)$, we randomly generated $\boldsymbol{a}_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$ for $i = 1, \ldots, m$, and investigated $\eta = 1, 10$. Figure 4 shows that OGM with adaptive restart converges faster than FGM with the adaptive restart. The benefit of adaptive restart is dramatic here; apparently FGM and OGM enter a locally strongly convex region after about 100–200 iterations, where adaptive restart then provides a fast linear rate.

6.2.2 Sparse Linear Regression

Consider the following cost function used for sparse linear regression:

$$ f(\boldsymbol{x}) = \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2, \quad \phi(\boldsymbol{x}) = \tau \|\boldsymbol{x}\|_1, $$
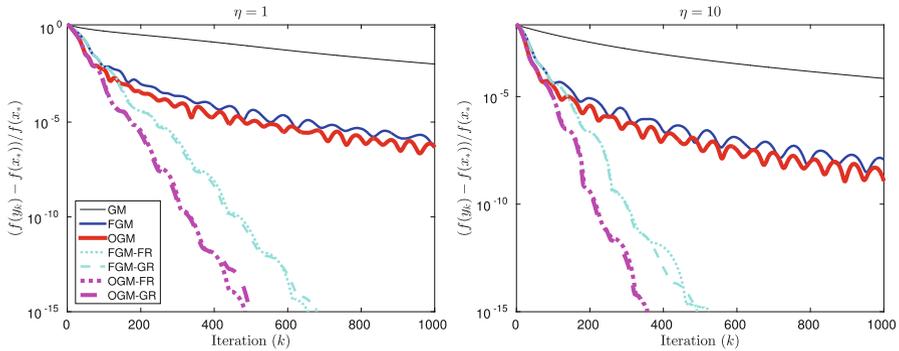
Fig. 4 Minimizing a smooth but nonstrongly convex Log-Sum-Exp function
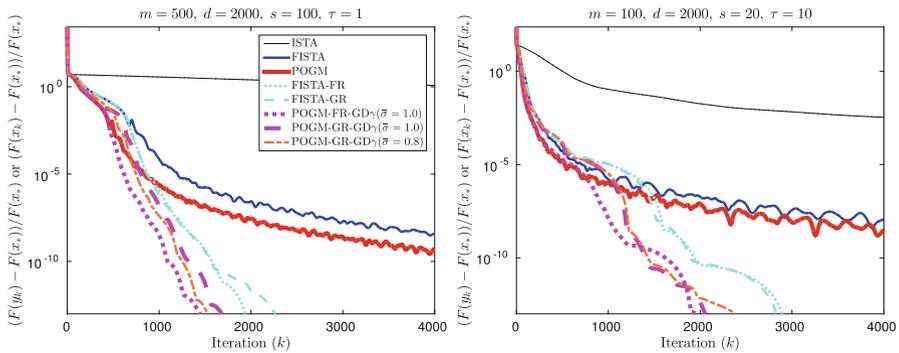
Fig. 5 Solving a sparse linear regression problem. (ISTA is a proximal variant of GM.)

for $A \in \mathbb{R}^{m \times d}$, where $L = \lambda_{\max}(A^\top A)$ and the parameter τ balances between the measurement error and signal sparsity. The proximity operator becomes a soft-thresholding operator, e.g., $\text{prox}_{\zeta_{k+1}\phi}(x) = \text{sgn}(x) \odot \max\{|x| - \zeta_{k+1}\tau, 0\}$, where $\odot$ is an element-wise multiplication. The minimization seeks a sparse solution x_*, and often the cost function is strongly convex with respect to the nonzero elements of x_*. Thus we expect to benefit from adaptive restarting.

For each choice of (m, d, s, τ) in Fig. 5, we generated an s-sparse true vector x_{true} by taking the s largest entries of a randomly generated vector. We then simulated $b = Ax_{\text{true}} + \varepsilon$, where the entries of matrix A and vector ε were sampled from a zero-mean normal distribution with variances 1 and 0.1, respectively. Figure 5 illustrates that POGM with adaptive schemes provides acceleration over FISTA with adaptive restart. While Sect. 3.4 discusses the undesirable overshooting behavior that a secondary sequence of OGM (or POGM) may encounter, these examples rarely encountered such behavior. Therefore, the choice of $\bar{\sigma}$ in the adaptive POGM was not significant in this experiment, unlike Sect. 6.1.2.

7 Conclusions

We introduced adaptive restarting schemes for the optimized gradient method (OGM) to heuristically provide a fast linear convergence rate, when the function is strongly convex or even when the function is not globally strongly convex. The method resets the momentum, when it makes a bad direction. We provided a heuristic dynamical system analysis to justify the practical acceleration of the adaptive scheme of OGM, by extending the existing analysis of the fast gradient method (FGM). On the way, we described a new accelerated gradient method named OGM-q for strongly convex quadratic problems. Numerical results illustrate that the proposed adaptive approach practically accelerates the convergence rate of OGM and, in particular, performs faster than FGM with adaptive restart. An interesting open problem is to determine the worst-case rates for OGM (and FGM) with adaptive restart.

References

1. Cevher, V., Becker, S., Schmidt, M.: Convex optimization for big data: scalable, randomized, and parallel algorithms for big data analytics. IEEE Signal Process. Mag. **31**(5), 32–43 (2014). https://doi.org/10.1109/MSP.2014.2329397
2. 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)
3. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer, Boston (2004). https://doi.org/10.1007/978-1-4419-8853-9
4. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009). https://doi.org/10.1137/080716542
5. O'Donoghue, B., Candès, E.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**(3), 715–32 (2015). https://doi.org/10.1007/s10208-013-9150-3
6. Giselsson, P., Boyd, S.: Monotonicity and restart in fast gradient methods. In: Proceedings of Conference on Decision and Control, pp. 5058–5063 (2014). https://doi.org/10.1109/CDC.2014.7040179
7. Su, W., Boyd, S., Candès, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J. Mach. Learn. Res. **17**(153), 1–43 (2016)
8. Muckley, M.J., Noll, D.C., Fessler, J.A.: Fast parallel MR image reconstruction via B1-based, adaptive restart, iterative soft thresholding algorithms (BARISTA). IEEE Trans. Med. Imaging **34**(2), 578–88 (2015). https://doi.org/10.1109/TMI.2014.2363034
9. Monteiro, R.D.C., Ortiz, C., Svaiter, B.F.: An adaptive accelerated first-order method for convex optimization. Comput. Optim. Appl. **64**(1), 31–73 (2016). https://doi.org/10.1007/s10589-015-9802-0
10. Kim, D., Fessler, J.A.: Optimized first-order methods for smooth convex minimization. Math. Program. **159**(1), 81–107 (2016). https://doi.org/10.1007/s10107-015-0949-3
11. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: a novel approach. Math. Program. **145**(1–2), 451–82 (2014). https://doi.org/10.1007/s10107-013-0653-0
12. Drori, Y.: The exact information-based complexity of smooth convex minimization. J. Complex. **39**, 1–16 (2017). https://doi.org/10.1016/j.jco.2016.11.001
13. Kim, D., Fessler, J.A.: Generalizing the optimized gradient method for smooth convex minimization (2016). arXiv:1607.06764
14. Kim, D., Fessler, J.A.: On the convergence analysis of the optimized gradient methods. J. Optim. Theory Appl. **172**(1), 187–205 (2017). https://doi.org/10.1007/s10957-016-1018-7
15. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Exact worst-case convergence rates of the proximal gradient method for composite convex minimization (2017). arXiv:1705.04398

16. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–61 (2013). https://doi.org/10.1007/s10107-012-0629-5

17. Chambolle, A., Dossal, C.: On the convergence of the iterates of the fast iterative shrinkage/thresholding algorithm. J. Optim. Theory Appl. **166**(3), 968–82 (2015). https://doi.org/10.1007/s10957-015-0746-4

18. Van Scoy, B., Freeman, R.A., Lynch, K.M.: The fastest known globally convergent first-order method for minimizing strongly convex functions. IEEE Control Syst. Lett. **2**(1), 49–54 (2018). https://doi.org/10.1109/LCSYS.2017.2722406

19. Polyak, B.T.: Introduction to Optimization. Optimization Software Inc, New York (1987)

20. Lessard, L., Recht, B., Packard, A.: Analysis and design of optimization algorithms via integral quadratic constraints. SIAM J. Optim. **26**(1), 57–95 (2016). https://doi.org/10.1137/15M1009597

21. Chiang, A.: Fundamental Methods of Mathematical Economics. McGraw-Hill, New York (1984)

22. Powell, M.J.D.: Restart procedures for the conjugate gradient method. Math. Program. **12**(1), 241–54 (1977). https://doi.org/10.1007/bf01593790

23. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York (2006). https://doi.org/10.1007/978-0-387-40065-5

24. Nemirovski, A.: Efficient methods in convex programming (1994). http://www2.isye.gatech.edu/~nemirovs/Lect_EMCO.pdf . Lecture notes

25. Fercoq, O., Qu, Z.: Restarting accelerated gradient methods with a rough strong convexity estimate (2016). arXiv:1609.07358

26. Fercoq, O., Qu, Z.: Adaptive restart of accelerated gradient methods under local quadratic growth condition (2017). arXiv:1709.02300

27. Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H., (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer, New York, NY (2011). https://doi.org/10.1007/978-1-4419-9569-8_10

28. Kim, D., Fessler, J.A.: Adaptive restart of the optimized gradient method for convex optimization (2017). arXiv:1703.04641