Mathematical Programming (2022) 195:735–781
https://doi.org/10.1007/s10107-021-01713-3

FULL LENGTH PAPER

Series A

From differential equation solvers to accelerated first-order methods for convex optimization

Hao Luo[1] · Long Chen[2]

Received: 13 June 2019 / Accepted: 19 September 2021 / Published online: 12 October 2021

Abstract

Convergence analysis of accelerated first-order methods for convex optimization problems are developed from the point of view of ordinary differential equation solvers. A new dynamical system, called Nesterov accelerated gradient (NAG) flow, is derived from the connection between acceleration mechanism and A-stability of ODE solvers, and the exponential decay of a tailored Lyapunov function along with the solution trajectory is proved. Numerical discretizations of NAG flow are then considered and convergence rates are established via a discrete Lyapunov function. The proposed differential equation solver approach can not only cover existing accelerated methods, such as FISTA, Güler's proximal algorithm and Nesterov's accelerated gradient method, but also produce new algorithms for composite convex optimization that possess accelerated convergence rates. Both the convex and the strongly convex cases are handled in a unified way in our approach.

Keywords Accelerated first-order methods · Ordinary differential equation · Convergence analysis · Convex optimization · Lyapunov function · Exponential decay

Mathematics Subject Classification 37N40 · 65L20 · 65B99 · 90C25

Long Chen was supported in part by the National Science Foundation under grants DMS-1913080 and DMS-2012465. Hao Luo was supported by the China Scholarship Council (CSC) joint Ph.D. student scholarship (Grant 201806240132).

✉ Long Chen
chenlong@math.uci.edu

Hao Luo
galeolev@foxmail.com

1 Department of Mathematics, Sichuan University, Chengdu 610065, China

2 Department of Mathematics, University of California at Irvine, Irvine, CA 92697, USA

1 Introduction

We consider iterative methods for solving the unconstrained minimization problem

$$\min_{x \in V} f(x), \tag{1}$$

where V is a Hilbert space, and $f : V \to \mathbb{R} \cup \{+\infty\}$ is a properly closed convex function. We shall first consider smooth f on the entire space V and later focus on the composite case $f = h + g$ where both h (smooth) and g (non-smooth) are convex on some (simple) closed convex set $Q \subseteq V$. We are mainly interested in the development and analysis of accelerated first-order methods.

Suppose V is equipped with the inner product $(\cdot, \cdot)$ and the correspondingly induced norm $\|\cdot\|$. We use $\langle \cdot, \cdot \rangle$ to denote the duality pair between V^* and V, where V^* is the continuous dual space of V and is endowed with the conventional dual norm $\|\cdot\|_*$. For any interval $I \subseteq \mathbb{R}$, denote by $C^k(I; V)$ the space of all k-times continuous differentiable V-valued functions on I, and the superscript k is dropped when $k = 0$. Let $\Omega \subseteq V$ be some closed convex subset, we say $f \in \mathcal{S}^1_\mu(\Omega)$ if it is continuous differentiable on Ω and there exists $\mu \geq 0$ such that

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{\mu}{2} \|x - y\|^2 \quad \forall x, y \in \Omega. \tag{2}$$

We call (2) the μ-convexity of f and when $\mu > 0$, we say f is strongly convex. We also write $f \in \mathcal{S}^{1,1}_{\mu,L}(\Omega)$ if $f \in \mathcal{S}^1_\mu(\Omega)$ and ∇f is Lipschitz continuous on Ω: there exists $0 < L < \infty$ such that

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\| \quad \forall x, y \in \Omega. \tag{3}$$

By [29, Theorem 2.1.5], this implies the inequality

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{L}{2} \|x - y\|^2 \quad \forall x, y \in \Omega. \tag{4}$$

For $\Omega = V$, we shall write $\mathcal{S}^1_\mu(\Omega)$ and $\mathcal{S}^{1,1}_{\mu,L}(\Omega)$ as $\mathcal{S}^1_\mu$ and $\mathcal{S}^{1,1}_{\mu,L}$, respectively.

The above functional classes are what we work with in this paper. As for the optimization problem (1), we also care about the global minimizer(s) of f. For strongly convex case, it is well-known that the minimizer exists uniquely. However, for convex case, to promise the existence of minimizers, additional assumption, such as the coercivity condition, is usually imposed. Throughout, we denote by $\operatorname{argmin} f$ the set of global minimizers of (1) and assume it is nonempty.

One approach to derive the gradient descent (GD) method is discretizing an ordinary differential equation (ODE), i.e., the so-called gradient flow:

$$x'(t) = -\nabla f(x(t)), \quad t > 0. \tag{5}$$

Here we introduce an artificial time variable t and x' is the derivative taken with respect to t. For ease of notation, in the sequel, we shall omit t when no confusion arises. The simplest forward (explicit) Euler method with step size $\eta_k > 0$ leads to the GD method

$$x_{k+1} = x_k - \eta_k \nabla f(x_k).$$

In the terminology of numerical analysis, it is well-known that this method is *conditionally A-stable* (cf. Sect. 2), and for $f \in \mathcal{S}_{\mu,L}^{1,1}$ with $0 \le \mu \le L < \infty$, the step size $\eta_k = 1/L$ is allowed to get the rate (see [29, Chapter 2])

$$O\left(\min\left\{L/k, (1+\mu/L)^{-k}\right\}\right). \tag{6}$$

One can also consider the backward (implicit) Euler method

$$x_{k+1} = x_k - \eta_k \nabla f(x_{k+1}), \tag{7}$$

which is *unconditionally A-stable* (cf. Sect. 2) and coincides with the well-known proximal point algorithm (PPA) [33]

$$x_{k+1} = \mathbf{prox}_{\eta_k f}(x_k) := \operatorname*{argmin}_{y \in V} \left(f(y) + \frac{1}{2\eta_k}\|y - x_k\|^2 \right). \tag{8}$$

Note that this method allows f to be nonsmooth and possesses linear convergence rate even for convex functions, as long as $\eta_k \ge \eta > 0$ for all $k > 0$.

1.1 Main results

Let us start from the quadratic objective $f(x) = \frac{1}{2}x^\top Ax$ over $\mathbb{R}^d$, for which the gradient flow (5) reads simply as

$$x' = -Ax, \tag{9}$$

where A is symmetric positive semi-definite and makes $f \in \mathcal{S}_{\mu,L}^{1,1}$. Instead of solving (9), we turn to a general linear ODE system

$$y' = Gy. \tag{10}$$

Briefly speaking, our main idea is to seek such a system (10) with some asymmetric block matrix G that transforms the spectrum of A from the real line to the complex plane and reduces the condition number from $\kappa(A) = L/\mu$ to $\kappa(G) = O(\sqrt{L/\mu})$. Afterwards, accelerated gradient methods may be constructed from A-stable methods for solving (10) with a significant larger step size and consequently improve the contraction rate from $O((1-\mu/L)^k)$ to $O((1-\sqrt{\mu/L})^k)$. Furthermore, to handle the convex case $\mu = 0$, we combine the transformation idea with suitable time scaling technique; for more details, we refer to Sect. 2.

One successful and important transformation is given below

$$G = \begin{pmatrix} -I & I \\ \mu/\gamma - A/\gamma & -\mu/\gamma\, I \end{pmatrix}, \tag{11}$$

where the built-in scaling factor γ is positive and satisfies

$$\gamma' = \mu - \gamma, \quad \gamma(0) = \gamma_0 > 0. \tag{12}$$

Based on this, for general $f \in \mathcal{S}_\mu^1$ with $\mu \geq 0$, we replace A in (11) with ∇f and write $y = (x, v)$ to obtain a first-order dynamical system:

$$\begin{cases} x' = v - x, \\ v' = \dfrac{\mu}{\gamma}(x - v) - \dfrac{1}{\gamma}\nabla f(x). \end{cases} \tag{13}$$

Eliminating v, we arrive at a second-order ODE of x:

$$\gamma x'' + (\mu + \gamma)\, x' + \nabla f(x) = 0, \tag{14}$$

which is actually a heavy ball model (cf. (21)) with variable damping coefficients in front of x'' and x'. Thanks to the scaling factor γ, we can handle both the convex case ($\mu = 0$) and the strongly convex case ($\mu > 0$) in a unified way. Moreover, we shall prove the exponential decay property

$$\mathcal{L}(t) \leq e^{-t}\mathcal{L}(0), \quad t > 0, \tag{15}$$

for a tailored Lyapunov function

$$\mathcal{L}(t) = f(x(t)) - f(x^*) + \frac{\gamma(t)}{2}\left\| v(t) - x^* \right\|^2, \quad t > 0, \tag{16}$$

where $x^* \in \operatorname{argmin} f$ is a global minimizer of f.

Accelerated gradient methods based on numerical discretizations of the dynamical system (13) with $f \in \mathcal{S}_{\mu,L}^{1,1}$ are then considered and analyzed by means of a discrete version of the Lyapunov function (16). It will be shown that the implicit scheme (see (72)) possesses linear convergence rate as long as the time step size is uniformly bounded below. This matches the exponential decay rate (15) in the continuous level. Also, for convex case $\mu = 0$, this implicit method amounts to an accelerated PPA, that is very close to Güler's PPA [20] and enjoys the same rate $O(1/k^2)$ (cf. Theorem 4). In Sect. 5, for semi-implicit schemes with suitable corrections (either an extrapolation or a gradient step), we prove the following convergence rate

$$O\left(\min\left\{L/k^2, (1 + \sqrt{\mu/L})^{-k}\right\}\right), \tag{17}$$

which is optimal in the sense of [29]. Moreover, we can recover Nesterov's optimal method [27,29] *exactly* from a semi-implicit scheme with a gradient descent correction; see Sect. 6. Therefore, instead of using estimate sequence, our ODE approach provides an alternative derivation of Nesterov's method and hopefully more intuitive for understanding the acceleration mechanism. From this point of view, we name both (13) and (14) as *Nesterov accelerated gradient (NAG) flow*.

As a proof of concepts, we also generalize our NAG flow to the composite case

$$\min_{x\in Q} f(x) := \min_{x\in Q} \left[h(x) + g(x)\right], \tag{18}$$

where $Q \subseteq V$ is a (simple) closed convex set, $h \in \mathcal{S}^{1,1}_{\mu,L}(Q)$ with $0 \le \mu \le L < \infty$ and $g : V \to \mathbb{R} \cup \{+\infty\}$ is proper, closed, and convex. We use $\mathbf{dom}\, g$ to denote the effective domain of g and assume that $Q \cap \mathbf{dom}\, g \neq \emptyset$. Treating (18) as an unconstrained minimization of $F = f + i_Q$ where i_Q denotes the indicator function of Q, the generalized version of (14) is a second-order differential inclusion

$$\gamma x'' + (\mu + \gamma)\, x' + \partial F(x) \ni 0. \tag{19}$$

We shall give the existence of the solution to (19) in proper sense and then obtain the exponential decay (15) for almost all $t > 0$.

For the unconstrained case $Q = V$, by using the tool of composite gradient mapping [29, Chapter 2], a semi-implicit scheme with correction for the generalized NAG flow (19) is presented and leads to an accelerated proximal gradient method (APGM); see Algorithm 2. We also give a simplified variant that is closely related to FISTA [12]. For the constrained problem (18), an accelerated forward-backward method is proposed in Algorithm 4. Both two algorithms call the proximal operation of g (over Q) only once in each iteration, and they are proved to share the same accelerated convergence rate (17).

The rest of this paper is organized as follows. In the continuing of the introduction, we will review some existing works devoting to the accelerated gradient methods from the ODE point of view. Next, in Sect. 2, we shall explain the acceleration mechanism from A-stability theory of ODE solvers and derive our NAG flow as well. Then in Sect. 3 we focus on the NAG flow and prove its exponential decay. After that, accelerated gradient methods based on numerical discretizations of NAG flow are proposed and analyzed in Sects. 4, 5 and 6. Finally, in Sect. 7, we extend the our NAG flow to composite optimization and propose two new accelerated methods with convergence analysis.

1.2 Related works

The well-known momentum method can be traced back to the 1960s. In [34], Polyak studied the heavy ball (HB) method

$$x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1}) \tag{20}$$

and its continuous analogue, the heavy ball dynamical system:

$$x'' + \alpha_1 x' + \alpha_2 \nabla f(x) = 0. \tag{21}$$

Local linear convergence results for (20) via spectrum analysis were established in [34, Theorem 9]. Note that the HB method (20) adds a momentum term up to the gradient step and is sensitive to its parameters. For $f \in \mathcal{S}^{1,1}_{\mu,L}$, it shares the same theoretical convergence rate (6) as the gradient descent method; see [18,40]. To our best knowledge, no work has established the global accelerated rate (17) for the original HB method (20). Recently, Nguyen et al. [26] developed the so-called accelerated residual method which combines (20) with an extra gradient descent step:

$$\begin{cases} y_k = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}), \\ x_{k+1} = y_k - \dfrac{\alpha}{\beta + 1} \nabla f(y_k). \end{cases}$$

Numerically, they verified the efficiency and usefulness of this method with a restart strategy. We refer to [1,3,11,19] for further investigations of the HB system (21).

To understand an accelerated gradient method with the rate $O(1/k^2)$ proposed by Nesterov [27], Su, Boyd and Candès [37] derived the following second-order ODE

$$x'' + \frac{\alpha}{t} x' + \nabla f(x) = 0, \quad t > 0, \tag{22}$$

where $\alpha > 0$ and $f \in \mathcal{S}^{1,1}_{0,L}$. If $\alpha \geq 3$ or $1 < \alpha < 3$ and $(f - f(x^*))^{(\alpha-1)/2}$ is convex, they proved the decay rate $O(t^{-2})$. If $\alpha \geq 3$ and f is strongly convex, they also obtained a faster rate $O(t^{-2\alpha/3})$. Later on, Aujol and Dossal [10] established a generic result:

$$f(x(t)) - f(x^*) \leq \begin{cases} Ct^{-2}, & \text{if } \alpha \geq 2\beta + 1, \\ Ct^{-2\alpha/(2\beta+1)}, & \text{if } 0 < \alpha < 2\beta + 1, \end{cases} \tag{23}$$

where $\beta > 0$ and $(f - f(x^*))^{\beta}$ is convex. Almost at the same time, Attouch et al. [8] obtained the estimate (23) for $\beta = 1$ and considered numerical discretizations for (22) with the convergence rate $O(k^{-\min\{2,2\alpha/3\}})$. Also, Vassilis et al. [42] studied the non-smooth version of (22):

$$x'' + \frac{\alpha}{t} x' + \partial f(x) \ni 0. \tag{24}$$

They proved that the solution trajectory of (24) converges to a minimizer of f and derived the decay estimate (23) for $\beta = 1$. For more works and generalizations related to the model (22) and the corresponding algorithms, we refer to [2,5–7,14] and references therein.

Recently, Wibisono et al. [43] introduced a Lagrangian

$$\mathcal{E}(y, w, t) = \frac{e^{\int_0^t \alpha(s)\,\mathrm{d}s}}{\alpha(t)\beta(t)} \left(\frac{\beta(t)}{2} \|w\|^2 - \alpha^2(t) f(y) \right), \tag{25}$$

for smooth and convex f, where $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous and $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies

$$\beta' \geq -\alpha\beta, \quad \beta(0) = \beta_0 > 0. \tag{26}$$

The Lagrangian (25) itself introduces a variational problem, the Euler–Lagrange equation to which is

$$\begin{cases} y' = \alpha(w - y), \\ \beta w' = -\alpha \nabla f(y). \end{cases} \tag{27}$$

They then established the convergence rate (cf. [43, Theorem 2.1])

$$f(y(t)) - f(x^*) \leq e^{-\int_0^t \alpha(s)\,\mathrm{d}s} \mathcal{L}(0), \tag{28}$$

by means of the Lyapunov function

$$\mathcal{L}(t) = e^{\int_0^t \alpha(s)\,\mathrm{d}s} \left[f(y(t)) - f(x^*) \right] + \frac{1}{2} \left\| w(t) - x^* \right\|^2.$$

Following this work, for $f \in \mathcal{S}_\mu^1$ with $\mu > 0$, Wilson et al. [44] introduced another Lagrangian whose Euler–Lagrange equations reads as

$$\begin{cases} y' = \alpha(w - y), \\ \mu w' = \mu\alpha(y - w) - \alpha \nabla f(y), \end{cases} \tag{29}$$

with the same scaling function α in (25). They proved the decay estimate (28) as well, by using the Lyapunov function

$$\mathcal{L}(t) = e^{\int_0^t \alpha(s)\,\mathrm{d}s} \left[f(y(t)) - f(x^*) + \frac{\mu}{2} \left\| w(t) - x^* \right\|^2 \right]. \tag{30}$$

When $\alpha = \sqrt{\mu}$, (29) gives the following model

$$y'' + 2\sqrt{\mu} y' + \nabla f(y) = 0, \tag{31}$$

which reduces to an HB system (cf. (21)); see also Siegel [38].

In addition, Siegel [38] and Wilson et al. [44] proposed two semi-explicit schemes for (31) individually. Both of their schemes are supplemented with an extra gradient descent step and share the same linear convergence rate $O((1 - \sqrt{\mu/L})^k)$.

Recently, introducing the so-called duality gap which is the difference of appropriate upper and lower bound approximations for the objective function, Diakonikolas and Orecchia [17] presented a general framework for the construction and analysis of continuous time dynamical systems and the corresponding numerical discretizations. They recovered several existing ODE models such as the gradient flow (5), the mirror descent dynamic system and its accelerated version. We mention that the derivation of our NAG flow and analyses of discrete algorithms are fundamentally different from their duality gap technique.

2 Stability of ODE solvers and acceleration

In what follows, for any $M \in \mathbb{R}^{d\times d}$, $\sigma(M)$ denotes the spectrum of M, i.e., the set of all eigenvalues of M. The spectral radius is then defined by $\rho(M) := \max_{\lambda\in\sigma(M)} |\lambda|$, and when M is invertible, its condition number $\kappa(M) := \rho(M^{-1})\rho(M)$. If $\sigma(M) \subset \mathbb{R}$, then $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ stand for the minimum and maximum of $\sigma(M)$, respectively. Moreover, $\|\cdot\|_2$ is the usual 2-norm for vectors and matrices.

To present our main idea as simple as possible, in this section, unless other specified, we restrict ourselves to the quadratic objective $f(x) = \frac{1}{2}x^\top Ax$, where A is a symmetric matrix with the bound

$$0 \le \mu := \lambda_{\min}(A) \le \lambda \le \lambda_{\max}(A) := L \quad \forall \lambda \in \sigma(A).$$

For this model example, $\nabla f(x) = Ax$ and the gradient flow (5) reads as $x' = -Ax$. The global minimal is achieved at $x^* = 0$, and when $\mu > 0$, the condition number of A is $\kappa(A) = L/\mu$.

2.1 A-stability of ODE solvers

Let $G \in \mathbb{R}^{d\times d}$ and assume $\mathfrak{Re}(\lambda) < 0$ for all $\lambda \in \sigma(G)$. For the linear ODE system

$$y' = Gy, \quad y(0) = y_0 \in \mathbb{R}^d, \tag{32}$$

it is not hard to derive that $\|y(t)\|_2 \to 0$ as $t \to \infty$ (see [13, Theorem 7] for instance). Hence $y^* = 0$ is an equilibrium of the dynamic system (32).

We now recall the concept of A-stability of ODE solves [23,39]. A one-step method ϕ for (32) with step size $\alpha > 0$ can be formally written as

$$y_{k+1} = E_\phi(\alpha, G)y_k. \tag{33}$$

As $y^* = 0$ is an equilibrium point, (33) also gives the error equation. The scheme ϕ is called *absolute stable* or *A-stable* if $\rho(E_\phi(\alpha, G)) < 1$ from which the asymptotic convergence $y_k \to 0$ follows (cf. [16, Theorem 6.1]). If $\rho(E_\phi(\alpha, G)) < 1$ holds for all $\alpha > 0$, then it is called *unconditionally A-stable*, and if $\rho(E_\phi(\alpha, G)) < 1$ for any $\alpha \in I$, where I is an interval of the positive half line, then the scheme is called *conditionally A-stable*.

If $E_\phi(\alpha, G)$ is normal, then $\|E_\phi(\alpha, G)\|_2 = \rho(E_\phi(\alpha, G))$. Therefore for A-stable methods the linear convergence follows directly from the norm contraction

$$\|y_{k+1}\|_2 \leq \rho(E_\phi(\alpha, G))\,\|y_k\|_2\,. \tag{34}$$

In general cases, however, bounding the spectral radius by one does not imply the norm contraction, i.e., (34) may not be true when $E_\phi(\alpha, G)$ is non-normal, even if (33) is A-stable. Nevertheless, we shall continue using the tool of A-stability through spectral analysis and comment on its limitation in Sect. 2.6.

2.2 Implicit and explicit Euler methods

It is well known that the implicit Euler (IE) method

$$\frac{y_{k+1} - y_k}{\alpha} = G y_{k+1}$$

is unconditionally A-stable. Indeed, $E_{\mathrm{IE}}(\alpha, G) = (I - \alpha G)^{-1}$ and $\rho(E_{\mathrm{IE}}(\alpha, G)) < 1$ for all $\alpha > 0$ since all eigenvalues of αG lie on the left of the complex plane and their distances to 1 are larger than one. Moreover, as it has no restriction on the step size, the implicit Euler method can achieve faster convergent rate by time rescaling which is equivalent to choosing a large step size.

In contract, the explicit Euler method

$$\frac{y_{k+1} - y_k}{\alpha} = G y_k \tag{35}$$

is only conditionally A-stable. Let us consider the case $G = -A$ with $\mu > 0$. Then (35) is exactly the gradient descent method for minimizing $\frac{1}{2}x^\top Ax$. It is not hard to obtain that

$$\rho(E_{\mathrm{GD}}(\alpha, -A)) = \rho(I - \alpha A) = \max\left\{\,|1 - \alpha\mu|\,,\ |1 - \alpha L|\,\right\}. \tag{36}$$

Hence $\rho(E_{\mathrm{GD}}(\alpha, -A)) < 1$ provided $0 < \alpha < 2/L$. Thanks to the symmetry of A, we have $\|E_{\mathrm{GD}}(\alpha, -A)\|_2 = \rho(E_{\mathrm{GD}}(\alpha, -A))$ and the norm convergence with linear rate follows. Moreover, based on (36), a standard argument outputs the optimal choice $\alpha^* = 2/(\mu + L)$, which gives the minimal spectrum

$$\|E_{\mathrm{GD}}(\alpha^*, -A)\|_2 = \min_{\alpha > 0} \rho(I - \alpha A) = \frac{\kappa(A) - 1}{\kappa(A) + 1}. \tag{37}$$

A quasi-optimal but simpler choice is $\alpha_* = 1/L$ which yields

$$\|E_{\mathrm{GD}}(\alpha_*, -A)\|_2 = \rho(I - \alpha_* A) = 1 - \frac{1}{\kappa(A)}. \tag{38}$$

We formulate the convergence rates (37) and (38) in terms of the condition number $\kappa(A)$ as it is invariant to the rescaling of A, i.e., $\kappa(cA) = \kappa(A)$ for any real number $c \neq 0$. To be A-stable, one has to choose $0 < \alpha < 2/\lambda_{\max}(A)$. It seems that a simple rescaling to cA can reduce $\lambda_{\max}(cA)$ and thus enlarge the range of the step size. However, the condition number $\kappa(cA) = \kappa(A)$ is invariant. From this we see that for the GD method (35), the simple rescaling cA is in vain.

The magnitude of the step size is relative to $\min|\lambda(G)|$. To fix the discussion, we chose $G = -A/\mu$ in (35) so that $\lambda_{\min}(A/\mu) = 1$. Then in order for the explicit Euler method to be A-stable it is equivalent to choose $\alpha = O(1/\kappa(A))$ which leads to the contraction rate $1 - 1/\kappa(A)$. Consequently for ill-conditioned problems, a tiny step size proportional to $1/\kappa(A)$ is required.

Rather than the rescaling, our main intuition is to seek some transformation G of A, that reduces $\kappa(A)$ to $\kappa(G) = O(\sqrt{\kappa(A)})$. We wish to construct explicit A-stable methods which can enlarge the step size from $O(1/\kappa(A))$ to $O(1/\sqrt{\kappa(A)})$ and consequently improve the contraction rate from $1 - 1/\kappa(A)$ to $O(1 - 1/\sqrt{\kappa(A)})$.

2.3 Transformation to the complex plane

Let us first consider the case $\mu > 0$ and embed A into some 2×2 block matrix G with a rotation built-in. Specifically, we construct two candidates

$$G_{\text{HB}} = \begin{pmatrix} 0 & I \\ -A/\mu & -2I \end{pmatrix} \quad \text{and} \quad G_{\text{NAG}} = \begin{pmatrix} -I & I \\ I - A/\mu & -I \end{pmatrix}. \tag{39}$$

Due to the asymmetrical fact, $\sigma(A)$ will be transformed from the real line to the complex plane. This may shrink the condition number; see the following result.

Proposition 1 *For $G = G_{\text{HB}}$ or G_{NAG} given in (39), it satisfies $\mathfrak{Re}(\lambda) < 0$ for any $\lambda \in \sigma(G)$, which promises the decay property $\|y(t)\|_2 \to 0$ for the system $y' = Gy$. Moreover, we have $\kappa(G_{\text{HB}}) = \kappa(G_{\text{NAG}}) = \sqrt{\kappa(A)}$.*

Proof Let us first consider $G = G_{\text{HB}}$. As A is symmetric, we can write $A = U\Lambda U^\top$ with unitary matrix U and diagonal matrix Λ consisting of eigenvalues of A. By applying the similar transform to G with the block diagonal matrix $\text{diag}(U, U)$, it suffices to consider eigenvalues of

$$R_{\text{HB}} = \begin{pmatrix} 0 & 1 \\ -\theta & -2 \end{pmatrix}, \quad \theta \in \sigma(A/\mu).$$

It is clear that $\det R_{\text{HB}} = \theta$ and $\text{tr}\, R_{\text{HB}} = -2 < 0$. In addition, since $\left|\text{tr}\, R_{\text{HB}}\right|^2 \le 4\det R_{\text{HB}}$, any eigenvalue $\lambda_R \in \sigma(R_{\text{HB}})$ is a complex number and

$$\mathfrak{Re}(\lambda_R) = -1, \quad |\lambda_R| = \sqrt{\det R_{\text{HB}}} = \sqrt{\theta}.$$

As $1 = \lambda_{\min}(A/\mu) \le \theta \le \lambda_{\max}(A/\mu) = \kappa(A)$, we conclude $\kappa(G_{\text{HB}}) = \sqrt{\kappa(A)}$.

Apply the similar transformation with $P = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, we observe that

$$R_{\text{NAG}} = P R_{\text{HB}} P^{-1} = \begin{pmatrix} -1 & 1 \\ 1-\theta & -1 \end{pmatrix}.$$

So $\sigma(R_{\text{NAG}}) = \sigma(R_{\text{HB}})$ and consequently $\kappa(G_{\text{NAG}}) = \sqrt{\kappa(A)}$. This completes the proof of this proposition. □

We write $y = (x, v)^\top$ and eliminate v in $y' = Gy$ to get a second order ODE of x, in which we replace Ax by general form $\nabla f(x)$. Both G_{HB} and G_{NAG} yield the same thing

$$\mu x'' + 2\mu x' + \nabla f(x) = 0, \tag{40}$$

which is a special case of the HB model (cf. (21)).

Note that we can find a lot of transformations G and derive corresponding ODE models. Indeed, given any G that meets our demand, both cG and QGQ^{-1} are acceptable candidates, where $c > 0$ and Q is some invertible matrix. We are not going further deep beyond those two transformations given in (39) for the strongly convex case $\mu > 0$ but aim to combine the transformation with a refined time scaling to propose another one for convex case $\mu = 0$ in Sect. 2.5.

2.4 Acceleration from a Gauss–Seidel splitting

We now consider numerical discretization for (32) with $G = G_{\text{HB}}$ and G_{NAG} given in (39). As discussed in Sect. 2.2, the implicit Euler method is unconditionally A-stable. But computing $(I - \alpha G)^{-1}$ needs significant effort and may not be practical.

One may hope that the explicit Euler method $y_{k+1} = (I + \alpha G) y_k$ will be A-stable with step size $\alpha = O(1/\kappa(G)) = O(1/\sqrt{\kappa(A)})$. Unfortunately, unlike the discussion for (35) with $G = -A$, where $\sigma(I - \alpha A)$ lies on the real line and $\rho(I - \alpha A)$ can be easily shrunk by choosing $\alpha = 1/\rho(A)$ (cf. (36)), the general asymmetric G spreads the spectrum on the complex plane. For both $G = G_{\text{HB}}$ and $G = G_{\text{NAG}}$, we have $\Re(\lambda) = -1$ for all $\lambda \in \sigma(G)$. Denote by $r = \rho(G)$. Then $\rho^2(I + \alpha G) = (1-\alpha)^2 + \alpha^2(r^2 - 1)$. To be A-stable, requiring $\rho(I + \alpha G) < 1$ is equivalent to letting $0 < \alpha < 2/r^2 = O(1/\kappa(A))$, where small step size $\alpha = O(1/\kappa(A))$ is still needed. The optimal choice $\alpha^* = r^{-2}$ only gives

$$\rho(I + \alpha^* G) = 1 - \alpha^* = 1 - O(1/\kappa(A)),$$

where no acceleration has been obtained.

We then expect that an explicit scheme closer to the implicit Euler method will hopefully have better stability with a larger step size.

Motivated by the Gauss–Seidel (GS) method [45] for computing $(I - \alpha G)^{-1}$, we consider the matrix splitting $G = M + N$ with M being the lower triangular part of G

(including the diagonal) and $N = G - M$, and propose the following Gauss–Seidel splitting scheme

$$\frac{y_{k+1} - y_k}{\alpha} = My_{k+1} + Ny_k \tag{41}$$

which gives the relation

$$y_{k+1} = E(\alpha, G)y_k, \quad E(\alpha, G) := (I - \alpha M)^{-1}(I + \alpha N). \tag{42}$$

Note that for $G = G_{\text{HB}}$ and G_{NAG}, the scheme (41) is still explicit as the lower triangular block matrix $I - \alpha M$ can be inverted easily, without involving A^{-1}.

The spectrum bound is given below and for the algebraic proof details, we refer to "Appendix A".

Theorem 1 *For $G = G_{\text{HB}}$ or G_{NAG} given in (39), if $0 < \alpha \le 2/\sqrt{\kappa(A)}$, then the Gauss–Seidel splitting scheme (41) is A-stable and*

$$\rho(E(\alpha, G)) \le \frac{1}{\sqrt{1 + 2\alpha}}.$$

2.5 Dynamic time rescaling for the convex case

The ODE model (40) given in Sect. 2.3 cannot treat the case $\mu = 0$ and the previous spectral analysis fails. Equivalently the condition number $\kappa(A)$ is infinity and the spectrum bound becomes 1. To conquer this, a careful rescaling is needed. Throughout this subsection, we assume $\mu = 0$.

For the gradient flow

$$x'(t) = -\nabla f(x(t)), \tag{43}$$

one can easily establish the sub-linear rate $f(x(t)) - f(x^*) \le C/t$; see [37]. To recover the exponential rate, we introduce a time rescaling $t(s) = e^s$ and let $y(s) = x(t(s))$. Then (43) becomes the following rescaled gradient flow

$$\gamma(s)y'(s) = -\nabla f(y(s)), \tag{44}$$

with the scaling factor $\gamma(s) = e^s$. Besides, the previous sublinear rate $f(x(t)) - f(x^*) \le C/t$ turns into $f(y(s)) - f(x^*) \le Ce^{-s}$. That is in the continuous level, we can achieve the exponential decay through suitable rescaling of time even for $\mu = 0$.

Now let us go back to our model case $f(x) = \frac{1}{2}x^\top Ax$ with $\mu = 0$ and $\lambda_{\max}(A) = L$. Coupled with the transformation G_{NAG}, we consider

$$y' = G(\gamma)\, y, \quad G(\gamma) = \begin{pmatrix} -I & I \\ -A/\gamma & O \end{pmatrix}, \tag{45}$$

where $y = (x, v)^\top$ and

$$\gamma' = -\gamma, \quad \gamma(0) = \gamma_0 > 0. \tag{46}$$

This gives a second-order ODE in terms of x:

$$\gamma x'' + \gamma x' + \nabla f(x) = 0, \tag{47}$$

which is in the HB type but with variable damping coefficients.

Obviously, the implicit Euler method for solving (45) is still unconditional A-stable. We now apply the GS splitting (41) to (45) and get

$$y_{k+1} = E(\alpha_k, G(\gamma_{k+1}))y_k, \tag{48}$$

where $E(\alpha_k, G(\gamma_{k+1}))$ is defined in (42). The equation (46) is discretized by

$$\gamma_{k+1} = \gamma_k - \alpha_k\gamma_{k+1}. \tag{49}$$

Eliminating v_k in (48) will give an HB method with variable coefficients

$$x_{k+1} = x_k - \frac{\alpha_k\alpha_{k-1}}{\gamma_k + \alpha_k\gamma_k}\nabla f(x_k) + \frac{\alpha_k}{\alpha_{k-1} + \alpha_k\alpha_{k-1}}(x_k - x_{k-1}).$$

Instead of studying the spectrum bound $E(\alpha_k, G(\gamma_{k+1}))$ which is 1, we apply the scaling technique to obtain a regularized matrix

$$\widetilde{E}_k = \begin{pmatrix} I & O \\ O & \gamma_{k+1}I \end{pmatrix} E(\alpha_k, G(\gamma_{k+1})) \begin{pmatrix} I & O \\ O & \gamma_k I \end{pmatrix}^{-1},$$

which is nearly similar with $E(\alpha_k, G(\gamma_{k+1}))$. Set $z_k = \begin{pmatrix} I & O \\ O & \gamma_k I \end{pmatrix} y_k$, then the discrete system (48) for $\{y_k\}$ becomes

$$z_{k+1} = \widetilde{E}_k z_k, \tag{50}$$

With a careful chosen step size, the spectrum bound of $\widetilde{E}_k$ is given below and for the algebraic proof details, we refer to "Appendix A". We note that, the step size in Theorem 2 is only to agree with the setting of Lemma B2 and for general choice $L\alpha_k^2/\gamma_k = O(1)$ and suitable initial value γ_0, it is possible to maintain the spectrum bound (51) together with the decay estimate (52).

Theorem 2 *If $\gamma_0 = L$ and $L\alpha_k^2 = \gamma_k(1 + \alpha_k)$, then both the scheme (48) and its equivalent form (50) are A-stable and we have*

$$\rho(\widetilde{E}_k) = \frac{\gamma_{k+1}}{\gamma_k} = \frac{1}{1 + \alpha_k}, \tag{51}$$

which further implies that

$$\prod_{i=0}^{k-1} \rho(\widetilde{E}_i) = \frac{\gamma_k}{\gamma_0} = O(k^{-2}). \tag{52}$$

2.6 Limitation of spectral analysis

For quadratic objective f, both the ODE models (40) and (47) are linear and the spectrum bound of $E(\alpha, G)$ for the Gauss–Seidel splitting (42) is derived. But as pointed out in the beginning, for A-stable methods, bounding the spectral radius by one is not sufficient for the norm convergence if the matrix $E(\alpha, G)$ is non-normal; see convincible examples in [23, Appendix D.2] and [23, Appendix D.4].

Moving beyond quadratic f and nonlinear ODE systems, transient growth or instability of perturbed problems can easily lead to nonlinear instabilities. Particularly, for the HB system (21), it is shown in [22] that the parameters optimized for linear ODE models does not guarantee the global convergence for a nonlinear system.

To provide rigorous convergence analysis for both continuous and discrete levels, in the sequel we shall introduce the tool of Lyapunov function. Following many related works [6,37,43], we first analyze some proper ODEs via a Lyapunov function, then construct optimization algorithms from numerical discretizations of continuous models and use a discrete Lyapunov function to establish the convergence rates of the proposed algorithms.

3 Nesterov accelerated gradient flow

3.1 Continuous problem

In the previous section, we have obtained two ODE models for quadratic objective $f(x) = \frac{1}{2}x^\top Ax$ with $\mu > 0$ and $\mu = 0$, respectively. To handle those two cases in a unified way, we combine G_{NAG} in (39) with $G(\gamma)$ in (45) and consider the following transformation

$$G = \begin{pmatrix} -I & I \\ \mu/\gamma - A/\gamma & -\mu/\gamma\, I \end{pmatrix}, \tag{53}$$

where

$$\gamma' = \mu - \gamma, \quad \gamma(0) = \gamma_0 > 0. \tag{54}$$

One can solve the above equation and obtain

$$\gamma(t) = \mu + (\gamma_0 - \mu)\mathrm{e}^{-t}, \quad t \geq 0.$$

Since $\gamma_0 > 0$, we have that $\gamma(t) > 0$ for all $t \geq 0$ and $\gamma(t)$ converges to μ exponentially and monotonically as $t \to +\infty$. In particular, if $\gamma_0 = \mu > 0$, then $\gamma(t) = \mu$. Therefore, when $\mu = 0$, (53) reduces to (45) and when $\gamma_0 = \mu > 0$, (53) recovers (39) indeed. Correspondingly, the transform (53) gives the system

$$\begin{cases} x' = v - x, \\ \gamma v' = \mu(x - v) - Ax. \end{cases} \tag{55}$$

Heuristically, for general $f \in \mathcal{S}_\mu^1$ with $\mu \geq 0$, we just replace Ax in (55) with $\nabla f(x)$ and obtain our NAG flow

$$\begin{cases} x' = v - x, \\ \gamma v' = \mu(x - v) - \nabla f(x), \end{cases} \tag{56}$$

with initial conditions $x(0) = x_0$ and $v(0) = v_0$. The equivalent second-order ODE (will also be abbreviated as NAG flow) reads as follows

$$\gamma x'' + (\mu + \gamma)x' + \nabla f(x) = 0, \tag{57}$$

with initial conditions $x(0) = x_0$ and $x'(0) = v_0 - x_0$. Clearly, if $\gamma_0 = \mu > 0$, then (57) becomes (40), and if $\mu = 0$, then (57) coincides with (47).

Motivated by (30), we introduce a Lyapunov function for (56):

$$\mathcal{L}(t) := f(x(t)) - f(x^*) + \frac{\gamma(t)}{2} \left\| v(t) - x^* \right\|^2, \quad t \geq 0. \tag{58}$$

In addition, we need the following lemma, which is trivial but very useful for the convergence analysis in both of the continuous and discrete levels.

Lemma 1 *For any $u, v, w \in V$, we have*

$$2(u - v, v - w) = \|u - w\|^2 - \|u - v\|^2 - \|v - w\|^2.$$

We first present the well-posedness of (57) and prove the exponential decay property of the Lyapunov function (58).

Lemma 2 *If $f \in \mathcal{S}_{\mu,L}^{1,1}$ with $\mu \geq 0$, then the NAG flow (57) admits a unique solution $x \in C^2([0, \infty); V)$ and moreover*

$$\mathcal{L}'(t) \leq -\mathcal{L}(t) - \frac{\mu}{2}\|x'(t)\|^2, \tag{59}$$

which implies that

$$\mathcal{L}(t) + \frac{\mu}{2} \int_0^t e^{s-t} \left\| x'(s) \right\|^2 \mathrm{d}s \leq \mathrm{e}^{-t} \mathcal{L}(0), \quad t \geq 0. \tag{60}$$

Proof Basically, as ∇f is Lipschitz continuous, applying the standard existence and uniqueness results of ODE (see [9, Theorem 4.1.4]) yields the fact that the system (56) admits a unique classical solution $(x, v) \in C^1([0,\infty); V) \times C^1([0,\infty); V)$. This implies that $x' = v - x \in C^1([0,\infty); V)$, and therefore $x \in C^2([0,\infty); V)$ is also the unique solution to our NAG flow (57).

It remains to prove (59), which yields the exponential decay (60) immediately. A straightforward calculation yields that

$$\mathcal{L}'(t) = \left\langle \nabla f(x), x' \right\rangle + \frac{\gamma'}{2} \left\| v - x^* \right\|^2 + \gamma \left\langle v', v - x^* \right\rangle,$$

and by (54) and (56), we replace γ' and v' by their right hand side terms and obtain

$$\mathcal{L}'(t) = \left\langle \nabla f(x), x' \right\rangle + \frac{\mu - \gamma}{2} \left\| v - x^* \right\|^2 + \left\langle \mu(x - v) - \nabla f(x), v - x^* \right\rangle. \tag{61}$$

Let us focus on the last term. Thanks to Lemma 1,

$$\mu(x - v, v - x^*) = \frac{\mu}{2} \left(\left\| x - x^* \right\|^2 - \| x - v \|^2 - \left\| v - x^* \right\|^2 \right),$$

and the gradient term is split as follows

$$- \left\langle \nabla f(x), v - x^* \right\rangle = - \left\langle \nabla f(x), v - x \right\rangle - \left\langle \nabla f(x), x - x^* \right\rangle. \tag{62}$$

By the relation $x' = v - x$, the first term in (62) becomes $\left\langle -\nabla f(x), x' \right\rangle$ which cancels the first term in (61). Combining all identities together gives

$$\mathcal{L}'(t) = \frac{\mu}{2} \left\| x - x^* \right\|^2 - \left\langle \nabla f(x), x - x^* \right\rangle - \frac{\gamma}{2} \left\| v - x^* \right\|^2 - \frac{\mu}{2} \| x' \|^2. \tag{63}$$

As f is μ-strongly convex (cf.(2)), there holds

$$\frac{\mu}{2} \left\| x - x^* \right\|^2 - \left\langle \nabla f(x), x - x^* \right\rangle \le f(x^*) - f(x),$$

and plugging this into (63) implies that

$$\mathcal{L}'(t) \le -\mathcal{L}(t) - \frac{\mu}{2} \| x'(t) \|^2,$$

which proves (59) and thus completes the proof of this lemma. □

Remark 1 According to the proof of Lemma 2, the Eq. (54) for γ can be relaxed to $\gamma' \le \mu - \gamma$. This makes (61) and (63) become inequality but leaves the final estimate (59) invariant. □

3.2 Rescaling property

Based on our NAG flow (56) (or (57)), it is possible to use time scaling technique to construct more ODE systems with any desirable convergence rate. It is worth distinguishing the connection and difference with existing dynamical models.

Specifically, let α be any continuous nonnegative function on $\mathbb{R}_+$, and consider the time rescaling

$$t(\tau) = \int_0^\tau \alpha(s)\,\mathrm{d}s, \quad \tau > 0. \tag{64}$$

Set $y(\tau) = x(t(\tau))$, $w(\tau) = v(t(\tau))$ and $\beta(\tau) = \gamma(t(\tau))$, then it is clear that

$$y'(\tau) = t'(\tau)x'(t(\tau)) = \alpha(\tau)x'(t(\tau)),$$

Similarly, $w'(\tau) = \alpha(\tau)v'(t(\tau))$ and plugging those facts into (56) gives the scaled NAG flow

$$\begin{cases} y' = \alpha(w - y), \\ \beta w' = \mu\alpha\big(y - w\big) - \alpha\nabla f(y), \end{cases} \tag{65}$$

with initial conditions $y(0) = x_0$ and $y'(0) = \alpha(0)x'(0)$. By Remark 1, the Eq. (54) can be replaced by $\gamma' \le \mu - \gamma$, which becomes

$$\beta' \le \alpha(\mu - \beta), \quad \beta(0) = \gamma_0. \tag{66}$$

Correspondingly, the Lyapunov function (58) reads as follows

$$\widetilde{\mathcal{L}}(\tau) := f(y(\tau)) - f(x^*) + \frac{\beta(\tau)}{2}\left\|w(\tau) - x^*\right\|^2, \quad \tau \ge 0.$$

Analogously to (59), we can prove

$$\widetilde{\mathcal{L}}' \le -\alpha\widetilde{\mathcal{L}} - \frac{\mu\alpha}{2}\|w - y\|^2,$$

which implies that

$$\widetilde{\mathcal{L}}(\tau) \le \mathrm{e}^{-\int_0^\tau \alpha(s)\,\mathrm{d}s}\widetilde{\mathcal{L}}(0), \quad \tau \ge 0. \tag{67}$$

Therefore, larger scaling factor α promises faster decay rate.

We note that the scaled NAG flow (65) is very close to the two models (27) and (29), which are derived in [43] and [44] respectively, via the variational perspective. Indeed, they differs mainly from the coefficient of w'. By (66), an elementary calculation gives

$$\beta(\tau) \le \mu + (\gamma_0 - \mu)\mathrm{e}^{-\int_0^\tau \alpha(s)\,\mathrm{d}s}, \quad \tau \ge 0.$$

Therefore, (65) chooses variable coefficient $\beta(\tau)$ for $\mu \geq 0$, while (27) considers dynamically changing coefficient (26) only for $\mu = 0$ and (29) adopts fixed parameter $\mu > 0$. For strongly convex case $\mu > 0$, if we take $\beta = \mu$, which satisfies (66), then the scaled system (65) coincides with (29). For convex case $\mu = 0$, if both (27) and (66) are equalities, then (65) agrees with (27). Hence, we conclude that our NAG flow system is more tight and provides a unified way to handle $\mu = 0$ and $\mu > 0$.

Now, let us look at a concrete rescaling example. Let the scaling factor α satisfy

$$2\alpha' \leq \mu - \alpha^2, \quad \alpha(0) = \sqrt{\gamma_0}. \tag{68}$$

For instance, the following choice is allowed:

$$\alpha(\tau) = \frac{\sqrt{\gamma_0}\, b}{\sqrt{\gamma_0}\, \tau + b}, \quad 0 < b \leq 2. \tag{69}$$

For the equality case of (68), we have a closed-form solution

$$\alpha(\tau) = \begin{cases} \dfrac{2\sqrt{\gamma_0}}{\sqrt{\gamma_0}\, \tau + 2}, & \text{if } \mu = 0, \\ \sqrt{\mu} \cdot \dfrac{e^{\sqrt{\mu}\, \tau} - \alpha_\mu}{e^{\sqrt{\mu}\, \tau} + \alpha_\mu}, & \text{if } \mu > 0, \end{cases} \tag{70}$$

where

$$\alpha_\mu = \frac{\sqrt{\mu} - \sqrt{\gamma_0}}{\sqrt{\mu} + \sqrt{\gamma_0}} \in (-1, 1).$$

We now set $\beta = \alpha^2$ which fulfills (66) by our assumption (68), then the scaled NAG flow (65) gives a new HB system

$$y'' + \frac{1}{\alpha}\left(\mu + \alpha^2 - \alpha'\right) y' + \nabla f(y) = 0. \tag{71}$$

According to (67), we have the estimate

$$\widetilde{\mathcal{L}}(\tau) \leq \begin{cases} \dfrac{b^b \widetilde{\mathcal{L}}(0)}{(\sqrt{\gamma_0}\tau + b)^b}, & \text{if } \alpha \text{ satisfies (69)}, \\ \dfrac{(1 + \alpha_\mu)^2 \widetilde{\mathcal{L}}(0)}{\left(e^{\sqrt{\mu}\tau/2} + \alpha_\mu e^{-\sqrt{\mu}\tau/2}\right)^2}, & \text{if } \alpha \text{ satisfies (70) and } \mu > 0. \end{cases}$$

Particularly, if $\mu > 0$ and α satisfies (70) with $\gamma_0 = \sqrt{\mu}$, then $\alpha(\tau) = \sqrt{\mu}$ and (71) recovers (31) with the same rate $O(e^{-\sqrt{\mu}\tau})$. Moreover, if $\mu = 0$ and α satisfies (69) with $\gamma_0 = 4$ and $b = 2$, then $\alpha(\tau) = 2/(\tau + 1)$ and (71) becomes

$$y'' + \frac{3}{\tau + 1} y' + \nabla f(y) = 0, \quad \tau > 0,$$

which gives the decay rate $O(\tau^{-2})$ and coincides with the prevailing ODE model (22) derived in [37].

4 An implicit scheme

Exponential decay of an implicit discretization for solving (56) can be established, which is more or less straightforward since one can easily follow the proof from the continuous problem. However, the implicit scheme requires efficient solver or proximal calculation and may not be practical sometimes. It is presented here to bridge the analysis from the continuous level to semi-implicit and explicit schemes.

Consider the following implicit scheme

$$\begin{cases} \dfrac{x_{k+1}-x_k}{\alpha_k} = v_{k+1} - x_{k+1}, \\ \dfrac{v_{k+1}-v_k}{\alpha_k} = \dfrac{\mu}{\gamma_k}(x_{k+1} - v_{k+1}) - \dfrac{1}{\gamma_k}\nabla f(x_{k+1}), \end{cases} \tag{72}$$

where $\alpha_k > 0$ denotes the time step size to discretize the time derivative and the parameter Eq. (54) is also discretized implicitly

$$\frac{\gamma_{k+1}-\gamma_k}{\alpha_k} = \mu - \gamma_{k+1}, \quad \gamma_0 > 0. \tag{73}$$

We shall present the convergence result for the implicit scheme (72)–(73). To do so, we introduce a suitable Lyapunov function

$$\mathcal{L}_k := f(x_k) - f(x^*) + \frac{\gamma_k}{2}\left\|v_k - x^*\right\|^2, \tag{74}$$

which is clearly a discrete analogue to the continuous one (58).

Theorem 3 *If $f \in \mathcal{S}_\mu^1$ with $\mu \geq 0$, then for the scheme (72) with $\alpha_k > 0$, we have*

$$\mathcal{L}_{k+1} \leq \frac{\mathcal{L}_k}{1+\alpha_k}, \quad k \in \mathbb{N}.$$

Proof It suffices to prove

$$\mathcal{L}_{k+1} - \mathcal{L}_k \leq -\alpha_k \mathcal{L}_{k+1}. \tag{75}$$

Let us mimic the proof of Lemma 2. Instead of the derivative, we compute the difference as follows

$$\begin{aligned}\mathcal{L}_{k+1}-\mathcal{L}_k =&f(x_{k+1})-f(x_k)+\frac{\gamma_{k+1}-\gamma_k}{2}\left\|v_{k+1}-x^*\right\|^2\\&+\frac{\gamma_k}{2}\left(\left\|v_{k+1}-x^*\right\|^2-\left\|v_k-x^*\right\|^2\right)\\=&f(x_{k+1})-f(x_k)+\frac{\alpha_k}{2}(\mu-\gamma_{k+1})\left\|v_{k+1}-x^*\right\|^2\\&+\gamma_k\left(v_{k+1}-v_k,(v_{k+1}+v_k)/2-x^*\right).\end{aligned}$$

Analogously to the continuous level, we focus on the last term

$$\begin{aligned}&\gamma_k\left(v_{k+1}-v_k,(v_{k+1}+v_k)/2-x^*\right)\\&\quad=\gamma_k\left(v_{k+1}-v_k,v_{k+1}-x^*\right)-\frac{\gamma_k}{2}\left\|v_{k+1}-v_k\right\|^2.\end{aligned}$$

By (72), it follows that

$$\begin{aligned}&\gamma_k\left(v_{k+1}-v_k,v_{k+1}-x^*\right)\\&\quad=\mu\alpha_k\left(x_{k+1}-v_{k+1},v_{k+1}-x^*\right)-\alpha_k\left\langle\nabla f(x_{k+1}),v_{k+1}-x^*\right\rangle,\end{aligned}$$

and we use Lemma 1 to split the cross term into squares:

$$\begin{aligned}&2\left(x_{k+1}-v_{k+1},v_{k+1}-x^*\right)\\&\quad=\left\|x_{k+1}-x^*\right\|^2-\left\|x_{k+1}-v_{k+1}\right\|^2-\left\|v_{k+1}-x^*\right\|^2.\end{aligned}$$

For the gradient term, we have $v_{k+1}-x^*=v_{k+1}-x_{k+1}+x_{k+1}-x^*$ and use (72) to obtain

$$\begin{aligned}&-\alpha_k\left\langle\nabla f(x_{k+1}),v_{k+1}-x^*\right\rangle\\&\quad=-\left\langle\nabla f(x_{k+1}),x_{k+1}-x_k\right\rangle-\alpha_k\left\langle\nabla f(x_{k+1}),x_{k+1}-x^*\right\rangle.\end{aligned}$$

Consequently, using the μ-strongly convex property (cf.(2)) of f and dropping surplus negative square terms, we see

$$\mathcal{L}_{k+1}-\mathcal{L}_k\leq-\alpha_k\mathcal{L}_{k+1}.$$

This proves (75) and concludes the proof of this theorem. □

We observe from Theorem 3 that the fully implicit scheme (72) achieves linear convergence rate as long as $\alpha_k\geq\alpha>0$ for all $k>0$ and larger α_k yields faster convergence rate. We also mention that (72) can be rewritten as

$$\left\{\begin{aligned}x_{k+1}&=\mathbf{prox}_{\eta_k f}(y_k),\\v_{k+1}&=x_{k+1}+\frac{x_{k+1}-x_k}{\alpha_k},\end{aligned}\right.\tag{76}$$

where the proximal operator $\mathbf{prox}_{\eta_k f}$ has been introduced in (8) and

$$\gamma_{k+1} = \frac{\gamma_k + \mu\alpha_k}{1+\alpha_k}, \ \eta_k = \frac{\alpha_k^2}{\gamma_k + (\mu + \gamma_k)\alpha_k}, \ y_k = \frac{\gamma_k\alpha_k v_k + (\gamma_k + \mu\alpha_k)x_k}{\gamma_k + (\mu + \gamma_k)\alpha_k}.$$

Therefore, it allows f to be nonsmooth and we claim that Theorem 3 still holds true in this case. One just replaces the gradient $\nabla f(x_{k+1})$ with the subgradient $(y_k - x_{k+1})/\eta_k \in \partial f(x_{k+1})$; see (105) and (112).

For convex case, i.e., $\mu = 0$, our method (76) is very close to Güler's proximal point algorithm [20]

$$\begin{cases} x_{k+1} = \mathbf{prox}_{\eta_k f}(y_k), \quad \eta_k = \alpha_k^2/\gamma_{k+1}, \\ v_{k+1} = x_k + \dfrac{x_{k+1} - x_k}{\alpha_k}, \end{cases}$$

where $\gamma_{k+1} - \gamma_k = -\alpha_k\gamma_k$ and $y_k = \alpha_k v_k + (1-\alpha_k)x_k$. Indeed, with suitable step size, they share the similar rate; see [20, Theorem 2.3] and Theorem 4 below.

Theorem 4 *If f is proper, closed and convex and we choose $\alpha_k^2 = \eta_k\gamma_k(1+\alpha_k)$ with $\eta_k > 0$, then for the proximal point algorithm (76) with $\mu = 0$, we have*

$$\frac{\mathcal{L}_0}{\left(1 + \sum_{i=0}^{k-1}\sqrt{\gamma_0\eta_i}\right)^2} \le \mathcal{L}_k \le \frac{4\mathcal{L}_0}{\left(2 + \sum_{i=0}^{k-1}\sqrt{\gamma_0\eta_i}\right)^2}, \tag{77}$$

which means if $\sum_{k=0}^{\infty}\sqrt{\eta_k} = \infty$ then $\mathcal{L}_k \to 0$ as $k \to \infty$. Moreover, it holds that

$$\mathcal{L}_k \le \frac{4}{\sum_{i=0}^{k-1}\sqrt{\eta_i}}\left(\frac{1}{\gamma_0}\left(f(x_0) - f(x^*)\right) + \frac{1}{2}\left\|v_0 - x^*\right\|^2\right). \tag{78}$$

Proof For convenience and later use, define a sequence $\{\rho_k\}$ by that

$$\rho_0 = 1, \quad \rho_k := \prod_{i=0}^{k-1}\frac{1}{1+\alpha_i}, \quad k \ge 1. \tag{79}$$

As mentioned above, Theorem 3 holds true for such a nonsmooth f and thus it is evident that $\mathcal{L}_k \le \rho_k\mathcal{L}_0$. Invoking Lemma B2 proves (77) and it is trivial to obtain (78) from (77). This finishes the proof. □

Remark 2 Note that the sequence $\{\gamma_k\}$ in (73) is bounded: $0 < \gamma_k \le \max\{\mu, \gamma_0\}$ and $\gamma_k \to \mu$ as $k \to \infty$. Hence, even for large γ_0, the Lyapunov function $\mathcal{L}_k$ is asymptotically bounded as $k \to \infty$. In addition, from (77) and (78), we see that, for small γ_0, the convergence rate depends on γ_0 but large γ_0 does not pollute the final rate. This fact also holds true for all the forthcoming convergence bounds. □

5 Gauss–Seidel splitting with corrections

This section considers the Gauss–Seidel splitting (41), which is a semi-implicit discretization. In Sect. 2.4, we have established the spectrum bound $O(1-\sqrt{\mu/L})$ with step size $\alpha_k = O(\sqrt{\mu/L})$ for quadratic objectives. However, as we summarized in Sect. 2.6, spectral analysis is not sufficient for (norm) convergence.

Indeed, in the sequel, we further show that, for the discrete Lyapunov function (74), with any step size $\alpha_k > 0$, the naive discretization (41), reformulated as (80), does not lead to the contraction property like (75). This motivates us to add some proper correction steps.

5.1 The Gauss–Seidel splitting

Recall the Gauss–Seidel splitting (41): given step size $\alpha_k > 0$ and previous result (x_k, v_k), compute (x_{k+1}, v_{k+1}) from

$$\left\{\begin{aligned} \frac{x_{k+1}-x_k}{\alpha_k} &= v_k - x_{k+1}, \\ \frac{v_{k+1}-v_k}{\alpha_k} &= \frac{\mu}{\gamma_k}(x_{k+1}-v_{k+1}) - \frac{1}{\gamma_k}\nabla f(x_{k+1}). \end{aligned}\right. \tag{80}$$

In addition, the parameter Eq. (54) of γ is still discretized implicitly via (73).

Lemma 3 *If $f \in \mathcal{S}_\mu^1$ with $\mu \geq 0$, then for (80) with any step size $\alpha_k > 0$, we have*

$$\mathcal{L}_{k+1} - \mathcal{L}_k \leq -\alpha_k \mathcal{L}_{k+1} - \frac{\gamma_k}{2}\|v_{k+1}-v_k\|^2 - \alpha_k \langle \nabla f(x_{k+1}), v_{k+1}-v_k\rangle, \tag{81}$$

and

$$\mathcal{L}_{k+1} - \mathcal{L}_k \leq -\alpha_k \mathcal{L}_{k+1} + \frac{\alpha_k^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2. \tag{82}$$

Proof Following the proof of Theorem 3, we start from the difference

$$\begin{aligned} \mathcal{L}_{k+1} - \mathcal{L}_k =& f(x_{k+1}) - f(x_k) - \frac{\alpha_k\gamma_{k+1}}{2}\left\|v_{k+1}-x^*\right\|^2 \\ & - \frac{\mu\alpha_k}{2}\|x_{k+1}-v_{k+1}\|^2 - \frac{\gamma_k}{2}\|v_{k+1}-v_k\|^2 \\ & + \frac{\mu\alpha_k}{2}\left\|x_{k+1}-x^*\right\|^2 - \alpha_k\left\langle \nabla f(x_{k+1}), v_{k+1}-x^*\right\rangle. \end{aligned}$$

Using the update for x_{k+1} in (80), we split the gradient term as below

$$
\begin{aligned}
&- \alpha_k \left\langle \nabla f(x_{k+1}), v_{k+1} - x^* \right\rangle \\
&\quad = -\alpha_k \left\langle \nabla f(x_{k+1}), v_{k+1} - v_k \right\rangle - \left\langle \nabla f(x_{k+1}), \alpha_k (v_k - x_{k+1}) \right\rangle \\
&\qquad - \alpha_k \left\langle \nabla f(x_{k+1}), x_{k+1} - x^* \right\rangle \\
&\quad = -\alpha_k \left\langle \nabla f(x_{k+1}), v_{k+1} - v_k \right\rangle - \left\langle \nabla f(x_{k+1}), x_{k+1} - x_k \right\rangle \\
&\qquad - \alpha_k \left\langle \nabla f(x_{k+1}), x_{k+1} - x^* \right\rangle.
\end{aligned}
$$

As $f \in \mathcal{S}_{\mu}^1$, we obtain that

$$
\begin{aligned}
\mathcal{L}_{k+1} - \mathcal{L}_k \le & - \alpha_k \mathcal{L}_{k+1} - \frac{\gamma_k}{2} \left\| v_{k+1} - v_k \right\|^2 - \alpha_k \left\langle \nabla f(x_{k+1}), v_{k+1} - v_k \right\rangle \\
& - \frac{\mu \alpha_k}{2} \left\| x_{k+1} - v_{k+1} \right\|^2 - \frac{\mu}{2} \left\| x_{k+1} - x_k \right\|^2 .
\end{aligned}
$$

Ignoring all the negative terms of the second line, the above estimate implies (81).

As we see, different from (75), the estimate (81) contains a combination of a negative term and another cross term. Obviously, an easy application of Cauchy-Schwarz inequality yields

$$
-\frac{\gamma_k}{2} \left\| v_{k+1} - v_k \right\|^2 - \alpha_k \left\langle \nabla f(x_{k+1}), v_{k+1} - v_k \right\rangle \le \frac{\alpha_k^2}{2\gamma_k} \left\| \nabla f(x_{k+1}) \right\|_*^2 .
$$

This proves another bound (82) that only involves a positive gradient norm. □

5.2 A predictor–corrector method

To conquer the cross term $-\alpha_k \left\langle \nabla f(x_{k+1}), v_{k+1} - v_k \right\rangle$ in (81), we add an extra extrapolation step to (80) which can be thought as an semi-implicit discretization of $x' = v - x$ with the newest update v_{k+1}. More precisely, consider

$$
\left\{
\begin{aligned}
\frac{y_k - x_k}{\alpha_k} &= v_k - y_k, \\
\frac{v_{k+1} - v_k}{\alpha_k} &= \frac{\mu}{\gamma_k}(y_k - v_{k+1}) - \frac{1}{\gamma_k} \nabla f(y_k), \\
\frac{x_{k+1} - x_k}{\alpha_k} &= v_{k+1} - x_{k+1}.
\end{aligned}
\right.
\tag{83}
$$

This is in line with the spirit of the predictor-corrector method for ODE solvers [39, Section 3.8]. The variable y_k is the predictor produced by an explicit scheme and x_{k+1} is the corrector by an implicit scheme. It can be also thought of as a symmetric Gauss–Seidel iteration for approximating the implicit Euler method. Again, the parameter Eq. (54) of γ is still discretized via (73).

As the first two steps of (83) agree with (80), with x_{k+1} being y_k, recalling the estimate (81), we have

$$\widehat{\mathcal{L}}_k - \mathcal{L}_k \le -\alpha_k \widehat{\mathcal{L}}_k - \frac{\gamma_k}{2} \|v_{k+1} - v_k\|^2 - \alpha_k \langle \nabla f(y_k), v_{k+1} - v_k \rangle,$$

where

$$\widehat{\mathcal{L}}_k := f(y_k) - f(x^*) + \frac{\gamma_{k+1}}{2} \left\| v_{k+1} - x^* \right\|^2. \tag{84}$$

Therefore, it follows that

$$\widehat{\mathcal{L}}_k \le \frac{\mathcal{L}_k}{1+\alpha_k} - \frac{\gamma_k}{2(1+\alpha_k)} \|v_{k+1} - v_k\|^2 - \frac{\alpha_k}{1+\alpha_k} \langle \nabla f(y_k), v_{k+1} - v_k \rangle.$$

From the update for y_k and x_{k+1} in (83), we find the relation

$$x_{k+1} - y_k = \frac{\alpha_k}{1+\alpha_k}(v_{k+1} - v_k),$$

and if $f \in \mathcal{S}^{1,1}_{\mu,L}$, then there comes the estimate (cf. (4))

$$\begin{aligned}
\mathcal{L}_{k+1} - \widehat{\mathcal{L}}_k &= f(x_{k+1}) - f(y_k) \\
&\le \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{L}{2} \|x_{k+1} - y_k\|^2 \\
&= \frac{\alpha_k}{1+\alpha_k} \langle \nabla f(y_k), v_{k+1} - v_k \rangle \\
&\quad + \frac{L\alpha_k^2}{2(1+\alpha_k)^2} \|v_{k+1} - v_k\|^2.
\end{aligned}$$

As a result, we obtain

$$\mathcal{L}_{k+1} \le \frac{\mathcal{L}_k}{1+\alpha_k} + \left(\frac{L\alpha_k^2}{2(1+\alpha_k)^2} - \frac{\gamma_k}{2(1+\alpha_k)} \right) \|v_{k+1} - v_k\|^2. \tag{85}$$

The second term vanishes if we choose suitable step size; see the theorem below.

Theorem 5 *Assume that $f \in \mathcal{S}^{1,1}_{\mu,L}$ with $0 \le \mu \le L < \infty$ and $L\alpha_k^2 = \gamma_k(1+\alpha_k)$, then for the predictor-corrector scheme (83) together with (73), we have*

$$\mathcal{L}_{k+1} \le \frac{\mathcal{L}_k}{1+\alpha_k}, \quad k \in \mathbb{N}, \tag{86}$$

where $\mathcal{L}_k$ is defined by (74). Consequently, for all $k \ge 0$,

$$\mathcal{L}_k \le \mathcal{L}_0 \times \min \left\{ \frac{4L}{(\sqrt{\gamma_0}\, k + 2\sqrt{L})^2}, \left(1 + \sqrt{\frac{\min\{\gamma_0, \mu\}}{L}}\right)^{-k} \right\}, \tag{87}$$

and moreover, for all $k \geq 1$,

$$\mathcal{L}_k \leq C_{\gamma_0,L} \times \min\left\{\frac{4}{k^2}, \left(1+\sqrt{\frac{\min\{\gamma_0,\mu\}}{L}}\right)^{1-k}\right\}, \tag{88}$$

where

$$C_{\gamma_0,L} := \frac{L}{\gamma_0}\big(f(x_0)-f(x^*)\big) + \frac{L}{2}\left\|v_0 - x^*\right\|^2. \tag{89}$$

Proof The inequality (85) suggests the choice $L\alpha_k^2 = \gamma_k(1+\alpha_k)$ and promises (86). Recalling the sequence $\{\rho_k\}$ defined by (79), we have $\mathcal{L}_k \leq \rho_k\mathcal{L}_0$. Hence, using Lemma B2 gives the decay estimate of ρ_k and proves (87).

It remains to check (88) for all $k \geq 1$.

From Lemma B2 we easily get

$$\rho_k\mathcal{L}_0 \leq \left(f(x_0)-f(x^*)+\frac{\gamma_0}{2}\left\|v_0-x^*\right\|^2\right) \times \frac{4L}{(\sqrt{\gamma_0}\,k+2\sqrt{L})^2} \leq \frac{4C_{\gamma_0,L}}{k^2} \tag{90}$$

On the other hand, by the relation $L\alpha_0^2 = \gamma_0(1+\alpha_0)$, it is evident that

$$\alpha_0 = \frac{1}{2L}\left(\gamma_0 + \sqrt{4\gamma_0 L + \gamma_0^2}\right),$$

which implies

$$\frac{1}{1+\alpha_0} = \frac{2L}{\gamma_0 + 2L + \sqrt{4\gamma_0 L + \gamma_0^2}} \leq \frac{L}{\gamma_0}.$$

The above estimate also indicates that

$$\rho_k\mathcal{L}_0 = \frac{\mathcal{L}_0}{1+\alpha_0}\frac{\rho_k}{\rho_1} \leq C_{\gamma_0,L}\frac{\rho_k}{\rho_1} = C_{\gamma_0,L} \times \prod_{i=1}^{k-1}\frac{1}{1+\alpha_i}.$$

Applying Lemma B2 shows $\alpha_k \geq \sqrt{\min\{\gamma_0,\mu\}/L}$ and it follows that

$$\rho_k\mathcal{L}_0 \leq C_{\gamma_0,L} \times \left(1+\sqrt{\min\{\gamma_0,\mu\}/L}\right)^{1-k}.$$

Collecting this estimate and (90) establishes the final rate (88) and thus completes the proof of this theorem. □

Remark 3 We mention that the estimate (88) verifies the claim made previously in Remark 2. That is, the convergence rate given in Theorem 5 depends on small γ_0 but is robust when $\gamma_0 \geq L$. □

5.3 Correction via a gradient step

Motivated by the estimate (82), we can also aim to cancel the squared gradient norm. One preferable choice is the gradient descent step and according to our discussion below, any other correction step satisfying the decay property (94) is acceptable. Note that the two numerical schemes proposed in [38,44] for the HB Eq. (31) also have additional gradient steps.

As what we did before, replace x_{k+1} by y_k in (80) and consider the following corrected scheme: given $\alpha_k > 0$ and (x_k, v_k), compute (x_{k+1}, v_{k+1}) from

$$\left\{\begin{aligned} \frac{y_k - x_k}{\alpha_k} &= v_k - y_k, \\ \frac{v_{k+1} - v_k}{\alpha_k} &= \frac{\mu}{\gamma_k}(y_k - v_{k+1}) - \frac{1}{\gamma_k}\nabla f(y_k), \\ x_{k+1} - y_k &= -\frac{1}{L}\nabla f(y_k). \end{aligned}\right. \tag{91}$$

The implicit discretization (73) for the parameter Eq. (54) keeps unchanged here. In the first equation y_k can be solved in terms of the known data (x_k, v_k). After that, we evaluate the gradient $\nabla f(y_k)$ once and use it to update (x_{k+1}, v_{k+1}).

Theorem 6 *Assume that $f \in \mathcal{S}^{1,1}_{\mu,L}$ with $0 \le \mu \le L < \infty$ and $L\alpha_k^2 = \gamma_k(1+\alpha_k)$, then for the corrected scheme (91) together with (73), we have*

$$\mathcal{L}_{k+1} \le \frac{\mathcal{L}_k}{1+\alpha_k}, \quad k \in \mathbb{N}, \tag{92}$$

where $\mathcal{L}_k$ is defined by (74), and both the two estimates (87) and (88) hold true here.

Proof According to (82) in Lemma 3, we have established that

$$\widehat{\mathcal{L}}_k - \mathcal{L}_k \le -\alpha_k \widehat{\mathcal{L}}_k + \frac{\alpha_k^2}{2\gamma_k}\left\|\nabla f(y_k)\right\|_*^2, \tag{93}$$

where $\widehat{\mathcal{L}}_k$ is defined by (84). Thanks to the additional gradient step in (91), we have the basic gradient descent inequality:

$$f(x_{k+1}) - f(y_k) \le -\frac{1}{2L}\left\|\nabla f(y_k)\right\|_*^2, \tag{94}$$

which comes from (4) since $f \in \mathcal{S}^{1,1}_{\mu,L}$ and implies that

$$\mathcal{L}_{k+1} \le \widehat{\mathcal{L}}_k - \frac{1}{2L}\left\|\nabla f(y_k)\right\|_*^2.$$

Plugging this into (93) gives

$$\mathcal{L}_{k+1} - \mathcal{L}_k \leq -\alpha_k \mathcal{L}_{k+1} + \frac{1}{2L\gamma_k}\left(L\alpha_k^2 - \gamma_k(1+\alpha_k)\right)\|\nabla f(y_k)\|_*^2 .$$

This together with the condition $L\alpha_k^2 = \gamma_k(1+\alpha_k)$ yields (92).

As we choose the same step size as Theorem 5, based on the contraction (92), it is trivial to conclude that the two estimates (87) and (88) hold true here indeed. This completes the proof of this theorem. □

6 A corrected semi-implicit scheme from NAG method

In this section, we consider another semi-implicit scheme which comes exactly from Nesterov accelerated gradient method.

6.1 NAG method

In [29, Chapter 2, General scheme of optimal method], by using the estimate sequence, Nesterov presented an accelerated gradient method for solving (1) with $f \in \mathcal{S}_{\mu,L}^{1,1}$, $0 \leq \mu \leq L < \infty$; see Algorithm 1 below.

Algorithm 1 Nesterov Accelerated Gradient (NAG) Method

Input: $x_0, v_0 \in V$ and $\gamma_0 > 0$.
1: **for** $k = 0, 1, \ldots$ **do**
2: Compute $\alpha_k \in (0, 1)$ from $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \mu\alpha_k$.
3: Update $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \mu\alpha_k$.
4: Set $y_k = \dfrac{\alpha_k\gamma_k v_k + \gamma_{k+1}x_k}{\gamma_k + \mu\alpha_k}$.
5: Update x_{k+1} such that $f(x_{k+1}) \leq f(y_k) - \dfrac{1}{2L}\|\nabla f(y_k)\|_*^2$.
6: Update $v_{k+1} = \dfrac{1}{\gamma_{k+1}}\left[(1-\alpha_k)\gamma_k v_k + \alpha_k(\mu y_k - \nabla f(y_k))\right]$.
7: **end for**

Note that we have many choices for x_{k+1} in step 5 of Algorithm 1. One noticeable example is the gradient descent step (see [29, Chapter 2, Constant Step Scheme, I]):

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k). \tag{95}$$

With this choice, the sequence $\{v_k\}$ in Algorithm 1 can be eliminated and y_{k+1} is updated by that (see [29, Chapter 2, Constant Step Scheme, II])

$$y_{k+1} = x_{k+1} + \frac{\alpha_k - \alpha_k^2}{\alpha_{k+1} + \alpha_k^2}(x_{k+1} - x_k),$$

where $\alpha_{k+1} \in (0, 1)$ is calculated from the quadratic equation

$$L\alpha_{k+1}^2 = L\alpha_k^2(1 - \alpha_{k+1}) + \mu\alpha_{k+1}.$$

If $\mu > 0$ and $\alpha_0 = \sqrt{\mu/L}$, then $\alpha_k = \sqrt{\mu/L}$; see [29, Chapter 2, Constant Step Scheme, III]. In particular, if $\mu = 0$, then Algorithm 1 (with x_{k+1} updated by (95)) coincides with the accelerated scheme proposed by Nesterov early in the 1980s [27].

6.2 NAG method as a corrected semi-implicit scheme

After simple calculations, we can rewrite Algorithm 1 as an equivalent form

$$\begin{cases} \dfrac{\gamma_{k+1} - \gamma_k}{\alpha_k} = \mu - \gamma_k, \\ \dfrac{y_k - x_k}{\alpha_k} = \dfrac{\gamma_k}{\gamma_{k+1}}(v_k - y_k), \\ \dfrac{v_{k+1} - v_k}{\alpha_k} = \dfrac{\mu}{\gamma_{k+1}}(y_k - v_k) - \dfrac{1}{\gamma_{k+1}}\nabla f(y_k), \end{cases} \tag{96}$$

where in addition we update x_{k+1} satisfying

$$f(x_{k+1}) \le f(y_k) - \frac{1}{2L}\|\nabla f(y_k)\|_*^2. \tag{97}$$

Surprisingly, (96) formulates a semi-implicit discretization for our NAG flow (56) with a correction step (97) and an explicit discretization for the Eq. (54) of γ. Similarly to (91), we can adopt the gradient descent step which promises (97).

Based on subtle algebraic calculations of the estimate sequence, Nesterov [29, Chapter 2] proved the convergence rate of Algorithm 1. In the following, we give an alternative proof by using the Lyapunov function (74).

Theorem 7 *Assume that $f \in \mathcal{S}_{\mu,L}^{1,1}$ with $0 \le \mu \le L < \infty$. If $L\alpha_k^2 = \gamma_{k+1}$, then for Algorithm 1, i.e., the scheme (96) together with (97), we have $0 < \alpha_k \le 1$ and*

$$\mathcal{L}_{k+1} \le (1 - \alpha_k)\mathcal{L}_k, \quad k \in \mathbb{N}, \tag{98}$$

where $\mathcal{L}_k$ is defined by (74). Consequently for all $k \ge 0$,

$$\mathcal{L}_k \le \mathcal{L}_0 \times \min\left\{\frac{4L}{(\sqrt{\gamma_0}\,k + 2\sqrt{L})^2}, \left(1 - \sqrt{\frac{\min\{\gamma_1, \mu\}}{L}}\right)^k\right\}. \tag{99}$$

Moreover, for all $k \ge 1$,

$$\mathcal{L}_k \le C_{\gamma_0, L} \times \min\left\{\frac{4}{k^2}, \left(1 - \sqrt{\frac{\min\{\gamma_1, \mu\}}{L}}\right)^{k-1}\right\}, \tag{100}$$

where $C_{\gamma_0,L}$ has been defined in (89).

Proof Let us first prove (98). By (96), we find

$$\begin{cases} v_k = y_k + \dfrac{\gamma_{k+1}}{\alpha_k\gamma_k}(y_k - x_k), \\ v_{k+1} = y_k + \dfrac{1-\alpha_k}{\alpha_k}(y_k - x_k) - \dfrac{\alpha_k}{\gamma_{k+1}}\nabla f(y_k), \end{cases}$$

and a direct computation gives

$$\begin{aligned} &\frac{\gamma_{k+1}}{2}\left\|v_{k+1} - x^*\right\|^2 - \frac{\gamma_k}{2}(1-\alpha_k)\left\|v_k - x^*\right\|^2 \\ &\quad = \alpha_k\left(\left\langle\nabla f(y_k), x^* - y_k\right\rangle + \frac{\mu}{2}\left\|x^* - y_k\right\|^2\right) \\ &\qquad + (1-\alpha_k)\left(\left\langle\nabla f(y_k), x_k - y_k\right\rangle + \frac{\mu}{2}\left\|x_k - y_k\right\|^2\right) \\ &\qquad + \frac{\alpha_k^2}{2\gamma_{k+1}}\left\|\nabla f(y_k)\right\|_*^2 - \frac{\mu(1-\alpha_k)}{2\alpha_k\gamma_k}(\gamma_k + \mu\alpha_k)\left\|y_k - x_k\right\|^2. \end{aligned}$$

Dropping the negative term $-\left\|y_k - x_k\right\|^2$ and using the μ-convexity of f imply that

$$\begin{aligned} &\frac{\gamma_{k+1}}{2}\left\|v_{k+1} - x^*\right\|^2 - \frac{\gamma_k}{2}(1-\alpha_k)\left\|v_k - x^*\right\|^2 \\ &\quad \le \alpha_k\left(f(x^*) - f(y_k)\right) + (1-\alpha_k)\left(f(x_k) - f(y_k)\right) + \frac{\alpha_k^2}{2\gamma_{k+1}}\left\|\nabla f(y_k)\right\|_*^2, \end{aligned}$$

and we get the inequality

$$\mathcal{L}_{k+1} - (1-\alpha_k)\mathcal{L}_k \le f(x_{k+1}) - f(y_k) + \frac{\alpha_k^2}{2\gamma_{k+1}}\left\|\nabla f(y_k)\right\|_*^2.$$

Consequently, by (97) and the relation $L\alpha_k^2 = \gamma_{k+1}$, the right hand side of the above inequality is negative, which proves (98).

In this case, we modify (79) as follows

$$\rho_0 = 1, \quad \rho_k := \prod_{i=0}^{k-1}(1-\alpha_i), \quad k \ge 1, \tag{101}$$

then by (98) it is clear that $\mathcal{L}_k \le \rho_k\mathcal{L}_0$, and invoking Lemma B1 proves (99). As the proof of (100) is very similar with that of (88), we omit the details here and conclude the proof of this theorem. □

Remark 4 Similar to our corrected schemes (83) and (91), NAG method (i.e., Algorithm 1) generates a three-term sequence $\{(x_k, y_k, v_k)\}$ as well. If $\mu = 0$, then they share the same convergence rate bound

$$\mathcal{L}_k \le \frac{4L\mathcal{L}_0}{(\sqrt{\gamma_0}\,k + 2\sqrt{L})^2},$$

and when $\gamma_0 = \mu > 0$, we have

$$\mathcal{L}_k \le \mathcal{L}_0 \times \begin{cases} (1 - \sqrt{\mu/L})^k, & \text{for NAG method,} \\ (1 + \sqrt{\mu/L})^{-k}, & \text{for (91) and (83) .} \end{cases} \tag{102}$$

In view of the trivial fact

$$1 - \epsilon = \frac{1}{1+\epsilon} - \frac{\epsilon^2}{1+\epsilon}, \quad \epsilon = \sqrt{\mu/L} \le 1,$$

we see the rates in (102) are asymptotically the same and NAG method can achieve a slightly better convergence rate. However, we note that they share the same computational complexity

$$O\left(\min\left\{\sqrt{L/\epsilon},\ \sqrt{L/\mu}\cdot|\ln\epsilon|\right\}\right),$$

which is optimal, in the sense that [29] it achieves the complexity lower bound of first-order algorithms for the function class $\mathcal{S}^{1,1}_{\mu,L}$ with $0 \le \mu \le L < \infty$. □

Remark 5 Unlike the gradient descent method, the function value $f(x_k)$ of accelerated gradient methods may not decrease in each step. It is the discrete Lyapunov function $\mathcal{L}_k$ that is always decreasing; see (86), (92) and (98). □

Remark 6 To reduce the function value, one can adopt the restating strategy [31]. Specifically, given (γ_0, v_0, x_0), if $f(x_k)$ is increasing after k-iteration, then set $k = 0$ and restart the iteration process with another initial guess $(\tilde{\gamma}_0, \tilde{v}_0, \tilde{x}_0)$. By Theorems 5, 6 and 7, when $f \in \mathcal{S}^{1,1}_{0,L}$ and $\gamma_0 = L$, $v_0 = x_0$, we only have the sublinear convergence rate

$$f(x_k) - f(x^*) \le \frac{4}{k^2}\left(f(x_0) - f(x^*) + \frac{L}{2}\left\|x_0 - x^*\right\|^2\right) \le \frac{4L}{k^2}\left\|x_0 - x^*\right\|^2, \tag{103}$$

where we used (4), which promises

$$f(x_0) - f(x^*) \le \frac{L}{2}\left\|x_0 - x^*\right\|^2.$$

Additionally, assume f satisfies the *quadratic growth condition* with $\sigma > 0$:

$$f(x) - f(x^*) \ge \sigma\,\mathrm{dist}^2(x, \operatorname{argmin} f) \quad \forall\, x \in V,$$

where $\operatorname{dist}(x, \operatorname{argmin} f) = \inf_{x^* \in \operatorname{argmin} f} \|x - x^*\|$. As (103) holds for all $x^* \in \operatorname{argmin} f$, we have immediately that

$$f(x_k) - f(x^*) \le \frac{4L}{k^2} \operatorname{dist}^2(x, \operatorname{argmin} f) \le \frac{4L}{\sigma k^2}(f(x_0) - f(x^*)).$$

Therefore, as analyzed in [30], if we consider fixed restart technique [31] every k steps, then after $N = nk$ steps we will get

$$f(x_N) - f(x^*) \le \left(\frac{4L}{\sigma k^2}\right)^n (f(x_0) - f(x^*)).$$

Evidently, the optimal choice $k_\# = e\sqrt{4L/\sigma}$ yields the linear rate

$$f(x_N) - f(x^*) \le e^{-2N/k_\#}(f(x_0) - f(x^*)).$$

If the parameter σ is unknown, one can use the adaptive restart technique [31].

When f is quadratic and convex, changing γ_k from L to μ periodically will smoothing out error in different frequencies and can further optimize the constant in front of the accelerated rate. That is, the dynamically changing parameter $\{\gamma_k\}$ hopefully outperforms the fixed one $\gamma_k = \mu$. For general nonlinear convex functions, a rigorous justification of the restart strategy is under investigation. □

7 Composite convex optimization

In this part we mainly focus on the composite optimization

$$\min_{x \in Q} f(x) := \min_{x \in Q} \left[h(x) + g(x)\right], \tag{104}$$

where $Q \subseteq V$ is a simple closed convex set, $h \in \mathcal{S}^{1,1}_{\mu,L}(Q)$ with $0 \le \mu \le L < \infty$ and $g : V \to \mathbb{R} \cup \{+\infty\}$ is proper, closed and convex, and $Q \cap \mathbf{dom}\, g \neq \emptyset$. In general g is not differentiable but its subdifferential ∂g exists as a set-valued function. More precisely, the subdifferential $\partial g(x)$ of g at x is defined by that

$$\partial g(x) := \left\{p \in V^* : g(y) \ge g(x) + \langle p, y - x\rangle \quad \forall\, y \in V\right\}. \tag{105}$$

Remark 7 For the case that $h \in \mathcal{S}^{1,1}_{0,L}(Q)$ and g is μ-strongly convex with $\mu \ge 0$, we can split $h + g$ as $(h(x) + \frac{\mu}{2}\|x\|^2) + (g(x) - \frac{\mu}{2}\|x\|^2)$, which reduces to our current assumption for (104). □

We shall apply our ODE solver approach to the problem (104). The first step is to generalize the dynamical system (56) to the current nonsmooth setting. Basically, we set $F = f + i_Q$ with i_Q being the indicator function of Q and obtain a differential inclusion for minimizing F on V, which is equivalent to minimize f over Q. After that, optimization methods (see Algorithms 2 and 4) for solving the original problem (104) with the accelerated convergence rate

$$O\left(\min\left\{L/k^2, (1+\sqrt{\mu/L})^{-k}\right\}\right)$$

are proposed from numerical discretizations of the continuous model (106). This is a proof of the effective and usefulness of our NAG flow model (106) and the ODE solver approach, by which we can construct new accelerated methods easily.

7.1 Continuous model

For minimizing a nonsmooth function F over V, our NAG flow (56) becomes a differential inclusion

$$\begin{cases} x' = v - x, \\ \gamma v' \in \mu(x - v) - \partial F(x). \end{cases} \tag{106}$$

To ensure solution existence, suitable initial conditions shall be imposed later. Correspondingly, the second-order ODE (57) reads as a second-order differential inclusion

$$\gamma x'' + (\mu + \gamma)x' + \partial F(x) \ni 0. \tag{107}$$

As the subdifferential ∂F is a set-valued maximal monotone operator, classical C^2 solution to (107) may not exist because discontinuity can occur in x'. Therefore, the concept of *energy-conserving solution* has been introduced in [15,32,36].

Let us assume the initial data

$$x(0) = x_0 \in \mathbf{dom} F \quad \text{and} \quad x'(0) = x_1 \in \mathcal{T}_{\mathbf{dom} F}(x_0), \tag{108}$$

where $\mathcal{T}_{\mathbf{dom} F}(x_0)$ denotes the tangent cone of $\mathbf{dom} F$ at x_0:

$$\mathcal{T}_{\mathbf{dom} F}(x_0) := \overline{\bigcup_{\tau>0} \tau(x_0 - \overline{\mathbf{dom} F})}.$$

In addition, we shall introduce some vector-valued functional spaces. Given any interval $I \subset \mathbb{R}$, let $M(I; V)$ be the space of V-valued Radon measures on I; for any $m \in \mathbb{N}$ and $1 \le p \le \infty$, $W^{m,p}(I; V)$ denotes the standard V-valued Sobolev space [21]; the space of all V-valued functions with bounded variation is defined by $BV(I; V)$ [4]. Also, $W^{m,p}_{\mathrm{loc}}(I; V)$ and $BV_{\mathrm{loc}}(I; V)$ consist of all the sets $W^{m,p}(\omega; V)$ and $BV(\omega; V)$ respectively, where $\omega \subset I$ is any compact subset.

Definition 1 We call $x : [0, \infty) \to V$ an energy-conserving solution to (107) with initial data (108) if it satisfies the following.

1. $x \in W^{1,\infty}_{\mathrm{loc}}(0, \infty; V)$, $x(0) = x_0$ and $x(t) \in \mathbf{dom} F$ for all $t > 0$.
2. $x' \in BV_{\mathrm{loc}}([0, \infty); V)$, $x'(0+) = x_1$.

3. For almost all $t > 0$, there holds the energy equality:

$$F(x(t)) + \frac{\gamma(t)}{2} \left\| x'(t) \right\|^2 + \int_0^t \frac{\mu + 3\gamma(s)}{2} \left\| x'(s) \right\|^2 \mathrm{d}s = F(x_0) + \frac{\gamma_0}{2} \left\| x_1 \right\|^2.$$

4. There exists some $\nu \in M(0, \infty; V)$ such that

$$\gamma x'' + (\mu + \gamma) x' + \nu = 0$$

holds in the sense of distributions, and for any $T > 0$, we have

$$\int_0^T \big(F(y(t)) - F(x(t))\big)\mathrm{d}t \geq \langle \nu, y - x \rangle_{C([0,T];V)} \quad \text{for all } y \in C([0, T]; V).$$

In [25], the problem (107) has been extended to a general case

$$\gamma x'' + (\mu + \gamma) x' + \partial F(x) \ni \xi,$$

where ξ stands for small perturbation. Therefore, according to [25, Theorem 2.1], we have the existence of an energy-conserving solution to (107) and by [25, Theorems 2.2 and 2.3], we obtain the exponential decay, which is a nonsmooth version of (60).

Theorem 8 *Assume V is a finite dimensional Hilbert space. In the sense of Definition 1, the differential inclusion (107) admits an energy-conserving solution $x : [0, \infty) \to V$ satisfying*

$$F(x(t)) - F(x^*) + \frac{\gamma(t)}{2} \left\| x(t) + x'(t) - x^* \right\|^2 \leq 2\mathcal{L}_0 e^{-t}, \tag{109}$$

for almost all $t > 0$, where $\mathcal{L}_0 := F(x_0) - F(x^*) + \frac{\gamma_0}{2} \left\| x_0 + x_1 - x^* \right\|^2$.

Remark 8 If additionally $\mathbf{dom} F = V$, then $x \in W^{2,\infty}_{\text{loc}}(0, \infty; V) \cap C^1([0, \infty); V)$ and (109) holds for all $t > 0$. □

7.2 An APGM for unconstrained optimization

Let us first consider the unconstrained case $Q = V$, i.e.,

$$\min_{x \in V} f(x) := \min_{x \in V} \left[h(x) + g(x) \right], \tag{110}$$

where $f \in \mathcal{S}^{1,1}_{\mu,L}$ with $0 \leq \mu \leq L < \infty$ and $g : V \to \mathbb{R} \cup \{+\infty\}$ is a properly closed and convex function and possibly nonsmooth.

7.2.1 Gradient mapping

To treat the nonsmooth part g, we introduce the tool of gradient mapping. Following [29, Chapter 2], given any $\eta > 0$, the composite gradient mapping $\mathcal{G}_f(x, \eta)$ of f at x is defined by that

$$\mathcal{G}_f(x, \eta) := \frac{x - S_f(x, \eta)}{\eta} \quad x \in V, \tag{111}$$

where $S_f(x, \eta) := \mathbf{prox}_{\eta g}(x - \eta \nabla h(x))$ and the proximal operator $\mathbf{prox}_{\eta g}$ has been defined by (8). Note that $S_f(x, \eta)$ is clearly well-defined and so is $\mathcal{G}_f(x, \eta)$.

It is well known [33,35] that

$$\frac{x - \mathbf{prox}_{\eta g}(x)}{\eta} \in \partial g(\mathbf{prox}_{\eta g}(x)), \tag{112}$$

which yields the fact

$$\mathcal{G}_f(x, \eta) - \nabla h(x) \in \partial g(S_f(x, \eta)). \tag{113}$$

From this we conclude that the fixed-point set of $S_f(\cdot, \eta)$ is argmin f. Indeed, $x = S_f(x, \eta)$ if and only if $0 \in \partial f(x)$. We also observe from (113) that the gradient mapping (111) is defined reversely from the proximal-gradient step for minimizing $f = h + g$, i.e.,

$$\begin{aligned}\frac{S_f(x, \eta) - x}{\eta} &\in -\nabla h(x) - \partial g(S_f(x, \eta)) \\ &= -\mathcal{G}_f(x, \eta).\end{aligned}$$

Hence it plays the role of the gradient ∇f in the smooth case. Particularly, if $g = 0$, then $\mathcal{G}_f(x, \eta) = \nabla h(x)$ and $S_f(x, \eta) = x - \eta \nabla h(x)$ is nothing but a gradient step.

To move on, we present an auxiliary lemma, which is a key ingredient for our convergence analysis. As we will fix $\eta = 1/L$, for simplicity, we set $\mathcal{G}_f(x) := \mathcal{G}_f(x, 1/L)$ and $S_f(x) := S_f(x, 1/L)$.

Lemma 4 *Assume $f = h + g$, where $h \in \mathcal{S}^{1,1}_{\mu,L}$ with $0 \leq \mu \leq L < \infty$ and $g : V \to \mathbb{R} \cup \{+\infty\}$ is properly closed and convex. Then for any $x, y \in V$,*

$$f(y) \geq f(S_f(x)) + \left\langle \mathcal{G}_f(x), y - x \right\rangle + \frac{\mu}{2} \|y - x\|^2 + \frac{1}{2L} \left\| \mathcal{G}_f(x) \right\|^2. \tag{114}$$

Proof Since $h \in \mathcal{S}^{1,1}_{\mu,L}$, applying (2) and (4) gives

$$\begin{aligned}h(x) - h(y) + \langle \nabla h(x), y - x \rangle &\leq -\frac{\mu}{2} \|x - y\|^2, \\ h(S_f(x)) - h(x) + \left\langle \nabla h(x), x - S_f(x) \right\rangle &\leq \frac{L}{2} \left\| S_f(x) - x \right\|^2,\end{aligned}$$

which implies that

$$h(y) \geq h(S_f(x, \eta)) + \left\langle \nabla h(x), y - S_f(x) \right\rangle + \frac{\mu}{2} \|y - x\|^2 - \frac{1}{2L} \left\| \mathcal{G}_f(x) \right\|^2.$$

Observing (113), we get

$$g(y) \geq g(S_f(x)) + \left\langle \mathcal{G}_f(x) - \nabla h(x), y - S_f(x) \right\rangle.$$

Summing the above two inequalities and using the split

$$\begin{aligned} \left\langle \mathcal{G}_f(x), y - S_f(x) \right\rangle &= \left\langle \mathcal{G}_f(x), y - x \right\rangle + \left\langle \mathcal{G}_f(x), x - S_f(x) \right\rangle \\ &= \left\langle \mathcal{G}_f(x), y - x \right\rangle + \frac{1}{L} \left\| \mathcal{G}_f(x) \right\|^2, \end{aligned}$$

we finally arrive at (114) and end the proof of this lemma. □

Remark 9 For a fixed x, the right hand side of (114) defines a quadratic approximation of f at x, and it is strongly reminiscent of the quadratic lower bound approximation (2) for the smooth case. However, compared to (2), the constant is shifted from $f(x)$ to a lower value $f(S_f(x)) + \frac{1}{2L} \left\| \mathcal{G}_f(x) \right\|^2$. The first order part is $\mathcal{G}_f(x)$ instead of the subgradient at x. The quadratic part $\frac{\mu}{2} \|y - x\|^2$ is due to the μ-convexity. □

7.2.2 The proposed method

Based on the corrected semi-implicit scheme (91) for NAG flow (56), it is possible to generalize it to solve the differential inclusion (106). We simply replace the gradient $\nabla f(y_k)$ with the gradient mapping $\mathcal{G}_f(y_k)$ and set the correction as $x_{k+1} = S_f(y_k)$. More precisely, consider

$$\left\{ \begin{aligned} \frac{y_k - x_k}{\alpha_k} &= v_k - y_k, \\ x_{k+1} &= S_f(y_k), \\ \frac{v_{k+1} - v_k}{\alpha_k} &= \frac{\mu}{\gamma_k}(y_k - v_{k+1}) - \frac{1}{\gamma_k} \mathcal{G}_f(y_k), \\ \frac{\gamma_{k+1} - \gamma_k}{\alpha_k} &= \mu - \gamma_{k+1}. \end{aligned} \right. \tag{115}$$

Once $x_{k+1} = S_f(y_k) = \mathbf{prox}_{\eta g}(y_k - \eta \nabla h(y_k))$ is obtained, we can update v_{k+1} with known datum x_k, y_k, v_k and x_{k+1}. Thus in each iteration, (115) only calls the proximal operation $\mathbf{prox}_{\eta g}$ once.

We still use the step size $L\alpha_k^2 = \gamma_k(1 + \alpha_k)$ and summarize the semi-implicit scheme (115) in Algorithm 2, which is called semi-implicit APGM (Semi-APGM for short). Also, the convergence rate is derived via the discrete Lyapunov function (74).

Algorithm 2 Semi-APGM for solving $\min_{x\in V}\,[h(x)+g(x)]$

Input: $x_0, v_0 \in V,\ \gamma_0 > 0$ and $\eta = 1/L$.
1: **for** $k = 0, 1, \dots$ **do**
2: Compute $\alpha_k > 0$ such that $L\alpha_k^2 = \gamma_k\big(1+\alpha_k\big)$.
3: Update $\gamma_{k+1} = \dfrac{\gamma_k + \mu\alpha_k}{1+\alpha_k}$.
4: Set $y_k = \dfrac{x_k + \alpha_k v_k}{1+\alpha_k}$ and $w_k = \dfrac{\gamma_k v_k + \mu\alpha_k y_k}{\gamma_k + \mu\alpha_k}$.
5: Update $x_{k+1} = \mathbf{prox}_{\eta g}(y_k - \eta\nabla h(y_k))$.
6: Set $v_{k+1} = w_k + \dfrac{\gamma_k}{\gamma_{k+1}}\dfrac{x_{k+1}-y_k}{\alpha_k}$.
7: **end for**

Theorem 9 *For Algorithm 2, we have*

$$\mathcal{L}_{k+1} \le \frac{\mathcal{L}_k}{1+\alpha_k} \quad \forall\, k \in \mathbb{N}, \tag{116}$$

where $\mathcal{L}_k = f(x_k) - f(x^*) + \frac{\gamma_k}{2}\|v_k - x^*\|^2$*, and both (87) and (88) hold true here.*

Proof The proof of (116) is very similar to that of (92). Replacing x_{k+1} and its gradient $\nabla f(x_{k+1})$ in (80) respectively with y_k and $\mathcal{G}_f(y_k)$, we can proceed as the proof of Lemma 3 and use Lemma 4 to obtain

$$\begin{aligned}\widehat{\mathcal{L}}_k - \mathcal{L}_k \le{}& -\alpha_k\widehat{\mathcal{L}}_k + (1+\alpha_k)\left(f(y_k) - f(x_{k+1})\right)\\ &+ \frac{\alpha_k^2}{2\gamma_k}\left\|\mathcal{G}_f(y_k)\right\|^2 - \frac{1+\alpha_k}{2L}\left\|\mathcal{G}_f(y_k)\right\|^2,\end{aligned} \tag{117}$$

where $\widehat{\mathcal{L}}_k$ is defined by (84). Thanks to the relation $L\alpha_k^2 = \gamma_k(1+\alpha_k)$, the second line of (117) vanishes, and inserting the identity $f(y_k) - f(x_{k+1}) = \widehat{\mathcal{L}}_k - \mathcal{L}_{k+1}$ into (117) gives (116). Based on this, it is not hard to see that both (87) and (88) hold true. This finishes the proof of this theorem. ◻

We mention that with another choice

$$L\alpha_k^2 = \mu\alpha_k^2 + \gamma_k(1+\alpha_k),$$

we can drop the sequence $\{v_k\}$ from (115). The procedure is not straightforward but very similar to that of Nesterov's optimal method in [29, page 80]. We omit the details and only list the following algorithm.

Algorithm 3 Simplified Semi-APGM

Input: $x_0, y_0 \in V$, $\gamma_0 > 0$ and $\eta = 1/L$.
1: **for** $k = 0, 1, \dots$ **do**
2: Compute $\alpha_k > 0$ such that $L\alpha_k^2 = \mu\alpha_k^2 + \gamma_k(1+\alpha_k)$.
3: Update $\gamma_{k+1} = \dfrac{\gamma_k + \mu\alpha_k}{1+\alpha_k}$ and set $\beta_k = \frac{L\alpha_k}{\gamma_{k+1}(1+\alpha_k)}$.
4: Set $y_{k+1} = x_k + \beta_k(x_{k+1} - x_k)$.
5: Update $x_{k+1} = \mathbf{prox}_{\eta g}(y_k - \eta\nabla h(y_k))$.
6: **end for**

This can be viewed as a generalization of [29, Chapter 2, Constant Step Scheme, II] to problem (110). Particularly, for convex case $\mu = 0$, it is very close to FISTA [12]. Both of them share the same spirit: applying one proximal gradient step first and then using some extrapolation formulae. The difference comes only from the use of the two sequences $\{\alpha_k\}$ and $\{\beta_k\}$. We also claim that Algorithm 3 has the same accelerated convergence rate as Algorithm 2, i.e., $O(\min(L/k^2, (1+\sqrt{\mu/L})^{-k}))$. In contrast FISTA is designed for $\mu = 0$ and has only the sublinear rate $O(L/k^2)$.

We mention that, accelerated proximal gradient methods for solving (110) with only one evaluation of $\mathbf{prox}_{\eta g}$ in each iteration can be found in [38] (only for strongly convex case) and [24, Chapter 2, Algorithm 2.2] (for both convex and strongly convex cases).

Both Algorithms 2 and 3 cannot be applied directly to the general constraint case (104). The main issue comes from the definition (111) of the gradient mapping $\mathcal{G}_f(x, \eta)$, where we impose the restriction $x \in Q$ and calculate the proximal operator $\mathbf{prox}_{\eta g}$ over Q to obtain $S_f(x) \in Q$. For both two algorithms, we shall compute $x_{k+1} = S_f(y_k) = \mathbf{prox}_{\eta g}(y_k - \eta\nabla h(y_k))$. But the sequence $\{y_k\}$ in Algorithms 2 and 3 may be outside the constraint set. This is not acceptable because $\nabla h(y_k)$ might not exist: for instance, $Q = [0, \infty)$ and h is the entropy function.

The original FISTA [12] and the methods in [38] and [24, Chapter 2, Algorithm 2.2] mentioned above, cannot be applied to the constrained problem (104) either. This stimulates us to propose a new operator splitting scheme to conquer this problem.

7.3 An accelerated forward–backward method for constrained optimization

We now go back to the constrained problem (104). As mentioned above, the tool of gradient mapping is not convenient for us to handle this case. To avoid using it, we utilize the separable structure of $f = h + g$ and apply explicit and implicit schemes for h and g, respectively. This is the so-called *operator splitting* technique in ODE solvers and is also known as the forward-backward method.

Let us start from the predictor-corrector scheme (83) and rewrite it as follows

$$\begin{cases} y_k = \dfrac{x_k + \alpha_k v_k}{1+\alpha_k}, \quad w_k = \dfrac{\gamma_k v_k + \mu\alpha_k y_k}{\gamma_k + \mu\alpha_k}, \\ v_{k+1} = \mathop{\mathrm{argmin}}\limits_{v\in V} \left\{ \langle \nabla f(y_k), v\rangle + \dfrac{\gamma_k + \mu\alpha_k}{2\alpha_k}\|v - w_k\|^2 \right\}, \\ x_{k+1} = \dfrac{x_k + \alpha_k v_{k+1}}{1+\alpha_k}. \end{cases} \tag{118}$$

For minimizing $f = h + g$ over Q, we modify the above method as follows

$$\begin{cases} y_k = \dfrac{x_k + \alpha_k v_k}{1+\alpha_k}, \quad w_k = \dfrac{\gamma_k v_k + \mu\alpha_k y_k}{\gamma_k + \mu\alpha_k}, \\ v_{k+1} = \mathop{\mathrm{argmin}}\limits_{v\in Q} \left\{ g(v) + \langle \nabla h(y_k), v\rangle + \dfrac{\gamma_k + \mu\alpha_k}{2\alpha_k}\|v - w_k\|^2 \right\}, \\ x_{k+1} = \dfrac{x_k + \alpha_k v_{k+1}}{1+\alpha_k}, \end{cases} \tag{119}$$

where $x_0, v_0 \in Q$ and the parameter sequence $\{\gamma_k\}$ comes from the implicit discretization (73) of the Eq. (54). Clearly, as convex combinations are used, the method (119) preserves the three-term sequence $\{(x_k, y_k, v_k)\}$ in Q and it requires the proximal computation of g over Q only once in each iteration.

We choose $L\alpha_k^2 = \gamma_k(1+\alpha_k)$ as before and rewrite (119) in Algorithm 4, which is called semi-implicit accelerated forward-backward (Semi-AFB for short) method.

Algorithm 4 Semi-AFB method for solving $\min_{x\in Q}\,[h(x) + g(x)]$

Input: $x_0, v_0 \in Q,\ \gamma_0 > 0$ and $L > 0$.
1: **for** $k = 0, 1, \ldots$ **do**
2: Compute $\alpha_k > 0$ such that $L\alpha_k^2 = \gamma_k(1+\alpha_k)$.
3: Update $\gamma_{k+1} = \dfrac{\gamma_k + \mu\alpha_k}{1+\alpha_k}$.
4: Set $y_k = \dfrac{x_k + \alpha_k v_k}{1+\alpha_k}$ and $w_k = \dfrac{\gamma_k v_k + \mu\alpha_k y_k}{\gamma_k + \mu\alpha_k}$.
5: Update $v_{k+1} = \mathop{\mathrm{argmin}}\limits_{v\in Q} \left\{ g(v) + \langle \nabla h(y_k), v\rangle + \dfrac{\gamma_k + \mu\alpha_k}{2\alpha_k}\|v - w_k\|^2 \right\}$.
6: Update $x_{k+1} = \dfrac{x_k + \alpha_k v_{k+1}}{1+\alpha_k}$.
7: **end for**

In [41], Tseng considered problem (104) only with convex assumption, i.e., $\mu = 0$, and proposed an APGM that possesses the rate $O(L/k^2)$. By using the technique of estimate sequence, Nesterov [28] presented an accelerated method for solving (104) with the assumption that h is L-smooth over Q and g is μ-strongly convex with $\mu \geq 0$. Both our Algorithm 4 and Nesterov's method generate a three-term sequence $\{(x_k, y_k, v_k)\}$ and have the same accelerated rate $O(\min(L/k^2, (1+\sqrt{\mu/L})^{-k}))$; see [28, Theorem 6] and our Theorem 10. However, as mentioned in [12], the later used an accumulated history of the past iterations to build recursively a sequence of estimate functions, and in each iteration, to update x_{k+1} and v_{k+1}, Nesterov's method in [28] calls $\mathbf{prox}_g$ over Q twice.

Below, we shall establish the convergence rate of Algorithm 4 via the analysis of a Lyapunov function. It is well known [28, Eq (2.9)] that the first-order optimality condition for v_{k+1} in (119) is the variational inequality

$$\left\langle \nabla h(y_k) + \frac{\gamma_k + \mu\alpha_k}{\alpha_k}(v_{k+1} - w_k) + p_{k+1}, x - v_{k+1} \right\rangle \geq 0 \quad \forall x \in Q,$$

where $p_{k+1} \in \partial g(v_{k+1})$. Expanding w_k, we observe the relation

$$\begin{aligned} &\gamma_k \left(v_{k+1} - v_k, v_{k+1} - x\right) \\ &\quad \leq \mu\alpha_k \left(y_k - v_{k+1}, v_{k+1} - x\right) - \alpha_k \left\langle \nabla h(y_k) + p_{k+1}, v_{k+1} - x \right\rangle, \end{aligned} \tag{120}$$

where $x \in Q$ is arbitrary.

Theorem 10 *For Algorithm 4, we have*

$$\mathcal{L}_{k+1} \leq \frac{\mathcal{L}_k}{1 + \alpha_k} \quad \forall k \in \mathbb{N}, \tag{121}$$

where $\mathcal{L}_k = f(x_k) - f(x^*) + \frac{\gamma_k}{2}\|v_k - x^*\|^2$, *and both* (87) *and* (88) *hold true here.*

Proof As before, we calculate the difference

$$\begin{aligned} \mathcal{L}_{k+1} - \mathcal{L}_k =& f(x_{k+1}) - f(x_k) + \frac{\alpha_k}{2}(\mu - \gamma_{k+1}) \left\|v_{k+1} - x^*\right\|^2 \\ &+ \gamma_k \left(v_{k+1} - v_k, v_{k+1} - x^*\right) - \frac{\gamma_k}{2}\|v_{k+1} - v_k\|^2. \end{aligned}$$

Thanks to (120), we have

$$\begin{aligned} &\gamma_k \left(v_{k+1} - v_k, v_{k+1} - x^*\right) \\ &\quad \leq \mu\alpha_k \left(y_k - v_{k+1}, v_{k+1} - x^*\right) - \alpha_k \left\langle \nabla h(y_k) + p_{k+1}, v_{k+1} - x^*\right\rangle. \end{aligned} \tag{122}$$

where $p_{k+1} \in \partial g(v_{k+1})$.

By Lemma 1, the first term in (122) is split as follows

$$\begin{aligned} &2\mu\alpha_k \left(y_k - v_{k+1}, v_{k+1} - x^*\right) \\ &\quad = \mu\alpha_k \left(\left\|y_k - x^*\right\|^2 - \|y_k - v_{k+1}\|^2 - \left\|v_{k+1} - x^*\right\|^2\right). \end{aligned}$$

The gradient term in (122) is more subtle. Firstly, by convexity of g, we have

$$\begin{aligned} -\alpha_k \left\langle p_{k+1}, v_{k+1} - x^*\right\rangle &\leq -\alpha_k \left(g(v_{k+1}) - g(x^*)\right) \\ &= -\alpha_k \left(g(x_{k+1}) - g(x^*)\right) - \alpha_k \left(g(v_{k+1}) - g(x_{k+1})\right), \end{aligned}$$

and secondly, according to the update for y_k (see step 4 in Algorithm 4), we find

$$
\begin{aligned}
&- \alpha_k \left\langle \nabla h(y_k), v_{k+1} - x^* \right\rangle \\
&\quad = -\alpha_k \left\langle \nabla h(y_k), v_{k+1} - v_k \right\rangle - \alpha_k \left\langle \nabla h(y_k), v_k - x^* \right\rangle \\
&\quad = -\alpha_k \left\langle \nabla h(y_k), v_{k+1} - v_k \right\rangle - \left\langle \nabla h(y_k), y_k - x_k \right\rangle - \alpha_k \left\langle \nabla h(y_k), y_k - x^* \right\rangle .
\end{aligned}
$$

As h is μ-strongly convex on Q, by the fact $\{(x_k, y_k, v_k)\} \subset Q$, it follows that

$$
\begin{aligned}
&- \left\langle \nabla h(y_k), y_k - x_k \right\rangle - \alpha_k \left\langle \nabla h(y_k), y_k - x^* \right\rangle \\
&\quad \le h(x_k) - h(y_k) - \frac{\mu}{2} \left\| x_k - y_k \right\|^2 - \alpha_k \left(h(y_k) - h(x^*) \right) - \frac{\mu\alpha_k}{2} \left\| x^* - y_k \right\|^2 \\
&\quad = (1 + \alpha_k) \left(h(x_{k+1}) - h(y_k) \right) - \alpha_k \left(h(x_{k+1}) - h(x^*) \right) - \frac{\mu\alpha_k}{2} \left\| x^* - y_k \right\|^2 \\
&\qquad + h(x_k) - h(x_{k+1}) - \frac{\mu}{2} \left\| x_k - y_k \right\|^2 .
\end{aligned}
$$

Therefore, collecting all the estimates and dropping surplus negative terms related to $- \left\| x_k - y_k \right\|^2$ and $-\left\| y_k - v_{k+1} \right\|^2$, we get

$$
\begin{aligned}
&\mathcal{L}_{k+1} - \mathcal{L}_k \\
&\quad \le -\alpha_k \mathcal{L}_{k+1} + (1 + \alpha_k) \left(h(x_{k+1}) - h(y_k) \right) - \alpha_k \left\langle \nabla h(y_k), v_{k+1} - v_k \right\rangle \\
&\qquad - \frac{\gamma_k}{2} \left\| v_{k+1} - v_k \right\|^2 + g(x_{k+1}) - g(x_k) - \alpha_k \left(g(v_{k+1}) - g(x_{k+1}) \right) .
\end{aligned}
\tag{123}
$$

Let us consider the additional terms in (123). In view of (4), we have

$$
h(x_{k+1}) - h(y_k) \le \left\langle \nabla h(y_k), x_{k+1} - y_k \right\rangle + \frac{L}{2} \left\| x_{k+1} - y_k \right\|^2 .
$$

Thanks to the extrapolation step for x_{k+1} (see step 6 in Algorithm 4), we find a crucial relation

$$
x_{k+1} - y_k = \frac{\alpha_k}{1 + \alpha_k} (v_{k+1} - v_k),
$$

which gives that

$$
\begin{aligned}
&(1 + \alpha_k) \left(h(x_{k+1}) - h(y_k) \right) - \alpha_k \left\langle \nabla h(y_k), v_{k+1} - v_k \right\rangle - \frac{\gamma_k}{2} \left\| v_{k+1} - v_k \right\|^2 \\
&\quad \le \frac{L\alpha_k^2}{2(1 + \alpha_k)} \left\| v_{k+1} - v_k \right\|^2 - \frac{\gamma_k}{2} \left\| v_{k+1} - v_k \right\|^2 = 0,
\end{aligned}
$$

as $L\alpha_k^2 = \gamma_k (1 + \alpha_k)$. Moreover, since x_{k+1} is a convex combination of x_k and v_{k+1}, the estimate follows

$$
\begin{aligned}
&g(x_{k+1}) - g(x_k) - \alpha_k \left(g(v_{k+1}) - g(x_{k+1}) \right) \\
&\quad = (1 + \alpha_k) g(x_{k+1}) - g(x_k) - \alpha_k g(v_{k+1}) \le 0.
\end{aligned}
$$

Plugging this and the previous inequality into (123) gives

$$\mathcal{L}_{k+1} - \mathcal{L}_k \le -\alpha_k \mathcal{L}_{k+1},$$

which establishes (121).

By the relation $L\alpha_k^2 = \gamma_k(1+\alpha_k)$ and the contraction (121), it is clear that the two estimates (87) and (88) hold true. This completes the proof of this theorem. □

Acknowledgements The authors would like to thank the anonymous reviewers for valuable suggestions and careful comments, which significantly improved the qualify of an early version of the paper.

A Spectral analysis

Proof of Theorem 1 Let us start from the scalar case

$$R = \begin{pmatrix} -a & c \\ -b & -d \end{pmatrix},$$

where $a, b, c, d \ge 0$ and $\operatorname{tr} R < 0 < \det R$. Set

$$M = \begin{pmatrix} -a & 0 \\ -b & -d \end{pmatrix}, \quad N = \begin{pmatrix} 0 & c \\ 0 & 0 \end{pmatrix}.$$

By direct computations we have

$$E(\alpha, R) := (I - \alpha M)^{-1}(I + \alpha N) = \frac{1}{\delta}\begin{pmatrix} 1 + d\alpha & c\alpha(1 + d\alpha) \\ -b\alpha & 1 + a\alpha - bc\alpha^2 \end{pmatrix}, \tag{124}$$

where $\delta := (1 + a\alpha)(1 + d\alpha)$. Since $\operatorname{tr} R < 0$, we see that

$$0 < \det E(\alpha, R) = \frac{1}{\delta} = \frac{1}{1 + |\operatorname{tr} R|\,\alpha + ad\alpha^2} < 1.$$

Note that any eigenvalue θ of $E(\alpha, R)$ satisfies

$$\theta^2 - \operatorname{tr} E(\alpha, R)\theta + \det E(\alpha, R) = 0. \tag{125}$$

We now arrive at the following lemma, which says the spectrum of $E(\alpha, R)$ can be transformed to the circle $|\theta| = \sqrt{\det E(\alpha, R)} < 1$, with proper α. □

Lemma 5 *Assume*

$$R = \begin{pmatrix} -a & c \\ -b & -d \end{pmatrix},$$

with $a, b, c, d \geq 0$ *such that* $\operatorname{tr} R < 0 < \det R$. *Let* $E(\alpha, R)$ *be defined by (124). If* $\alpha > 0$ *satisfies*

$$|\operatorname{tr} R| - 2\sqrt{\det R} \leq bc\,\alpha \leq |\operatorname{tr} R| + 2\sqrt{\det R}, \tag{126}$$

then we have

$$\rho(E(\alpha, R)) = \frac{1}{\sqrt{1 + |\operatorname{tr} R|\,\alpha + ad\alpha^2}} < 1.$$

Proof If $\Delta = |\operatorname{tr} E(\alpha, R)|^2 - 4\det E(\alpha, R) \leq 0$, then any solution to (125) satisfies that $|\theta| = \sqrt{\det E(\alpha, R)}$ and the conclusion follows. By direct calculation, $\Delta \leq 0$ is equivalent to

$$\sqrt{\delta} - 1 \leq \alpha\sqrt{\det R} \leq \sqrt{\delta} + 1.$$

Square the inequality $\alpha\sqrt{\det R} - 1 \leq \sqrt{\delta}$ and cancel one α to get the upper bound in (126). The lower bound can be proved similarly. □

We now in the position of establishing Theorem 1. We first consider $G = G_{\text{HB}}$, for which we have

$$E(\alpha, G) = \frac{1}{1 + 2\alpha}\begin{pmatrix} (1 + 2\alpha)I & \alpha(1 + 2\alpha)I \\ -\alpha A/\mu & I - A\alpha^2/\mu \end{pmatrix}.$$

It is clear that $\theta \in \sigma(E(\alpha, G)) \Leftrightarrow \theta \in \sigma(E(\alpha, R(\lambda)))$, where $E(\alpha, R(\lambda))$ is defined by (124) with

$$R(\lambda) = \begin{pmatrix} 0 & 1 \\ -\lambda/\mu & -2 \end{pmatrix}, \quad \lambda \in \sigma(A).$$

As $|\operatorname{tr} R(\lambda)| \leq 2\sqrt{\det R(\lambda)}$, by Lemma 5, if

$$0 < \alpha \leq 2/\sqrt{\kappa(A)}, \tag{127}$$

then we can obtain

$$\rho(E(\alpha, G)) = \max_{\lambda \in \sigma(A)} \rho(E(\alpha, R(\lambda))) = \frac{1}{\sqrt{1 + 2\alpha}}.$$

Similarly, for $G = G_{\mathrm{NAG}}$ with condition (127), we can establish

$$\rho(E(\alpha, G)) = \max_{\lambda\in\sigma(A)} \rho(E(\alpha, R(\lambda))) = \frac{1}{\sqrt{1+2\alpha+\alpha^2}} \le \frac{1}{\sqrt{1+2\alpha}}.$$

Consequently, for both two cases, taking $\alpha = 2/\sqrt{\kappa(A)}$ yields the spectrum bound

$$\rho(E(\alpha, G)) \le \frac{1}{\sqrt{1+4/\sqrt{\kappa(A)}}} \le \frac{1}{1+1/\sqrt{\kappa(A)}}.$$

This concludes the proof of Theorem 1. □

Proof of Theorem 2 Observe that $\widetilde{E}_k$ is similar with

$$\begin{pmatrix} I & O \\ O & \gamma_k I \end{pmatrix}^{-1} \begin{pmatrix} I & O \\ O & \gamma_{k+1} I \end{pmatrix} E(\alpha_k, G(\gamma_{k+1})) = \frac{H_k}{1+\alpha_k},$$

where

$$H_k = \begin{pmatrix} I & \alpha_k I \\ -A\alpha_k/\gamma_k & I - A\alpha_k^2/\gamma_k \end{pmatrix}.$$

To prove (51), it is sufficient to verify $\rho(H_k) = 1$.

Given any eigenvalue $\theta \in \sigma(H_k)$, it solves

$$\theta^2 + (\lambda\alpha_k^2/\gamma_k - 2)\theta + 1 = 0,$$

with some $\lambda \in \sigma(A) \subset [0, L]$. By (49), $\{\gamma_k\}$ is decreasing and thus $\gamma_k \le \gamma_0 = L$. According to our choice $L\alpha_k^2 = \gamma_k(1+\alpha_k)$, we have $0 < \alpha_k \le 2$ and moreover $0 < \lambda\alpha_k^2/\gamma_k \le L\alpha_k^2/\gamma_k = 1 + \alpha_k \le 3$. This implies $\Delta = (\lambda\alpha_k^2/\gamma_k - 2)^2 - 4 \le 0$ for all $\lambda \in \sigma(A)$. Therefore, we conclude that $|\theta| = 1$ for all $\theta \in \sigma(H_k)$, which proves $\rho(H_k) = 1$ and thus establishes (51).

Thanks to Lemma B2, there holds

$$\frac{1}{(k+1)^2} \le \frac{\gamma_k}{\gamma_0} = \prod_{i=0}^{k-1} \frac{1}{1+\alpha_i} \le \frac{4}{(k+2)^2}.$$

This proves (52) and completes the proof of Theorem 2. □

B Decay rates

Lemma B1 *Let $\gamma_0 > 0$ and $\mu \ge 0$ be given and assume there is a real positive sequence $\{L_k\}$ such that $L_k \ge \mu$. Define $\{(\alpha_k, \gamma_k)\}$ by that*

$$\begin{cases} L_k\alpha_k^2 = \gamma_{k+1}, \quad \alpha_k > 0, \\ \gamma_{k+1} = (1-\alpha_k)\gamma_k + \mu\alpha_k. \end{cases} \tag{128}$$

Then we have $\gamma_k > 0, 0 < \alpha_k \le 1$ *and* $\alpha_k \ge \sqrt{\min\{\gamma_1, \mu\}/L}$, *where* $L := \sup_{k\in\mathbb{N}} L_k$. *Moreover, for all* $k \ge 1$,

$$\prod_{i=0}^{k-1}(1-\alpha_i) \le \min\left\{4\left(2+\sum_{i=0}^{k-1}\sqrt{\frac{\gamma_0}{L_i}}\right)^{-2}, \left(1-\sqrt{\frac{\min\{\gamma_1,\mu\}}{L}}\right)^{k}\right\}, \tag{129}$$

and if $\mu = 0$, *then we have the lower bound*

$$\prod_{i=0}^{k-1}(1-\alpha_i) \ge \left(1+\sum_{i=0}^{k-1}\sqrt{\frac{\gamma_0}{L_i}}\right)^{-2}. \tag{130}$$

Proof Let us first check that $0 < \alpha_k \le 1$ and $\gamma_k > 0$. Since $\gamma_0 > 0$, by (128) we have

$$L_0\alpha_0^2 = \gamma_1 = (1-\alpha_0)\gamma_0 + \mu\alpha_0,$$

from which we claim that $0 < \alpha_0 \le 1$. Thus by the second step in (128) we have $\gamma_1 > 0$. A sequential argument implies that $0 < \alpha_k \le 1$ and $\gamma_k > 0$ for all $k \ge 0$.

It is not hard to find the fact: if $\gamma_0 > \mu$, then $\mu < \gamma_{k+1} < \gamma_k$ and if $\gamma_0 < \mu$, then $\gamma_k < \gamma_{k+1} < \mu$. Particularly, if $\gamma_0 = \mu$, then $\gamma_k = \mu$. Based on this observation and the fact $L_k \le L$, we conclude that $\alpha_k \ge \sqrt{\min\{\gamma_1, \mu\}/L}$ and thus

$$\prod_{i=0}^{k-1}(1-\alpha_i) \le \left(1-\sqrt{\frac{\min\{\gamma_1,\mu\}}{L}}\right)^{k}.$$

Next, let us prove the estimate

$$\rho_k \le 4\left(2+\sum_{i=0}^{k-1}\sqrt{\frac{\gamma_0}{L_i}}\right)^{-2}, \tag{131}$$

where ρ_k is defined by (101). We start from the trivial equality

$$\frac{1}{\sqrt{\rho_{k+1}}} - \frac{1}{\sqrt{\rho_k}} = \frac{\sqrt{\rho_k}-\sqrt{\rho_{k+1}}}{\sqrt{\rho_k\rho_{k+1}}} = \frac{1-\sqrt{1-\alpha_k}}{\sqrt{\rho_{k+1}}} = \frac{\alpha_k}{\sqrt{\rho_{k+1}}(1+\sqrt{1-\alpha_k})}, \tag{132}$$

where we used the relation $\rho_{k+1} = \rho_k(1-\alpha_k)$. By (128), for any $i \ge 0$, it holds that

$$\gamma_{i+1} = (1-\alpha_i)\gamma_i + \mu\alpha_i \ge (1-\alpha_i)\gamma_i, \tag{133}$$

and multiplying the above inequality from $i = 0$ to $i = k-1$ gives $\rho_k \le \gamma_k/\gamma_0$. Plugging this into (132) and using the relation $L_k\alpha_k^2 = \gamma_{k+1}$ and the fact $0 < \alpha_k \le 1$ imply

$$\frac{1}{\sqrt{\rho_{k+1}}} - \frac{1}{\sqrt{\rho_k}} \ge \frac{\sqrt{\gamma_0}\alpha_k}{\sqrt{\gamma_{k+1}}(1+\sqrt{1-\alpha_k})} \ge \frac{\sqrt{\gamma_0}}{2\sqrt{L_k}},$$

which further indicates that

$$\frac{1}{\sqrt{\rho_k}} - \frac{1}{\sqrt{\rho_0}} \ge \sum_{i=0}^{k-1} \frac{\sqrt{\gamma_0}}{2\sqrt{L_i}}.$$

Therefore, a simple calculation proves (131) and concludes the proof of this lemma.

For $\mu = 0$, we have the relation $\rho_k = \gamma_k/\gamma_0$, and proceeding as the above derivation, it is not hard to establish the lower bound (130). This concludes the proof of this lemma. □

Similarly, we can establish the following result, the proof of which is omitted for simplicity.

Lemma B2 *Let $\gamma_0 > 0$ and $\mu \ge 0$ be given and assume there is a real positive sequence $\{L_k\}$ such that $L_k \ge \mu$. Define $\{(\alpha_k, \gamma_k)\}$ by that*

$$\begin{cases} \gamma_{k+1} = \gamma_k + \alpha_k(\mu - \gamma_{k+1}), \\ L_k\alpha_k^2 = \gamma_k(1+\alpha_k), \ \alpha_k > 0. \end{cases}$$

Then we have $\gamma_k > 0$ and $\alpha_k \ge \sqrt{\min\{\gamma_0, \mu\}/L}$, where $L := \sup_{k\in\mathbb{N}} L_k$. Moreover, for all $k \ge 1$,

$$\prod_{i=0}^{k-1} \frac{1}{1+\alpha_i} \le \min\left\{ 4\left(2 + \sum_{i=0}^{k-1}\sqrt{\frac{\gamma_0}{L_i}}\right)^{-2}, \left(1+\sqrt{\frac{\min\{\gamma_0,\mu\}}{L}}\right)^{-k}\right\},$$

and if $\mu = 0$, then we have the lower bound

$$\prod_{i=0}^{k-1} \frac{1}{1+\alpha_i} \ge \left(1 + \sum_{i=0}^{k-1}\sqrt{\frac{\gamma_0}{L_i}}\right)^{-2}.$$

References

1. Alvarez, F.: On the minimizing property of a second order dissipative system in Hilbert spaces. SIAM J. Control. Optim. **38**(4), 1102–1119 (2000)
2. Apidopoulos, V., Aujol, J.-F., Dossal, C.: Convergence rate of inertial Forward–Backward algorithm beyond Nesterov's rule. Math. Program. (2018)

3. Attouch, H., Goudou, X., Redont, P.: The heavy ball with friction method, I. The continuous dynamical system: Global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. Commun. Contemp. Math. **2**(1), 1–34 (2000)
4. Attouch, H., Buttazzo, G., Michaille, G.: Variational Analysis in Sobolev and BV Spaces. Society for Industrial and Applied Mathematics, MOS-SIAM Series on Optimization (2014)
5. Attouch, H., Chbani, Z.: Fast inertial dynamics and FISTA algorithms in convex optimization. Perturbation aspects. arXiv:1507.01367 (2015)
6. Attouch, H., Chbani, Z., Peypouquet, J., Redont, P.: Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. Math. Program. **168**(1–2), 123–175 (2016)
7. Attouch, H., Cabot, A.: Convergence rates of inertial forward-backward algorithms. SIAM J. Optim. **28**(1), 849–874 (2018)
8. Attouch, H., Chbani, Z., Riahi, H.: Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. ESAIM: Control Optim. Calc. Variat. **25**(2), 1 (2019)
9. Ahmad, S., Ambrosetti, A.: A Textbook on Ordinary Differential Equations, 2nd. UNITEXT - La Matematica per il 3+2, vol. 88. Springer, Cham (2015)
10. Aujol, J., Dossal, C.: Optimal rate of convergence of an ODE associated to the fast gradient descent schemes for $b>0$. hal-01547251v2:22, (2017)
11. Balti, M., May, R.: Asymptotic for the perturbed heavy ball system with vanishing damping term. Evol. Equ. Control Theory **6**(2),1 (2016)
12. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**(1), 183–202 (2009)
13. Bellman, R.: Stability Theory of Differential Equations. MeGraw-Hill Book Company (1953)
14. Cabot, A., Engler, H., Gadat, S.: On the long time behavior of second order differential equations with asymptotically small dissipation. Trans. Am. Math. Soc. **361**(11), 5983–6017 (2009)
15. Cabot, A., Paoli, L.: Asymptotics for some vibro-impact problems with a linear dissipation term. J. Math. Pures Appl. **87**(3), 291–323 (2007)
16. Demmel, J.: Applied Numerical Linear Algebra. Society for Industrial and Applied Mathematics (1997)
17. Diakonikolas, J., Orecchia, L.: The approximate duality gap technique: A unified theory of first-order methods. arXiv:1712.02485, (2018)
18. Ghadimi, E., Feyzmahdavian, H. R., Johansson, M.: Global convergence of the Heavy-ball method for convex optimization. In European Control Conference (ECC), pp. 310–315, (2015)
19. Goudou, X., Munier, J.: The gradient and heavy ball with friction dynamical systems: the quasiconvex case. Math. Program. **116**(1–2), 173–191 (2009)
20. Güler, O.: New proximal point algorithms for convex minimization. SIAM J. Optim. **2**(4), 649–664 (1992)
21. Kreuter, M.: Sobolev Spaces of Vector-Valued Functions. Master Thesis, Ulm University (2015)
22. Lessard, L., Recht, B., Packard, A.: Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints. SIAM J. Optim. **26**(1), 57–95 (2016)
23. LeVeque, R.: Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems. Society for Industrial and Applied Mathematics (2007)
24. Lin, Z., Li, H., Fang, C.: Accelerated Optimization for Machine Learning. Springer, Singapore (2020)
25. Luo, H.: Accelerated differential inclusion for convex optimization. arXiv:2103.06629 (2021)
26. Nguyen, N., Fernandez, P., Freund, R.M., Peraire, J.: Accelerated residual methods for the iterative solution of systems of equations. SIAM J. Sci. Comput. **40**(5), A3157–A3179 (2018)
27. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Sov. Math. Doklady **27**(2), 372–376 (1983)
28. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2012)
29. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, volume 87. Springer (2013)
30. Necoara, I., Nesterov, Y., Glineur, F.: Linear convergence of first order methods for non-strongly convex optimization. Math. Program. **175**(1), 69–107 (2019)
31. O'Donoghue, B., Candès, E.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**(3), 715–732 (2015)
32. Paoli, L.: An existence result for vibrations with unilateral constraints: case of a nonsmooth set of constraints. Math. Models Methods Appl. Sci. **10**(06), 815–831 (2000)
33. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. **1**(3), 127–239 (2014)

34. Polyak, B.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
35. Rockafellar, R.: Convex Analysis. Princeton University Press (1970)
36. Schatzman, M.: A class of nonlinear differential equations of second order in time. Nonlinear Anal. **2**(3), 355–373 (1978)
37. Su, W., Boyd, S., Candès, E.: A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. J. Mach. Learn. Res. **17**(153), 1–43 (2016)
38. Siegel, J.: Accelerated first-order methods: Differential equations and Lyapunov functions. arXiv preprint: arXiv:1903.05671 (2019)
39. Süli, E.: Numerical Solution of Ordinary Differential Equations. University of Oxford, Mathematical Institute (2010)
40. Sun, T., Yin, P., Li, D., Huang, C., Guan, L., Jiang, H.: Non-ergodic convergence analysis of heavy-ball algorithms. arXiv:1811.01777 (2018)
41. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. Unpublished manuscript (2008)
42. Vassilis, A., Jean-François, A., Charles, D.: The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case $b \leq 3$. SIAM J. Optim. **28**(1), 551–574 (2018)
43. Wibisono, A., Wilson, A., Jordan, M.: A variational perspective on accelerated methods in optimization. Proc. Natl. Acad. Sci. **113**(47), E7351–E7358 (2016)
44. Wilson, A., Recht, B., Jordan, M.: A Lyapunov analysis of momentum methods in optimization. arXiv preprint: arXiv:1611.02635 (2016)
45. Yousef, S.: Iterative Methods for Sparse Linear Systems. Society for Industrial and Applied Mathematics, USA (2003)