



Properties and practicability of convergence-guaranteed optimization methods derived from weak discrete gradients

Kansei Ushiyama¹ · Shun Sato¹ · Takayasu Matsuo¹

Received: 31 August 2023 / Accepted: 20 February 2024
© The Author(s) 2024

Abstract

The ordinary differential equation (ODE) models of optimization methods allow for concise proofs of convergence rates through discussions based on Lyapunov functions. The weak discrete gradient (wDG) framework discretizes ODEs while preserving the properties of convergence, serving as a foundation for deriving optimization methods. Although various optimization methods have been derived through wDG, their properties and practical applicability remain underexplored. Hence, this study elucidates these aspects through numerical experiments. Particularly, although wDG yields several implicit methods, we highlight the potential utility of these methods in scenarios where the objective function incorporates a regularization term.

Keywords Convex optimization · Proximal gradient method · Numerical analysis · Discrete gradient

1 Introduction

In this paper, we consider the unconstrained convex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x).$$

✉ Kansei Ushiyama
ushiyama-kansei074@g.ecc.u-tokyo.ac.jp
Shun Sato
shun@mist.i.u-tokyo.ac.jp
Takayasu Matsuo
matsuo@mist.i.u-tokyo.ac.jp

¹ Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku 113-8656, Tokyo, Japan

One approach to analyze and derive optimization methods for this problem is modeling them by ordinary differential equations (ODEs). This approach has been considered for a long time, starting with the modeling of the steepest descent method by gradient flow. Further, it has received more attention in recent years since the derivation of an ODE model of Nesterov’s accelerated gradient method (NAG) by Su–Boyd–Candès [1].

One of the advantages of considering ODE models is that we can analyze convergence rates of, for example, the objective function succinctly in continuous time via Lyapunov functions which express the rate explicitly. This framework of Lyapunov analysis has been organized by [2, 3], and applied to various ODEs (e.g., [4–6].)

The weak discrete gradient (wDG) is a concept proposed in [7] to transfer the Lyapunov analysis in continuous-time ODEs to discrete-time optimization methods. It represents a class of gradient evaluation methods that satisfy conditions necessary to reproduce the continuous proof in discrete systems. By discretizing ODEs using wDGs, we have an abstract optimization method that contains a variety of optimization methods. Additionally, selecting a wDG yields a concrete optimization method with a convergence rate guarantee. However, [7] did not discuss which specific wDG is suitable for optimization in practice, and only limited preliminary numerical examples are given in the appendix section.

Therefore, in this paper, we examine the properties and performance of the methods obtained by discretizing the following accelerated gradient flows with wDGs more intensively numerically:

$$\dot{x} = \frac{2}{t}(v - x), \quad \dot{v} = -\frac{t}{2}\nabla f(x) \quad (\text{for convex } f), \tag{1}$$

$$\dot{x} = \sqrt{\mu}(v - x), \quad \dot{v} = \sqrt{\mu}(x - v - \nabla f(x)/\mu) \quad (\text{for } \mu\text{-strongly convex } f). \tag{2}$$

A differentiable function f is said to be μ -strongly convex if it satisfies

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

and f is said to be L -smooth if it satisfies

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

The above ODEs are models for Nesterov’s famous accelerated gradient method for convex functions [8] and strongly convex functions [9].

In this paper, we discuss the following three issues based on numerical experiments:

- 1. Differences between wDG discretizations and other discretizations:** As wDG schemes for the accelerated gradient flow are different from NAG, we aim to compare them. Additionally, since the wDG schemes do not coincide with any typical numerical methods for ODEs, we compare them with the partitioned Runge–Kutta method, which is a natural choice for the ODE.

2. **Practicability of implicit methods:** Most wDG discretizations yield implicit schemes. In optimization, implicit methods are not favored because of their high computational complexity per iteration. In numerical analysis, however, it is common to employ implicit methods to solve “stiff” equations due to their high stability. In contrast to explicit methods, which can only compute numerical solutions with a small time step width due to their lower stability, the use of implicit methods allows for a much larger time step width. Consequently, despite the higher computational cost per step, the overall cost is reduced, enabling the attainment of numerical solutions at more advanced time points. We point out that a similar phenomenon occurs in optimization as well. Specifically, for “stiff” optimization problems, implicit methods can converge more rapidly in terms of overall computational cost compared to explicit methods.
3. **Efficiency of variants of the proximal gradient method derived by wDG discretizations:** We can create a new wDG by summing two wDGs. This property allows us to produce variants of the proximal gradient method with convergence rate guarantees. We verify their performance through numerical experiments.

Issues 1 and 2 will be discussed in Sect. 3 and Issue 3 in Sect. 4.

Before proceeding, let us summarize the insight in [7, Appendix I] here, in which simple numerical illustrations were given for both convex and strongly convex cases. For the convex case, the paper considered NAG for convex functions, and an explicit wDG scheme for convex functions, which is a variant of (5) below (this is almost the only explicit scheme we can naturally think of out of the wDG framework for convex functions). They were tested for a simple quartic function and the following was obtained: the performance was roughly the same for both methods, while the performance of the wDG scheme was slightly better asymptotically. For strongly convex case, we can think of various explicit wDG schemes. There, two explicit wDG schemes (which are called wDGEX_sc and wDGEX2_sc in the present paper; see below) and NAG for strongly convex functions were tested for a simple quadratic function. The numerical experiment suggested the following: again, the performance of these three methods was quite similar, while trajectories in the solution space were considerably different. In both convex and strongly convex cases, the actual convergence rates were much better than the theoretical estimates, which is not surprising since the theoretical estimate is the *estimate for the worst-case functions*.

In the numerical experiments carried out in the present study, we have the following settings. First, we mainly focus on strongly convex cases, since in the convex case, the only explicit wDG scheme was already tested in [7] as noted above, and few are left for investigation (an exception will be given in the last numerical example.) Second, we observe the convergence in the gradient norm, i.e., $\|\nabla f(x)\|_2$, instead of $f(x)$ itself, although the wDG framework provides convergence rate guarantees with respect to the latter. This is because we also deal with objective functions whose exact minimums are unknown. The profiles remain informative in view of the inequality $\|\nabla f(x)\|_2^2 \leq 2L(f(x) - f^*)$, which holds for any L -smooth convex functions. Finally, we assume the existence of the optimal solution x^* and the optimal value $f^* = f(x^*)$.

2 Preliminary: weak discrete gradient [7]

In this section, we briefly review the wDG.

In the Lyapunov analysis of ODEs, we derive convergence rates by showing the nonincrease of rate-revealing Lyapunov functions (by showing $(d/dt)E(t) \leq 0$). For accelerated gradient flows (1) and (2), the following functions serve as Lyapunov functions:

$$E(t) = t^2(f(x) - f^*) + \|v - x^*\|^2 \quad \text{for (1), and} \tag{3}$$

$$E(t) = e^{\sqrt{\mu}t} \left(f(x) - f^* + \frac{\mu}{2} \|v - x^*\|^2 \right) \quad \text{for (2).} \tag{4}$$

Since these E 's are non-increasing and the terms involving $\|v - x^*\|^2$ are nonnegative, we can derive immediately the convergence rate $f(x(t)) - f^* = O(1/t^2)$ and $O(e^{-\sqrt{\mu}t})$.

wDGs represent a class of gradient approximations that, given an optimization ODE and its Lyapunov function for the rate analysis, provides a discretization of the ODE that preserves its convergence rate.

Definition 1 (Weak discrete gradient) A gradient approximation $\bar{\nabla} f : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is said to be *weak discrete gradient of f* if there exists $\alpha \geq 0$ and β, γ with $\beta + \gamma \geq 0$ such that the following two conditions hold for all $x, y, z \in \mathbb{R}^d$:

$$f(y) - f(x) \leq \langle \bar{\nabla} f(y, z), y - x \rangle + \alpha \|y - z\|^2 - \beta \|z - x\|^2 - \gamma \|y - x\|^2, \\ \bar{\nabla} f(x, x) = \nabla f(x).$$

The following functions are wDGs.

Proposition 1 Suppose that $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a μ -strongly convex function. Let (L) and (SC) denote the additional assumptions: (L) f is L -smooth, and (SC) $\mu > 0$. Then, the following functions are wDGs:

1. If $\bar{\nabla} f(y, x) = \nabla f(x)$ and f satisfies (L), then $(\alpha, \beta, \gamma) = (L/2, \mu/2, 0)$.
2. If $\bar{\nabla} f(y, x) = \nabla f(y)$, then $(\alpha, \beta, \gamma) = (0, 0, \mu/2)$.
3. If $\bar{\nabla} f(y, x) = \nabla f(\frac{x+y}{2})$ and f satisfies (L), then $(\alpha, \beta, \gamma) = ((L + \mu)/8, \mu/4, \mu/4)$.
4. If $\bar{\nabla} f(y, x) = \nabla_{AVF} f(y, x)$ and f satisfies (L), then $(\alpha, \beta, \gamma) = (L/6 + \mu/12, \mu/4, \mu/4)$.
5. If $\bar{\nabla} f(y, x) = \nabla_G f(y, x)$ and f satisfies (L)(SC), then $(\alpha, \beta, \gamma) = ((L + \mu)/8 + (L - \mu)^2/16\mu, \mu/4, 0)$.
6. If $\bar{\nabla} f(y, x) = \nabla_{IA} f(y, x)$ and f satisfies (L)(SC), then $(\alpha, \beta, \gamma) = (dL^2/\mu - \mu/4, \mu/2, -\mu/4)$.

Here, Gonzalez discrete gradient [10] is as follows:

$$\nabla_G f(y, x) := \nabla f\left(\frac{y+x}{2}\right) + \frac{f(y) - f(x) - \langle \nabla f\left(\frac{y+x}{2}\right), y-x \rangle}{\|y-x\|^2} (y-x).$$

Itoh–Abe discrete gradient [11] is as follows:

$$\nabla_{IA} f(y, x) := \begin{bmatrix} \frac{f(y_1, x_2, x_3, \dots, x_d) - f(x_1, x_2, x_3, \dots, x_d)}{y_1 - x_1} \\ \frac{f(y_1, y_2, x_3, \dots, x_d) - f(y_1, x_2, x_3, \dots, x_d)}{y_2 - x_2} \\ \vdots \\ \frac{f(y_1, y_2, y_3, \dots, y_d) - f(y_1, y_2, y_3, \dots, x_d)}{y_d - x_d} \end{bmatrix},$$

and average vector field (AVF) [12] is as follows:

$$\nabla_{AVF} f(y, x) := \int_0^1 \nabla f(\tau y + (1 - \tau)x) d\tau.$$

Given an ODE and its rate-revealing Lyapunov function, we can derive a wDG scheme that ensures a decrease in the Lyapunov function. Here are the convergence rate theorems of wDG discretizations of ODE (1) and (2) with respect to Lyapunov functions (3) and (4). In the following schemes, δ^+ denotes the forward difference operator with time step h , e.g., $\delta^+ x^{(k)} := (x^{(k+1)} - x^{(k)})/h$.

Proposition 2 (Convex case, Theorem 5.4 in [7]) *Let $\bar{\nabla} f$ be a wDG of f and suppose that $\beta \geq 0, \gamma \geq 0$. Let f be a convex function that additionally satisfies the necessary conditions that the wDG requires. Let $\{(x^{(k)}, v^{(k)})\}$ be the sequence given by*

$$\begin{cases} \delta^+ x^{(k)} = \frac{\delta^+ A_k}{A_k} (v^{(k+1)} - x^{(k+1)}), \\ \delta^+ v^{(k)} = -\frac{\delta^+ A_k}{4} \bar{\nabla} f(x^{(k+1)}, z^{(k)}), \\ \frac{z^{(k)} - x^{(k)}}{h} = \frac{\delta^+ A_k}{A_{k+1}} (v^{(k)} - x^{(k)}) \end{cases} \tag{5}$$

with $(x^{(0)}, v^{(0)}) = (x_0, v_0)$, where $A_k := A(kh)$. Then, if $A_k = (kh)^2$ and $h \leq 1/\sqrt{2\alpha}$, the sequence satisfies

$$f(x^{(k)}) - f^* \leq \frac{2\|x_0 - x^*\|^2}{A_k}.$$

Proposition 3 (Strongly convex case, Theorem 5.5 in [7]) *Let $\bar{\nabla} f$ be a wDG of f and suppose that $\beta + \gamma > 0$. Let f be a strongly convex function that additionally satisfies*

the necessary conditions that the wDG requires. Let $\{(x^{(k)}, v^{(k)})\}$ be the sequence given by

$$\begin{cases} \delta^+ x^{(k)} = \sqrt{2(\beta + \gamma)}(v^{(k+1)} - x^{(k+1)}), \\ \delta^+ v^{(k)} = \sqrt{2(\beta + \gamma)}\left(\frac{\beta}{\beta + \gamma}z^{(k)} + \frac{\gamma}{\beta + \gamma}x^{(k+1)} - v^{(k+1)} - \frac{\bar{\nabla}f(x^{(k+1)}, z^{(k)})}{2(\beta + \gamma)}\right), \\ \frac{z^{(k)} - x^{(k)}}{h} = \sqrt{2(\beta + \gamma)}(x^{(k)} + v^{(k)} - 2z^{(k)}) \end{cases} \quad (6)$$

with $(x^{(0)}, v^{(0)}) = (x_0, v_0)$. Then, if $h \leq \bar{h} := \left(\sqrt{2}(\sqrt{\alpha + \gamma} - \sqrt{\beta + \gamma})\right)^{-1}$, the sequence satisfies

$$f(x^{(k)}) - f^* \leq \left(1 + \sqrt{2(\beta + \gamma)}h\right)^{-k} \left(f(x_0) - f^* + \beta\|v_0 - x^*\|^2\right).$$

In particular, the sequence satisfies

$$f(x^{(k)}) - f^* \leq \left(1 - \sqrt{\frac{\beta + \gamma}{\alpha + \gamma}}\right)^k \left(f(x_0) - f^* + \beta\|v_0 - x^*\|^2\right),$$

when the optimal step size $h = \bar{h}$ is employed.

3 Features of wDG schemes

3.1 Target methods

In this section, we discuss the features of the wDG method by numerically comparing them to existing optimization methods and numerical methods. We consider wDG methods derived from the accelerated gradient flow for strongly convex functions (2).

- wDGEX_sc: An explicit method obtained by the following instantiation: we first replace $\bar{\nabla}f(x^{(k+1)}, z^{(k)})$ with $\bar{\nabla}f(x^{(k+1)}, x^{(k)})$ in the abstract scheme (6), which gives a convergent wDG scheme with a worse convergence rate [7]. However, this scheme is more natural as a numerical method. We then set $\bar{\nabla}f(y, x) = \nabla f(x)$ (Item 1 of Proposition 1) to find

$$\begin{cases} \delta^+ x^{(k)} = \sqrt{2(\beta + \gamma)}(v^{(k+1)} - x^{(k+1)}), \\ \delta^+ v^{(k)} = \sqrt{2(\beta + \gamma)}\left(\frac{\beta}{\beta + \gamma}x^{(k)} + \frac{\gamma}{\beta + \gamma}x^{(k+1)} - v^{(k+1)} - \frac{\nabla f(x^{(k)})}{2(\beta + \gamma)}\right). \end{cases}$$

For L -smooth functions, the convergence rate is $f(x^{(k)}) - f^* \leq O\left((1 - \mu h)^k\right)$ if $h \leq 1/L$.

- wDGEX2_sc: An explicit method obtained by using $\bar{\nabla} f(y, x) = \nabla f(x)$ in the abstract scheme (6). The convergence rate and the step-size restriction are shown in the proposition.
- wDGagr_sc: An explicit method obtained by considering another Lyapunov function

$$E = f(x) - f^* + \frac{4}{9}\mu \left(\|v - x^*\|^2 - \frac{1}{2}\|v - x\|^2 \right).$$

This function is proposed in [13] and shown to decrease faster than (4); however this does not guarantee faster convergence of f since $\frac{4}{9}\mu(\|v - x^*\|^2 - \frac{1}{2}\|v - x\|^2)$ is not always positive. We can derive an abstract wDG method for the ODE that ensure the corresponding non-increase of E :

$$\left\{ \begin{array}{l} \delta^+ x^{(k)} = \sqrt{2(\beta + \gamma)}(v^{(k+1)} - x^{(k+1)}), \\ \delta^+ v^{(k)} = \sqrt{2(\beta + \gamma)} \left(\frac{2\beta}{\beta + \gamma} z^{(k)} + \frac{2\gamma}{\beta + \gamma} x^{(k+1)} - 2v^{(k+1)} - \frac{9\bar{\nabla} f(x^{(k+1)}, z^{(k)})}{8(\beta + \gamma)} \right), \\ \frac{z^{(k)} - x^{(k)}}{h} = \sqrt{2(\beta + \gamma)}(x^{(k)} + v^{(k)} - 2z^{(k)}). \end{array} \right.$$

We omit the proof of the non-increase property. We set $\bar{\nabla} f(y, x) = \nabla f(x)$ to obtain an explicit method. Owing to the aforementioned reason, the convergence rate of the method is not known. It is interesting to numerically observe how the fast decrease in the above E affects the actual performance.

- wDGIE_sc: An implicit method obtained by letting $\bar{\nabla} f(y, x) = \nabla f(y)$ (Item 2 of Proposition 1) in (6).
- wDGAVF_sc: An implicit method obtained by letting $\bar{\nabla} f(y, x) = \nabla_{AVF} f(y, x)$ (Item 4 of Proposition 1) in (6). Note that if f is quadratic, this method is the same as the implicit midpoint rule $\bar{\nabla} f(y, x) = \nabla f((y + x)/2)$.
- wDGIA_sc: An implicit method obtained by letting $\bar{\nabla} f(y, x) = \nabla_{IA} f(y, x)$ (Item 6 of Proposition 1) in (6).

We compare these methods to Nesterov’s accelerated gradient method for μ -strongly convex functions whose continuous limit is also (2):

- NAG_sc:

$$\left\{ \begin{array}{l} x^{(k+1)} = y^{(k)} - s \nabla f(y^{(k)}), \\ y^{(k+1)} = x^{(k+1)} + \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}}(x^{(k+1)} - x^{(k)}), \end{array} \right.$$

where s is a step size. For L -smooth functions, its convergence rate is $f(x^{(k)}) - f^* \leq O\left((1 - \sqrt{\mu s})^k\right)$ if $s \leq 1/L$.

wDGEX2_sc and NAG_sc have the same best convergence rate among those explicit methods when $h = 1/(\sqrt{L} - \sqrt{\mu})$ and $s = 1/L$ respectively.

We also compare wDG methods to a partitioned Runge–Kutta method and its modified version.

- pRK: An explicit partitioned Runge–Kutta method (cf. [14]) for (2)

$$\begin{cases} z^{(k)} - x^{(k)} = \frac{\sqrt{\mu}h}{2}(v^{(k)} - z^{(k)}), \\ v^{(k+1)} - v^{(k)} = \sqrt{\mu}h \left(z^{(k)} - v^{(k+1)} - \frac{\nabla f(z^{(k)})}{\mu} \right), \\ x^{(k+1)} - z^{(k)} = \frac{\sqrt{\mu}h}{2}(v^{(k+1)} - z^{(k)}). \end{cases} \tag{7}$$

The convergence rate is $O\left(\left(1 + \sqrt{\mu}h\right)^{-k}\right)$ for $h \leq 4\sqrt{\mu}/L$ (see Appendix A.1),

- pRK2: A modified version of (7)

$$\begin{cases} z^{(k)} - x^{(k)} = \left(1 - \frac{\sqrt{\mu}h}{2}\right)\sqrt{\mu}h(v^{(k)} - z^{(k)}), \\ v^{(k+1)} - v^{(k)} = \sqrt{\mu}h \left(z^{(k)} - \frac{1}{2}v^{(k)} - \frac{1}{2}v^{(k+1)} - \frac{\nabla f(z^{(k)})}{\mu} \right), \\ x^{(k+1)} - z^{(k)} = \frac{\mu h^2}{2}(v^{(k+1)} - z^{(k)}) + \left(1 - \frac{\sqrt{\mu}h}{2}\right)\sqrt{\mu}h^2 \frac{v^{(k+1)} - v^{(k)}}{h}. \end{cases} \tag{8}$$

The convergence rate is $O\left(\left(\frac{1 - \sqrt{\mu}h/2}{1 + \sqrt{\mu}h/2}\right)^k\right)$ for $h \leq \sqrt{1/L}$ (see Appendix A.2).

3.2 Numerical experiments

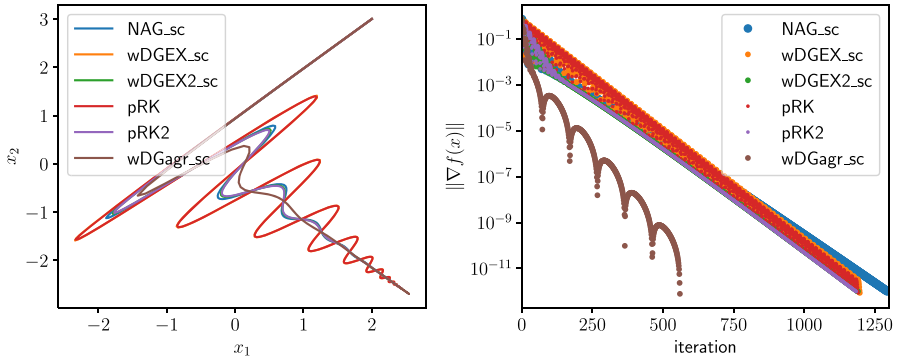
We compare the above methods in terms of trajectories and function decrease. The test problem is the 2-dimensional quadratic function minimization $\min_{x \in \mathbb{R}^2} f(x)$ where

$$f(x) = \frac{1}{2}x^\top Ax + b^\top x, \quad A = \begin{pmatrix} 0.101 & 0.099 \\ 0.099 & 0.101 \end{pmatrix}, \quad b = \begin{pmatrix} 0.01 \\ 0.02 \end{pmatrix}. \tag{9}$$

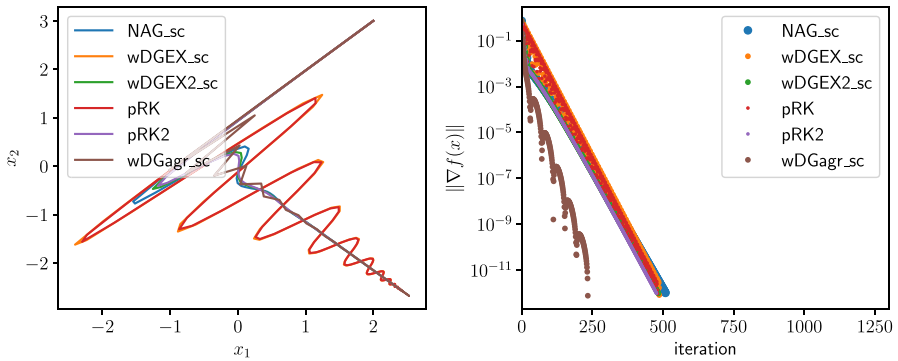
This function is 0.002-strongly convex and 0.2-smooth.

First, we compare explicit methods. Figure 1 shows the results when the baseline step size is set to $s_0 = 1/L$ for NAG_sc and $h_0 = 1/(\sqrt{L} - \sqrt{\mu})$ for the other methods, which are theoretical maximum step sizes for NAG_sc and wDG2_sc respectively. Each method is run with step sizes that are constant multiple of s_0, h_0 .

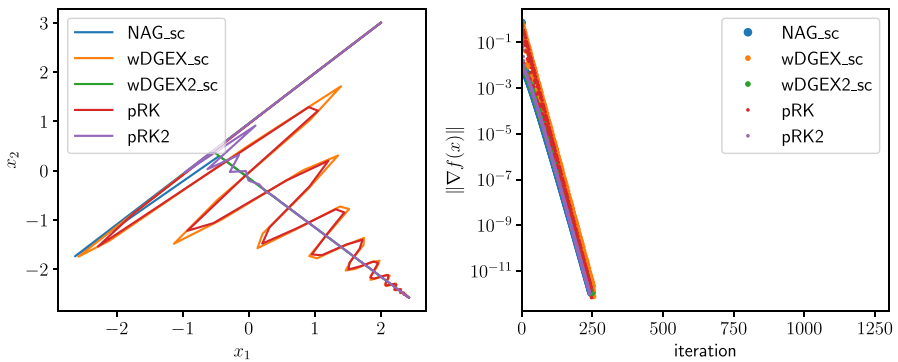
The trajectories can be roughly classified into those of (wDGEX_sc, pRK), (NAG_sc, wDGEX2_sc, pRK2) and (wDGagr_sc). The first group trajectories oscillate similarly to the original ODE trajectory, and the same level of oscillation is maintained for larger step sizes. The second group trajectories oscillate less and cease to oscillate as h increases. pRK2 has a smaller upper step size to converge than wDGEX_sc and wDGEX2_sc, which seems to explain that the solutions of pRK2 are more oscillatory near h_0 . wDGagr_sc shows unstable trajectories even at small step sizes, diverging at a step size of about $h = 0.6h_0$, which is still a step size where the Lyapunov function decreases theoretically.



(a) $h = 0.2h_0; s = 0.2^2 s_0$



(b) $h = 0.5h_0; s = 0.5^2 s_0$



(c) $h = h_0; s = s_0$

Fig. 1 Trajectories and convergences of $\|\nabla f(x^{(k)})\|$ by explicit methods for Problem (9). The initial point is (2, 3)

Meanwhile, there is almost no difference in the decrease in f for the methods other than wDGagr, except that the decrease by wDG_sc and pRK is not monotone. wDG_agr decreases f faster than the other methods for the step sizes that do not cause

divergence, but even at $h = 0.5h_0$, just before the divergence, it was not much faster than the other methods at $h = h_0$.

Next, we compare the implicit methods and NAG_sc. Figure 2 shows the results. The trajectories of NAG_sc and wDGIE_sc oscillate similarly and smaller as h goes to h_0 . wDGAVF_sc shows a larger oscillation than those two. The trajectory of wDGIA_sc is quite different from the other ones, showing a qualitatively different behavior from that of the original ODE.

Implicit methods demonstrate their strength when the objective function is a “stiff” function with either a large L or L/μ . In explicit methods, strict step-size constraints due to stiffness lead to very slow convergence. Conversely, implicit methods tend to have looser step-size constraints as long as the scheme is solvable. Hence, for implicit methods, the computational cost per step becomes significantly higher, but for stiff objective functions, this can pay off, resulting in faster overall convergence times compared to explicit methods. Let us consider the following optimization problem as an example, in which μ is very small.

$$f(x) = \frac{1}{2}x^\top Ax + \log\left(\sum_{i=1}^d \exp(0.05x_i)\right), \quad A = \begin{pmatrix} 1 & 1/2 & \cdots & 1/d \\ 1/2 & 1/3 & & \\ \vdots & & \ddots & \\ 1/d & & & 1/(2d-1) \end{pmatrix}, \tag{10}$$

where $x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ and $d = 10$.

For this problem, the results of applying WDG_sc, and WDGIE_sc are shown in Fig. 3. In terms of total computation time, the implicit method converges faster.

In the next section, we consider exploiting implicit wDG methods for not necessarily stiff problems by assuming f is a sum of convex functions.

4 Variants of proximal gradient method

4.1 Our interests

In this section, we consider the following problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \{f_1(x) + f_2(x)\}, \tag{11}$$

where f_1 is convex and L -smooth, and f_2 is convex and easy to compute its proximal map

$$\text{Prox}_{hf_2}(x) := \arg \min_{y \in \mathbb{R}^d} \left\{ f_2(y) + \frac{1}{2h} \|y - x\|^2 \right\}.$$

Examples of f_2 are L^1 , L^2 regularization terms $\lambda\|x\|_1$, $\lambda\|x\|_2^2$ and indicator function $\iota_C(x)$ of a convex set $C \subseteq \mathbb{R}^d$, i.e., $\iota_C(x) := 0$ if $x \in C$ and $\iota_C(x) := \infty$ if $x \notin C$.

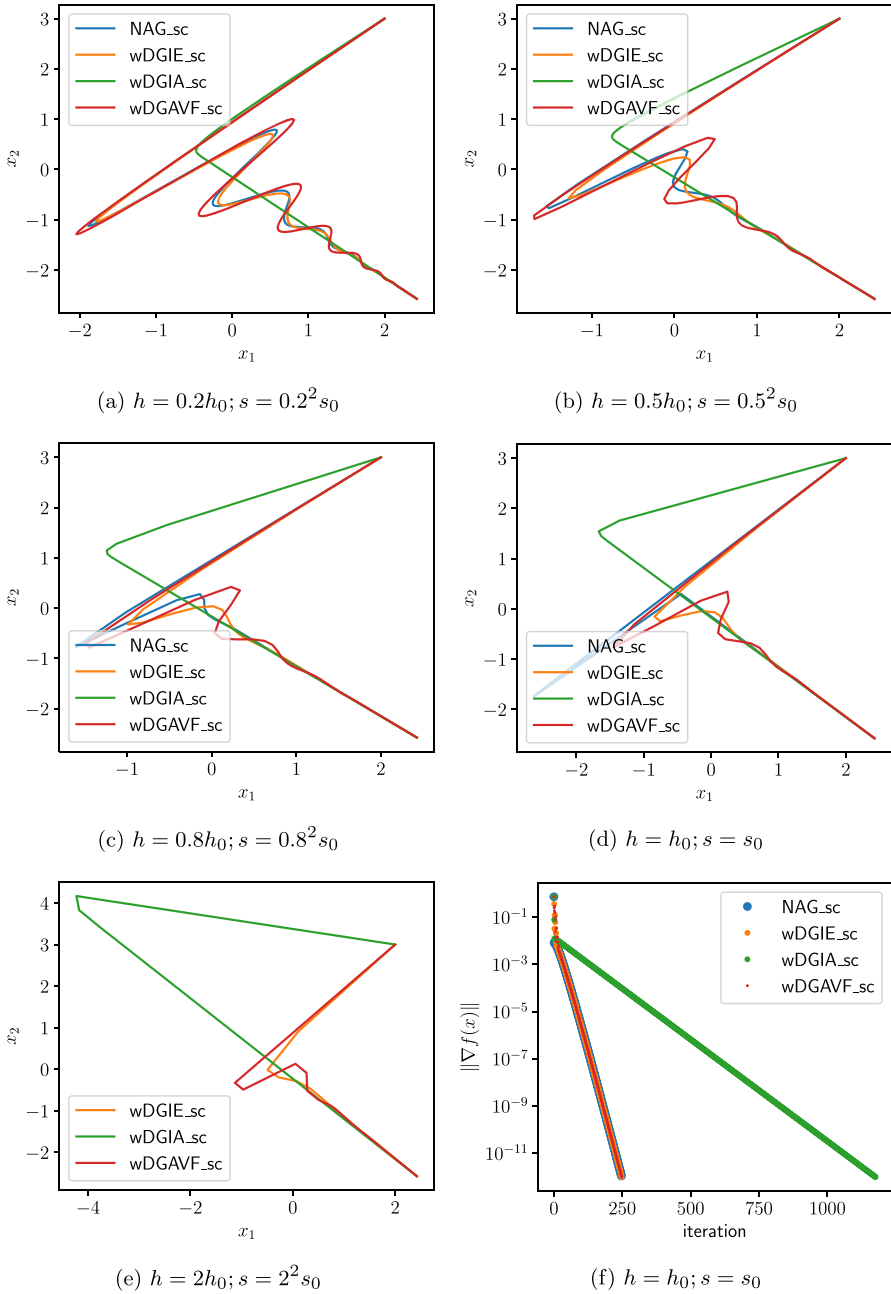


Fig. 2 Trajectories and convergences of $\|\nabla f(x^{(k)})\|$ by implicit methods for Problem (9). The initial point is (2, 3)

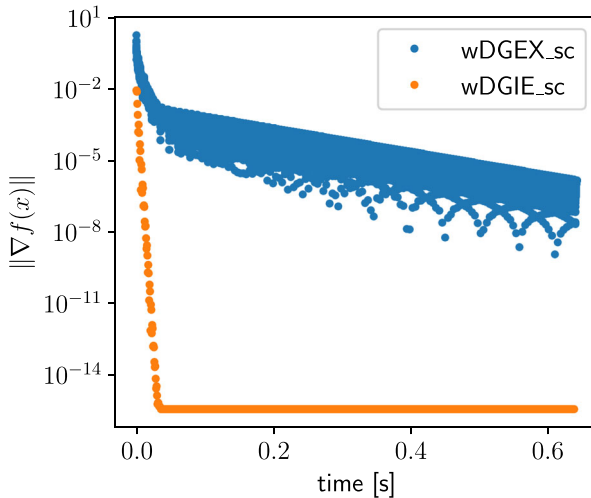


Fig. 3 Convergences of $\|\nabla f(x^{(k)})\|$ by wDGEX_sc and wDGIE_sc for Problem (10). The horizontal axis is the execution time; in the range shown in the figure, wDGEX_sc has 14,000 iterations and wDGIE_sc has 1,000 iterations. Experimental settings: Google colabotatory (OS: Ubuntu 22.04.2 LTS, CPU: Intel Xeon 2.20 GHz, Memory: 12 GiB, language: python 3.10.12)

The proximal gradient method defined below is a popular method to solve (11):

$$x^{(k+1)} = \text{Prox}_{hf_2}(x^{(k)} - h\nabla f_1(x^{(k)})).$$

For differentiable f_2 , computing this iteration is equivalent to solving the equation

$$\frac{x^{(k+1)} - x^{(k)}}{h} = -\nabla f_1(x^{(k)}) - \nabla f_2(x^{(k+1)}),$$

which is known as an IMEX-scheme for the gradient flow $\dot{x} = -\nabla f_1(x) - \nabla f_2(x)$, that is, the combination of the explicit Euler discretization of $-\nabla f_1$ and the implicit Euler discretization of $-\nabla f_2$. Since these types of discretization are members of the wDG and the sum of two wDGs is also a wDG, the convergence rate of this method is the result of the unified analysis for wDG schemes (this framework is valid for the non-differentiable f_2 using subdifferentials).

Now we consider using another wDG to discretize $-\nabla f_2$:

$$\frac{x^{(k+1)} - x^{(k)}}{h} = -\nabla f_1(x^{(k)}) - \bar{\nabla} f_2(x^{(k+1)}, x^{(k)}),$$

and assume that this scheme is easily solvable for $x^{(k+1)}$. As said before, most wDG methods are implicit and the high computational complexity generally discourages their application to large-scale problems except for highly stiff case, as discussed in the previous section. However, in the current setting, these implicit methods can

be efficiently computed, similar to the proximal gradient methods with L^1 or L^2 regularizations.

FISTA is the accelerated method of the proximal gradient method and its scheme is

$$\begin{cases} x^{(k+1)} = \text{Prox}_{hf_2}(y^{(k)} - h\nabla f_1(y^{(k)})) \\ y^{(k+1)} = x^{(k+1)} + \frac{k}{k+3}(x^{(k+1)} - x^{(k)}), \end{cases}$$

for the convex function f , and

$$\begin{cases} x^{(k+1)} = \text{Prox}_{hf_2}(y^{(k)} - h\nabla f_1(y^{(k)})) \\ y^{(k+1)} = x^{(k+1)} + \frac{1 - \sqrt{\mu h}}{1 + \sqrt{\mu h}}\nabla(x^{(k+1)} - x^{(k)}) \end{cases}$$

for strongly convex f . In this study, we distinguish these two methods by calling them FISTA_c and FISTA_sc.

The continuous limits of these methods are the accelerated gradient flows (1) and (2). In addition to the above gradient flow case, the abstract schemes (5) and (6) with the wDG defined by $\bar{\nabla} f(y, x) = \nabla f_1(x) + \bar{\nabla} f_2(y, x)$ produce new optimization methods similar to FISTA:

$$\begin{cases} x^{(k+1)} - x^{(k)} = \frac{2k+1}{k^2}(v^{(k+1)} - x^{(k+1)}), \\ v^{(k+1)} - v^{(k)} = -\frac{2k+1}{4}h^2(\nabla f_1(z^{(k)}) + \bar{\nabla} f_2(x^{(k+1)}, z^{(k)})), \\ z^{(k)} - x^{(k)} = \frac{2k+1}{(k+1)^2}(v^{(k)} - x^{(k)}) \end{cases} \tag{12}$$

for convex f , and

$$\begin{cases} x^{(k+1)} - x^{(k)} = \sqrt{2(\beta + \gamma)}h(v^{(k+1)} - x^{(k+1)}), \\ v^{(k+1)} - v^{(k)} = \sqrt{2(\beta + \gamma)}h\left(\frac{\beta}{\beta + \gamma}z^{(k)} + \frac{\gamma}{\beta + \gamma}x^{(k+1)} - v^{(k+1)} - \frac{1}{2(\beta + \gamma)}(\nabla f_1(z^{(k)}) + \bar{\nabla} f_2(x^{(k+1)}, z^{(k)}))\right), \\ z^{(k)} - x^{(k)} = \sqrt{2(\beta + \gamma)}h(x^{(k)} + v^{(k)} - 2z^{(k)}) \end{cases} \tag{13}$$

for strongly convex f . If $\bar{\nabla} f_2$ is wDG with parameters $(\alpha_2, \beta_2, \gamma_2)$, then $\bar{\nabla} f(y, x) = \nabla f_1(x) + \bar{\nabla} f_2(y, x)$ is wDG with parameters $(L_1/2 + \alpha_2, \mu_1/2 + \beta_2, \gamma_2)$, where we assume that f_1 is L_1 -smooth and μ_1 -strongly convex. Therefore, Proposition 2 implies that the scheme (12) satisfies $f(x^{(k)}) - f^* \leq O(1/k^2)$ if $h \leq 1/\sqrt{L_1 + 2\alpha_2}$, and Proposition 3 implies that the scheme (13) satisfies $f(x^{(k)}) - f^* \leq$

$O\left((1 + \sqrt{\mu_1 + 2\beta_2 + 2\gamma_2 h})^{-k}\right)$ if $h \leq (\sqrt{L_1 + 2\alpha_2 + 2\gamma_2} - \sqrt{\mu_1 + 2\beta_2 + 2\gamma_2})^{-1}$. In particular, by selecting $\bar{\nabla} f_2(x^{(k+1)}, z^{(k)}) = \nabla f_2(x^{(k+1)})$, these schemes have the same rate as FISTA but different iterative formulas, which we call IMEX_c/sc for (12)/(13). Using other wDGs for $-\nabla f_2$ in place of the implicit Euler method, we can create different methods. In this study, we consider AVF $\bar{\nabla} f_2(x^{(k+1)}, z^{(k)}) = \int_0^1 \nabla f_2(\tau x^{(k+1)} + (1 - \tau)z^{(k)})d\tau$ and we call them AVFEX_c/sc.

Our interest is the practical behavior of these wDG methods compared to FISTA. In the following section, we apply FISTA and wDG methods to several test problems. For strongly convex problems $\min f$, we use _sc methods, and for convex but not strongly convex problems, we use _c methods.

4.2 Numerical experiments

In this section, we consider several (strongly) convex functions as f_1 and two regularization terms as f_2 : L^2 regularization $f_2(x) = \frac{\lambda}{2}\|x\|_2^2$ and L^1 regularization $f_2(x) = \lambda\|x\|_1$. For L^2 regularization problems, we compare FISTA_sc, IMEX_sc and AVFEX_sc; note that adding the L^2 -regularization term makes any original convex objective functions ($f_1(x)$) strongly convex. For L^1 regularization problems, since $f_2(x) = \lambda\|x\|_1$ is not differentiable, $\nabla_{AVF} f_2$ cannot be a wDG, and we should abandon it. Thus, we only consider FISTA_sc and IMEX_sc when f_1 is strongly convex, and FISTA_c and IMEX_c when f_1 is not strongly convex (but convex). The used step size is the theoretical maximum value that guarantees the convergence rate.

We consider the following four problems.

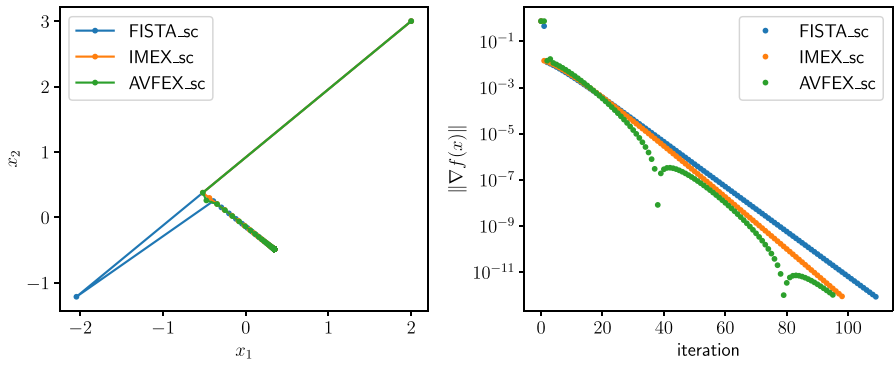
Problem 1

$$f_1(x) = \frac{1}{2}x^\top Ax + b^\top x, \quad \text{where } A = \begin{pmatrix} 0.101 & 0.099 \\ 0.099 & 0.101 \end{pmatrix}, \quad b = \begin{pmatrix} 0.01 \\ 0.02 \end{pmatrix}.$$

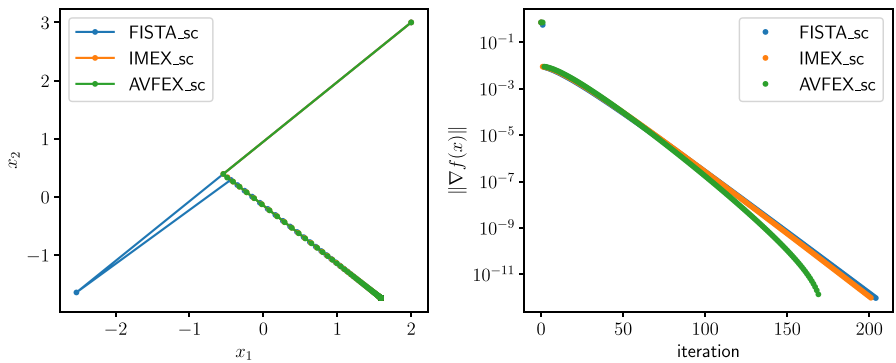
This function is 0.002-strongly convex and 0.2-smooth.

The result of L^2 regularization $f_2(x) = \frac{\lambda}{2}\|x\|_2^2$ is presented in Fig. 4. The shapes of trajectory are almost the same among the three methods and λ 's except that FISTA_sc overruns at the first step. For $\lambda = 0.01$, IMEX_sc and AVFEX_sc show slightly faster convergence than FISTA_sc, and the decrease by AVFEX_sc is oscillatory. For $\lambda = 0.001$, AVFEX_sc is faster than the other two methods. For $\lambda = 0.0001$, the three methods have the same convergence speed. Notably, the computational complexities per step of the tested methods are roughly the same, and the comparison of the convergence profiles against steps is fair. This notice applies to similar figures in this section.

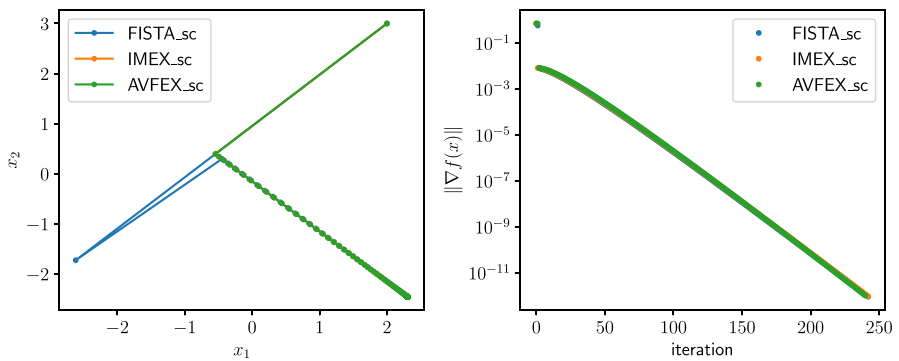
The result of L^1 regularization $f_2(x) = \lambda\|x\|_1$ is Fig. 5. In this case, FISTA_sc and IMEX_sc show similar trajectories and convergences in $\lambda = 0.01, 0.001$ and more λ 's.



(a) $\lambda = 0.01$



(b) $\lambda = 0.001$



(c) $\lambda = 0.0001$

Fig. 4 Trajectories and convergences of $\|\nabla f\|$ by FISTA_sc, IMEX_sc and AVFEX_sc for Problem 1 with the L^2 regularization

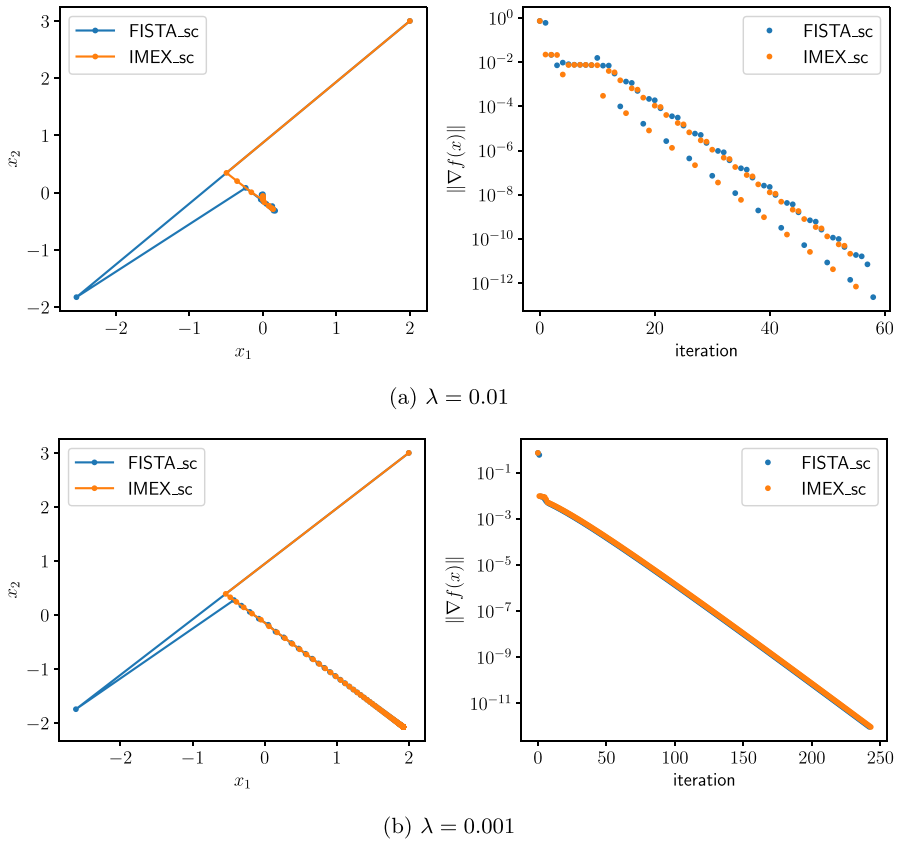


Fig. 5 Trajectories and convergences of $\|\nabla f\|$ by FISTA_sc and IMEX_sc for Problem 1 with the L^1 regularization

Problem 2

$$f_1(x) = \frac{10}{4} \left(\frac{1}{2} \left[x_1^2 + \sum_{i=1}^{1000} (x_i - x_{i+1})^2 + x_{1000}^2 \right] - x_1 \right).$$

This function has $L \approx 10$ and $\mu \approx 1.57 \times 10^{-2}$.

Figure 6 shows the convergence of $\|\nabla f\|$. Similar to Problem 1, in the L^2 regularization setting, IMEX_sc and AVFEX_sc are faster than FISTA_sc at large λ , and they have the same speed at small λ . In the L^1 regularization setting, there is no difference in the convergence of $\|\nabla f\|$ between FISTA_sc and IMEX_sc for $\lambda = 0.1$ and some more λ 's.

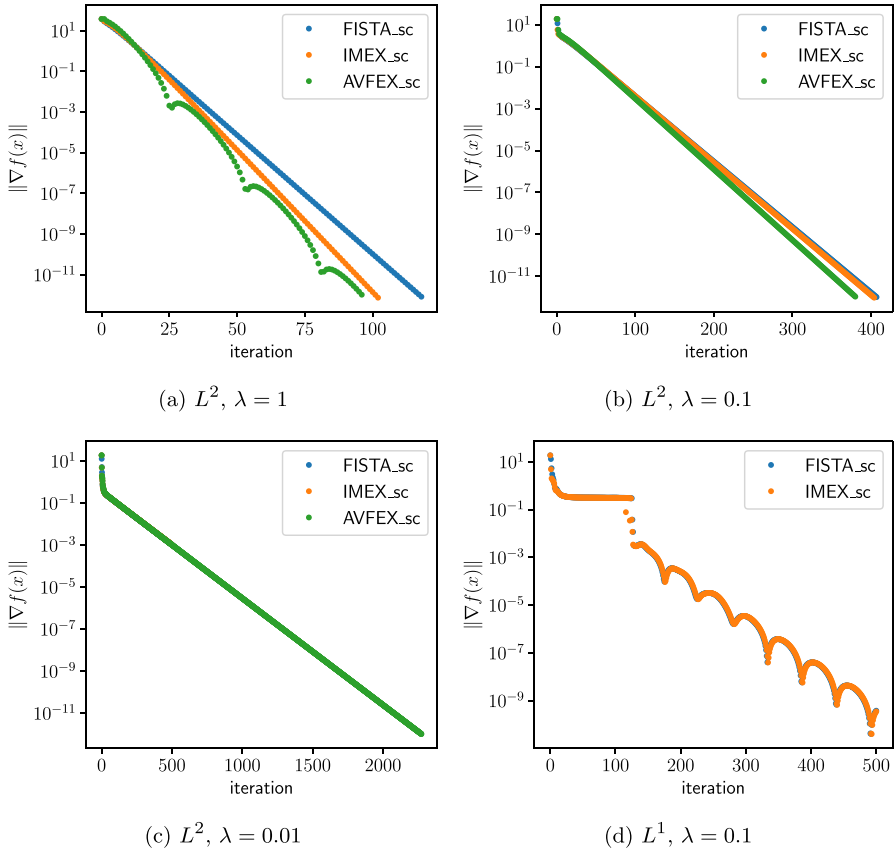


Fig. 6 Convergences of $\|\nabla f\|$ for Problem 2 by FISTA_sc, IMEX_sc and AVFEX_sc for the L^2 regularization (a–c) and by FISTA_sc and IMEX_sc for the L^1 regularization (d)

Problem 3 [15]

$$f_1(x) = \sum_{i=1}^{50} g(a_i^\top x - b_i), \quad g(x) = \begin{cases} \frac{1}{2}x^2e^{-r/x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

where $x, a_i \in \mathbb{R}^{1000}, b_i \in \mathbb{R} (i = 1, \dots, 50)$ and $r = 0.001$. Parameters a_i, b_i are chosen randomly. This is a convex but not strongly convex function. With the chosen parameters, L is approximately 1548. For this problem, we consider methods for strongly convex functions in the case of the L^2 -regularization, and those for convex functions in the case of the L^1 -regularization (as described before). This notice applies to Problem 4 as well.

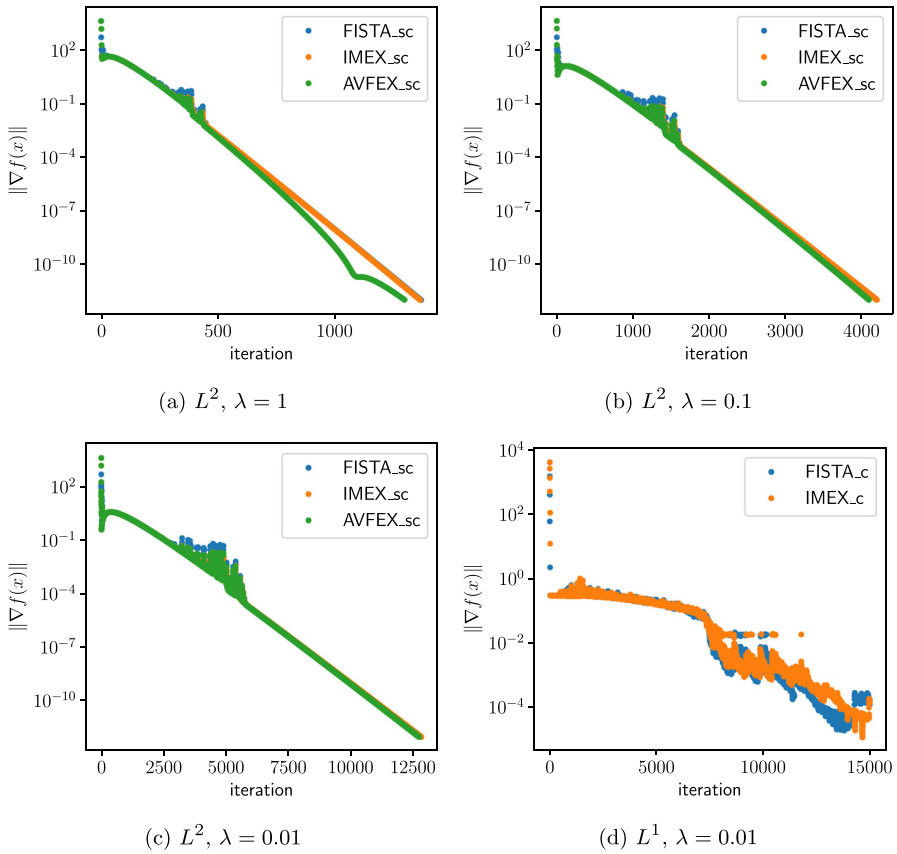


Fig. 7 Convergences of $\|\nabla f\|$ for Problem 3 by FISTA_sc, IMEX_sc and AVFEX_sc for the L^2 regularization (a–c) and by FISTA_c and IMEX_c for the L^1 regularization (d)

The results are shown in Fig. 7. In the L^2 regularization setting, at $\lambda = 1$, AVF has an advantage, but there is no significant difference between the methods in performance as in the previous two cases. For the L^1 setting, no _c method achieved successful optimization.

Problem 4

$$f_1(x) = \sum_{i=1}^{200} \log\left(1 + e^{-y_i \xi_i^T x}\right),$$

where $\xi_i \in \mathbb{R}^d$ and $y_i \in \{1, -1\}$ ($i = 1, \dots, 200$) are randomly chosen. This is a convex but not strongly convex function appearing in the logistic regression. With the chosen parameters, L is approximately 1549.

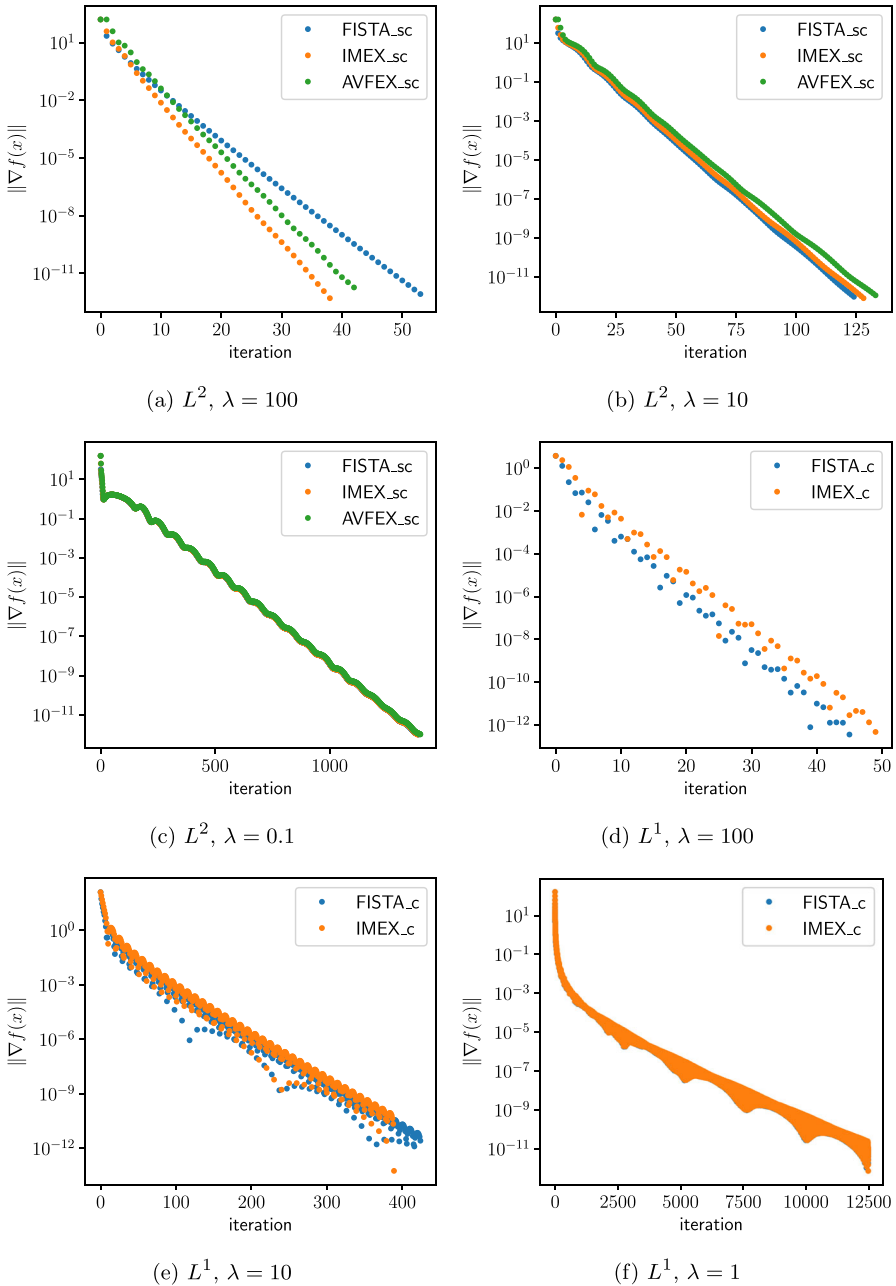


Fig. 8 Convergences of $\|\nabla f\|$ for Problem 2 by FISTA_sc, IMEX_sc and AVFEX_sc for the L^2 regularization (a–c) and by FISTA_c and IMEX_c for the L^1 regularization (d–f)

The results are shown in Fig. 8. The optimization algorithms exhibited different behavior for this problem compared to previous ones. For $\lambda = 100$, IMEX_sc is the fastest, followed by AVFEX_sc and FISTA_sc. However, for $\lambda = 10$, the order of speed is as follows: FISTA_sc, IMEX_sc, and AVFEX_sc. For $\lambda = 0.1$, there are no differences in speed.

5 Summary and discussions

In this section, we summarize the results of the numerical experiments to answer the three issues raised in the Introduction.

5.1 Issue 1: differences between wDGs and other methods

In the preliminary numerical experiment in [7], the advantage of wDG methods was not clear both for strongly convex and convex objective functions, as mentioned in Introduction. Below are the novel contributions of the present study in the context of strongly convex functions.

We first observed in Sect. 3.2 that the convergence profiles are roughly the same for all the tested methods except for wDGagr_sc. This means that wDG methods for the (strongly convex) NAG ODE perform well at the same level as NAG_sc. Hence, one might think that constructing optimization methods via the wDG framework is not really advantageous. However, one important observation is that wDGEX_sc, which is a wDG scheme that is quite natural as a numerical method for the NAG ODE, is competitive in comparison with other methods including NAG_sc (which is not a natural discretization). This is more preferable from a numerical analysis perspective.

Another important observation there is that wDGagr_sc converges much faster than other methods; however unfortunately a theoretical convergence estimate is yet to be given. Nevertheless, this verifies the potential ability of the wDG framework; if we succeed in finding better ODE and/or its Lyapunov function, we can immediately construct a method that can outperform existing methods. Although wDGagr_sc tends to be unstable for larger time steps, this does not mean it is inferior in this respect. By comparing Cases (b) and (c) in Fig. 1, in (b), wDGagr_sc converges more than twice faster than other methods. That is, it performs better than other methods in (c).

Additionally, we observe that even when convergence profiles are similar, the corresponding trajectories in the solution space can be different (recall Problem 1.) This encourages us to further explore various wDGs. A negative example in this respect is the wDGIA_sc, the implicit method based on the Itoh–Abe discrete gradient, which performs very poorly. This warns us that even though the IA discrete gradient is now drawing attention in the context of optimization (due to its possible high computational efficiency compared to other implicit methods), its use is not always encouraged.

In Sect. 4.2, we observe that the implicit wDG methods are as highly efficient as the standard proximal gradient methods (with L^2 - or L^1 -regularizations), and the convergence profiles are similar, or even slightly better (within the experiments in the

present study). In this sense, we should consider such wDG methods in the situations where FISTA is needed.

5.2 Issue 2: practicability of implicit wDG methods

This is investigated in the following two ways, in the case of strongly convex objective functions: First, in Sect. 3.2, it is confirmed that for some highly stiff objective functions, implicit optimization methods can actually be advantageous. Although this finding is widely known in the numerical analysis (of ODEs) field, it has not been established in the optimization field. Next, in Sect. 4.2, it is confirmed (as mentioned above) that implicit wDG methods are of practical use when we consider proximal gradient methods. Hence, implicit wDG methods are in fact practical under some circumstances. This greatly widens the scope of new optimization methods; we have various implicit ODE solvers in numerical analysis that have been proven to deserve consideration.

5.3 Issue 3: efficiency of proximal gradient-like wDG methods

The practicability of the proximal gradient-like methods in the case of strongly convex objective functions is discussed above. The last two problems (Problems 3 and 4) contained the case of (not strongly) convex functions. The results show that the behaviors are roughly the same. This means that there is no strong evidence to recommend wDG methods, but at the same time no reason to avoid using them. Further investigations with other wDGs might reveal meaningful differences.

6 Concluding remarks

In this paper, we conducted numerical experiments on methods based on wDG framework to validate their properties and utility. The results show that in most cases, wDG methods show competitive performance as compared to typical optimization methods (e.g., NAG and FISTA) and can even slightly outperform them. Additionally, the results show that implicit methods deserve further consideration. This verifies the possibility of wDG framework, where in many cases the resulting method becomes implicit. This is consistent with the fact that structure-preserving numerical methods are recommended to ODEs that are difficult to solve with generic methods. We hope that this paper encourages researchers in the optimization and numerical analysis communities to build on the framework and enrich optimization researches.

Appendix A Convergence analysis for pRK_sc and pRK2_sc method

Partitioned Runge–Kutta methods are numerical methods for ODEs in the partitioned form

$$\dot{y} = g_1(y, z), \quad \dot{z} = g_2(y, z). \quad (\text{A1})$$

Definition 2 (cf. [14]) Let b_i, a_{ij} and \hat{b}_i, \hat{a}_{ij} be the coefficients of two Runge–Kutta methods. A partitioned Runge–Kutta method for (A1) is given by

$$\begin{aligned} \phi_i &= y^{(0)} + h \sum_{j=1}^s a_{ij} g_1(\phi_j, \psi_j), & \psi_i &= z^{(0)} + h \sum_{j=1}^s \hat{a}_{ij} g_2(\phi_j, \psi_j), \\ y^{(1)} &= y^{(0)} + h \sum_{i=1}^s b_i g_1(\phi_i, \psi_i), & z^{(1)} &= z^{(0)} + h \sum_{i=1}^s \hat{b}_i g_2(\phi_i, \psi_i). \end{aligned}$$

Their convergence issues as numerical methods can be found, for example, in [16]. Below we discuss their convergence as optimization methods (in terms of $f(x) - f(x^*)$).

A.1 Convergence analysis for pRK method

We discretize the following system with a partitioned Runge–Kutta method:

$$\begin{cases} \dot{x} = v - x, \\ \dot{v} = x - v - \nabla f(x)/\mu. \end{cases}$$

This is the time-scaled version of the ODE (2). The partitioned Runge–Kutta method corresponding to the tableau

$$\begin{array}{c|cc} \alpha & \alpha & 0 \\ \alpha & \alpha & 0 \\ \hline 1 & \alpha & 1 - \alpha \end{array} \quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \beta & 1 - \beta \\ \hline 1 & \beta & 1 - \beta \end{array}$$

(the parameters $\alpha, \beta \in [0, 1]$ will be determined later) reads

$$\begin{cases} z^{(k)} - x^{(k)} = \alpha h(v^{(k)} - z^{(k)}), \\ v^{(k+1)} - v^{(k)} = h(z^{(k)} - \tilde{v} - \nabla f(z^{(k)})/\mu), \\ x^{(k+1)} - z^{(k)} = (1 - \alpha)h(v^{(k+1)} - z^{(k)}), \end{cases} \tag{A2}$$

where $\tilde{v} = (1 - \beta)v^{(k+1)} + \beta v^{(k)}$. By setting $\alpha = \beta = 1/2$ we have Stömer–Verlet method, which is symplectic. For this method, the wDG theory cannot be directly applied, but the convergence can be shown by a similar discussion. Let us consider the Lyapunov function:

$$E^{(k)} = f(x^{(k)}) - f^* + \frac{\mu}{2} \|v^{(k)} - x^*\|^2. \tag{A3}$$

Its time difference reads

$$\begin{aligned}
 E^{(k+1)} - E^{(k)} &= f(x^{(k+1)}) - f(x^{(k)}) + \frac{\mu}{2} (\|v^{(k+1)} - x^*\|^2 - \|v^{(k)} - x^*\|^2) \\
 &= f(x^{(k+1)}) - f(z^{(k)}) + f(z^{(k)}) - f(x^{(k)}) \\
 &\quad + \frac{\mu}{2} (\|v^{(k+1)} - x^*\|^2 - \|v^{(k)} - x^*\|^2) \\
 &\leq \langle \nabla f(z^{(k)}), x^{(k+1)} - z^{(k)} \rangle + \langle \nabla f(z^{(k)}), z^{(k)} - x^{(k)} \rangle \\
 &\quad + \mu((1 - \alpha)\langle v^{(k+1)} - v^{(k)}, v^{(k+1)} - x^* \rangle + \alpha\langle v^{(k+1)} - v^{(k)}, v^{(k)} - x^* \rangle) \\
 &\quad + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} (1 - 2\alpha)\|v^{(k+1)} - v^{(k)}\|^2 \\
 &= (1 - \alpha)h\langle \nabla f(z^{(k)}), v^{(k+1)} - z^{(k)} \rangle + \alpha h\langle \nabla f(z^{(k)}), v^{(k)} - z^{(k)} \rangle \\
 &\quad + \mu((1 - \alpha)h\langle z^{(k)} - \tilde{v} - \nabla f(z^{(k)})/\mu, v^{(k+1)} - x^* \rangle \\
 &\quad \quad \quad + \alpha h\langle z^{(k)} - \tilde{v} - \nabla f(z^{(k)})/\mu, v^{(k)} - x^* \rangle) \\
 &\quad + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} (1 - 2\alpha)\|v^{(k+1)} - v^{(k)}\|^2 \\
 &= (1 - \alpha)h(\langle \nabla f(z^{(k)}), x^* - z^{(k)} \rangle - \mu\langle \tilde{v} - z^{(k)}, v^{(k+1)} - x^* \rangle) \\
 &\quad + \alpha h(\langle \nabla f(z^{(k)}), x^* - z^{(k)} \rangle - \mu\langle \tilde{v} - z^{(k)}, v^{(k)} - x^* \rangle) \\
 &\quad + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} (1 - 2\alpha)\|v^{(k+1)} - v^{(k)}\|^2 \\
 &= (1 - \alpha)h \left(\langle \nabla f(z^{(k)}), x^* - z^{(k)} \rangle \right. \\
 &\quad \quad \quad \left. - \frac{\mu}{2} (\|v^{(k+1)} - z^{(k)}\|^2 + \|\tilde{v} - x^*\|^2 - \|z^{(k)} - x^*\|^2 - \|\tilde{v} - v^{(k+1)}\|^2) \right) \\
 &\quad + \alpha h \left(\langle \nabla f(z^{(k)}), x^* - z^{(k)} \rangle \right. \\
 &\quad \quad \quad \left. - \frac{\mu}{2} (\|v^{(k)} - z^{(k)}\|^2 + \|\tilde{v} - x^*\|^2 - \|z^{(k)} - x^*\|^2 - \|\tilde{v} - v^{(k)}\|^2) \right) \\
 &\quad + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} (1 - 2\alpha)\|v^{(k+1)} - v^{(k)}\|^2 \\
 &\leq -(1 - \alpha)h \left(f(z^{(k)}) - f^* + \frac{\mu}{2} \|\tilde{v} - x^*\|^2 \right) - \alpha h \left(f(z^{(k)}) - f^* + \frac{\mu}{2} \|\tilde{v} - x^*\|^2 \right) \\
 &\quad - \frac{\mu}{2} (1 - \alpha)h (\|v^{(k+1)} - z^{(k)}\|^2 - \|\tilde{v} - v^{(k+1)}\|^2) \\
 &\quad - \frac{\mu}{2} \alpha h (\|v^{(k)} - z^{(k)}\|^2 - \|\tilde{v} - v^{(k)}\|^2) \\
 &\quad + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} (1 - 2\alpha)\|v^{(k+1)} - v^{(k)}\|^2 \\
 &= -h \left(f(z^{(k)}) - f^* + \frac{\mu}{2} \|\tilde{v} - x^*\|^2 \right) \\
 &\quad - \frac{\mu}{2} h ((1 - \alpha)\|v^{(k+1)} - z^{(k)}\|^2 + \alpha\|v^{(k)} - z^{(k)}\|^2) \\
 &\quad + \frac{\mu}{2} h ((1 - \alpha)\|\tilde{v} - v^{(k+1)}\|^2 + \alpha\|\tilde{v} - v^{(k)}\|^2) \\
 &\quad + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} (1 - 2\alpha)\|v^{(k+1)} - v^{(k)}\|^2.
 \end{aligned}$$

The first inequality is from the smoothness of the objective function, and the second is from the strong convexity.

If we further note

$$\|\tilde{v} - x^*\|^2 = (1 - \beta)\|v^{(k+1)} - x^*\|^2 + \beta\|v^{(k)} - x^*\|^2 - \beta(1 - \beta)\|v^{(k+1)} - v^{(k)}\|^2,$$

which is visible from the definition of \tilde{v} , we obtain

$$\begin{aligned} &(1 - \alpha)\|\tilde{v} - v^{(k+1)}\|^2 + \alpha\|\tilde{v} - v^{(k)}\|^2 \\ &= ((1 - \alpha)\beta^2 + \alpha(1 - \beta)^2)\|v^{(k+1)} - v^{(k)}\|^2 \\ &= ((\beta - \alpha)^2 + \alpha(1 - \alpha))\|v^{(k+1)} - v^{(k)}\|^2. \end{aligned}$$

From this we see

$$\begin{aligned} &E^{(k+1)} - E^{(k)} \\ &\leq -(1 - \beta)h\left(f(x^{(k+1)}) - f^* + \frac{\mu}{2}\|v^{(k+1)} - x^*\|^2\right) \\ &\quad - \beta h\left(f(x^{(k)}) - f^* + \frac{\mu}{2}\|v^{(k)} - x^*\|^2\right) \\ &\quad + (1 - \beta)h(f(x^{(k+1)}) - f(z^{(k)})) + \beta h(f(x^{(k)}) - f(z^{(k)})) \\ &\quad + \frac{\mu}{2}\beta(1 - \beta)h\|v^{(k+1)} - v^{(k)}\|^2 \\ &\quad - \frac{\mu}{2}h((1 - \alpha)\|v^{(k+1)} - z^{(k)}\|^2 + \alpha\|v^{(k)} - z^{(k)}\|^2) \\ &\quad + \frac{\mu}{2}h((\beta - \alpha)^2 + \alpha(1 - \alpha))\|v^{(k+1)} - v^{(k)}\|^2 \\ &\quad + \frac{L}{2}\|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2}\|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2}(1 - 2\alpha)\|v^{(k+1)} - v^{(k)}\|^2. \end{aligned}$$

Let (err) denote the terms in the last five lines on the right side of the above inequality. If $(err) \leq 0$, then we have

$$(1 + (1 - \beta)h)E^{(k+1)} \leq (1 - \beta h)E^{(k)},$$

which implies the desired linear convergence.

Let us show $(err) \leq 0$, and reveal the constant of the linear convergence. Using

$$\begin{aligned} \nabla f(z^{(k)}) &= -\mu\left(\frac{v^{(k+1)} - v^{(k)}}{h} + \tilde{v} - z^{(k)}\right) \\ &= -\mu\left(\frac{v^{(k+1)} - v^{(k)}}{h} + (1 - \beta)v^{(k+1)} + \beta v^{(k)} - z^{(k)}\right) \\ &= -\mu\left(\frac{1 - \beta h}{h}(v^{(k+1)} - v^{(k)}) + v^{(k+1)} - z^{(k)}\right) \\ &= -\mu\left(\frac{1 + (1 - \beta)h}{h}(v^{(k+1)} - v^{(k)}) + v^{(k)} - z^{(k)}\right), \end{aligned} \tag{A4}$$

we see that the first line of (err) can be bounded as

$$\begin{aligned}
 & (1 - \beta)h(f(x^{(k+1)}) - f(z^{(k)})) + \beta h(f(x^{(k)}) - f(z^{(k)})) \\
 & \leq (1 - \beta)h\langle \nabla f(z^{(k)}), x^{(k+1)} - z^{(k)} \rangle - \beta h\langle \nabla f(z^{(k)}), z^{(k)} - x^{(k)} \rangle \\
 & \quad + \frac{L}{2}(1 - \beta)h\|x^{(k+1)} - z^{(k)}\|^2 + \frac{L}{2}\beta h\|x^{(k)} - z^{(k)}\|^2 \\
 & = -\mu(1 - \beta)(1 - \alpha)h^2 \left\langle \frac{1 - \beta h}{h}(v^{(k+1)} - v^{(k)}) + v^{(k+1)} - z^{(k)}, v^{(k+1)} - z^{(k)} \right\rangle \\
 & \quad + \mu\beta\alpha h^2 \left\langle \frac{1 + (1 - \beta)h}{h}(v^{(k+1)} - v^{(k)}) + v^{(k)} - z^{(k)}, v^{(k)} - z^{(k)} \right\rangle \\
 & \quad + \frac{L}{2}(1 - \beta)h\|x^{(k+1)} - z^{(k)}\|^2 + \frac{L}{2}\beta h\|x^{(k)} - z^{(k)}\|^2 \\
 & = -\mu(1 - \beta)(1 - \alpha)(1 - \beta h)h \frac{1}{2}(\|v^{(k+1)} - v^{(k)}\|^2 + \|v^{(k+1)} - z^{(k)}\|^2 - \|v^{(k)} - z^{(k)}\|^2) \\
 & \quad - \mu\beta\alpha(1 + (1 - \beta)h)h \frac{1}{2}(\|v^{(k+1)} - v^{(k)}\|^2 + \|v^{(k)} - z^{(k)}\|^2 - \|v^{(k+1)} - z^{(k)}\|^2) \\
 & \quad - \mu(1 - \beta)(1 - \alpha)h^2\|v^{(k+1)} - z^{(k)}\|^2 + \mu\beta\alpha h^2\|v^{(k)} - z^{(k)}\|^2 \\
 & \quad + \frac{L}{2}(1 - \beta)(1 - \alpha)^2 h^3\|v^{(k+1)} - z^{(k)}\|^2 + \frac{L}{2}\beta\alpha^2 h^3\|v^{(k)} - z^{(k)}\|^2.
 \end{aligned}$$

The inequality is from the discrete chain rule.

If we collect the coefficients of $\|v^{(k+1)} - v^{(k)}\|^2$ in (err), we find

$$\begin{aligned}
 & -\frac{\mu}{2}h[(1 - \beta)(1 - \alpha)(1 - \beta h) + \beta\alpha(1 + (1 - \beta)h)] \\
 & + \frac{\mu}{2}h[\beta(1 - \beta) + (\beta - \alpha)^2 + \alpha(1 - \alpha)] - \frac{\mu}{2}(1 - 2\alpha) \\
 & = \frac{\mu}{2}(1 - 2\alpha)(\beta h - 1)((1 - \beta)h + 1). \tag{A5}
 \end{aligned}$$

Similarly, the coefficient of $\|v^{(k+1)} - z^{(k)}\|^2$ is

$$\begin{aligned}
 & -\frac{\mu}{2}h[(1 - \beta)(1 - \alpha)(1 - \beta h) - \beta\alpha(1 + (1 - \beta)h) + 2(1 - \beta)(1 - \alpha)h] \\
 & + \frac{L}{2}(1 - \beta)(1 - \alpha)^2 h^3 - \frac{\mu}{2}h(1 - \alpha) + \frac{L}{2}(1 - \alpha)^2 h^2 \\
 & = \frac{1}{2}(1 + (1 - \beta)h)h(-\mu(2(1 - \alpha) - \beta) + L(1 - \alpha)^2 h), \tag{A6}
 \end{aligned}$$

and the coefficient of $\|v^{(k)} - z^{(k)}\|^2$ is

$$\begin{aligned}
 & \frac{\mu}{2}h[(1 - \beta)(1 - \alpha)(1 - \beta h) - \beta\alpha(1 + (1 - \beta)h) + 2\beta\alpha h] + \frac{L}{2}\beta\alpha^2 h^3 \\
 & - \frac{\mu}{2}h\alpha - \frac{\mu}{2}\alpha^2 h^2. \tag{A7}
 \end{aligned}$$

The three quantities above should be nonpositive. From these restrictions and also from (A5) and (A6), we have the conditions

$$\alpha \leq \frac{1}{2}, \quad h \leq \frac{1}{\beta}, \quad h \leq \frac{\mu(2(1-\alpha)-\beta)}{L(1-\alpha)^2}. \tag{A8}$$

If $\alpha > 1/2$, no h can satisfy this condition. This means h cannot be larger than $O(\mu/L)$. The restriction on h from (A8) becomes the mildest when $\beta = 0$. (This is interesting because this might imply that symplectic methods are not necessarily suitable for optimization.) In this case, the required condition for (A7) becomes

$$\frac{1-2\alpha}{\alpha^2} \leq h.$$

This condition vanishes by setting $\alpha = 1/2$.

In summary, when $\alpha = 1/2$, and $\beta = 0$, the step size condition becomes $h \leq 4\mu/L$, and the convergence rate is

$$E^{(k)} \leq \left(\frac{1-h\beta}{1+h(1-\beta)} \right)^k E^{(0)} = \left(\frac{1}{1+h} \right)^k E^{(0)}.$$

By further replacing h with $\sqrt{\mu}h$, we obtain the scheme corresponding to (7).

A.2 Convergence analysis for the pRK2 method

To obtain a better convergence rate, we modify (A2) by adding an extra term of order h^2 :

$$\begin{cases} z^{(k)} - x^{(k)} = \alpha h(v^{(k)} - z^{(k)}), \\ v^{(k+1)} - v^{(k)} = h(z^{(k)} - \tilde{v} - \nabla f(z^{(k)})/\mu), \\ x^{(k+1)} - z^{(k)} = (1-\alpha)h(v^{(k+1)} - z^{(k)}) + \alpha h^2 \frac{v^{(k+1)} - v^{(k)}}{h}, \end{cases}$$

where $\tilde{v} = (1-\beta)v^{(k+1)} + \beta v^{(k)}$.

For this, we consider the Lyapunov function (A3) again and perform the same discussion.

$$\begin{aligned} E^{(k+1)} - E^{(k)} &= f(x^{(k+1)}) - f(z^{(k)}) + f(z^{(k)}) - f(x^{(k)}) \\ &\quad + \frac{\mu}{2} (\|v^{(k+1)} - x^*\|^2 - \|v^{(k)} - x^*\|^2) \\ &\stackrel{\text{chain}}{\leq} \langle \nabla f(z^{(k)}), x^{(k+1)} - z^{(k)} \rangle + \langle \nabla f(z^{(k)}), z^{(k)} - x^{(k)} \rangle \\ &\quad + \mu \langle v^{(k+1)} - v^{(k)}, v^{(k+1)} - x^* \rangle \\ &\quad + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} \|v^{(k+1)} - v^{(k)}\|^2 \\ &= (1-\alpha)h \langle \nabla f(z^{(k)}), v^{(k+1)} - z^{(k)} \rangle + \alpha h \langle \nabla f(z^{(k)}), v^{(k)} - z^{(k)} \rangle \end{aligned}$$

$$\begin{aligned}
 & + \alpha \langle \nabla f(z^{(k)}), h(v^{(k+1)} - v^{(k)}) \rangle \\
 & + \mu h \langle z^{(k)} - \tilde{v} - \nabla f(z^{(k)})/\mu, v^{(k+1)} - x^* \rangle \\
 & + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} \|v^{(k+1)} - v^{(k)}\|^2 \\
 = & h \langle \nabla f(z^{(k)}), v^{(k+1)} - z^{(k)} \rangle + \mu h \langle z^{(k)} - \tilde{v} - \nabla f(z^{(k)})/\mu, v^{(k+1)} - x^* \rangle \\
 & + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} \|v^{(k+1)} - v^{(k)}\|^2 \\
 \leq_{sc} & -h \left(f(z^{(k)}) - f^* + \frac{\mu}{2} \|\tilde{v} - x^*\|^2 \right) \\
 & - \frac{\mu}{2} h \|v^{(k+1)} - z^{(k)}\|^2 + \frac{\mu}{2} h \|\tilde{v} - v^{(k+1)}\|^2 \\
 & + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} \|v^{(k+1)} - v^{(k)}\|^2 \\
 = & -(1 - \beta)h \left(f(x^{(k+1)}) - f^* + \frac{\mu}{2} \|v^{(k+1)} - x^*\|^2 \right) \\
 & - \beta h \left(f(x^{(k)}) - f^* + \frac{\mu}{2} \|v^{(k)} - x^*\|^2 \right) \\
 & + (1 - \beta)h(f(x^{(k+1)}) - f(z^{(k)})) + \beta h(f(x^{(k)}) - f(z^{(k)})) + \frac{\mu}{2} \beta(1 - \beta)h \|v^{(k+1)} - v^{(k)}\|^2 \\
 & - \frac{\mu}{2} h \|v^{(k+1)} - z^{(k)}\|^2 + \frac{\mu}{2} h \beta^2 \|v^{(k+1)} - v^{(k)}\|^2 \\
 & + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 - \frac{\mu}{2} \|z^{(k)} - x^{(k)}\|^2 - \frac{\mu}{2} \|v^{(k+1)} - v^{(k)}\|^2.
 \end{aligned}$$

Let (err) denote the terms in the last three lines on the right side. Using (A4), we have

$$\begin{aligned}
 & (1 - \beta)h(f(x^{(k+1)}) - f(z^{(k)})) + \beta h(f(x^{(k)}) - f(z^{(k)})) \\
 & \leq (1 - \beta)h \langle \nabla f(z^{(k)}), x^{(k+1)} - z^{(k)} \rangle - \beta h \langle \nabla f(z^{(k)}), z^{(k)} - x^{(k)} \rangle \\
 & \quad + \frac{L}{2} (1 - \beta)h \|x^{(k+1)} - z^{(k)}\|^2 + \frac{L}{2} \beta h \|x^{(k)} - z^{(k)}\|^2 \\
 = & -\mu(1 - \beta)h^2 \left\langle \frac{1 - \beta h}{h} (v^{(k+1)} - v^{(k)}) + v^{(k+1)} - z^{(k)}, \right. \\
 & \qquad \qquad \qquad \left. (1 - \alpha)(v^{(k+1)} - z^{(k)}) + \alpha(v^{(k+1)} - v^{(k)}) \right\rangle \\
 & + \mu\beta\alpha h^2 \left\langle \frac{1 + (1 - \beta)h}{h} (v^{(k+1)} - v^{(k)}) + v^{(k)} - z^{(k)}, v^{(k)} - z^{(k)} \right\rangle \\
 & + \frac{L}{2} (1 - \beta)h \|x^{(k+1)} - z^{(k)}\|^2 + \frac{L}{2} \beta h \|x^{(k)} - z^{(k)}\|^2 \\
 = & -\mu(1 - \beta)(1 - \alpha)(1 - \beta h)h \frac{1}{2} (\|v^{(k+1)} - v^{(k)}\|^2 + \|v^{(k+1)} - z^{(k)}\|^2 - \|v^{(k)} - z^{(k)}\|^2) \\
 & - \mu(1 - \beta)\alpha h (\|v^{(k+1)} - v^{(k)}\|^2 \\
 & \quad + \frac{1}{2} h ((1 - 2\beta)\|v^{(k+1)} - v^{(k)}\|^2 + \|v^{(k+1)} - z^{(k)}\|^2 - \|v^{(k)} - z^{(k)}\|^2)) \\
 & - \mu\beta\alpha(1 + (1 - \beta)h)h \frac{1}{2} (\|v^{(k+1)} - v^{(k)}\|^2 + \|v^{(k)} - z^{(k)}\|^2 - \|v^{(k+1)} - z^{(k)}\|^2) \\
 & - \mu(1 - \beta)(1 - \alpha)h^2 \|v^{(k+1)} - z^{(k)}\|^2 + \mu\beta\alpha h^2 \|v^{(k)} - z^{(k)}\|^2 \\
 & + \frac{L}{2} (1 - \beta)h \|x^{(k+1)} - z^{(k)}\|^2 + \frac{L}{2} \beta\alpha^2 h^3 \|v^{(k)} - z^{(k)}\|^2.
 \end{aligned}$$

Thus, we have

$$\begin{aligned}
 &(\text{err}) \\
 &\leq -\frac{\mu}{2}(1-\beta)(1-\alpha)(1-\beta h)h(\|v^{(k+1)}-v^{(k)}\|^2 + \|v^{(k+1)}-z^{(k)}\|^2 - \|v^{(k)}-z^{(k)}\|^2) \\
 &\quad -\mu(1-\beta)\alpha h(\|v^{(k+1)}-v^{(k)}\|^2 \\
 &\quad\quad +\frac{1}{2}h((1-2\beta)\|v^{(k+1)}-v^{(k)}\|^2 + \|v^{(k+1)}-z^{(k)}\|^2 - \|v^{(k)}-z^{(k)}\|^2)) \\
 &\quad -\frac{\mu}{2}\beta\alpha(1+(1-\beta)h)h(\|v^{(k+1)}-v^{(k)}\|^2 + \|v^{(k)}-z^{(k)}\|^2 - \|v^{(k+1)}-z^{(k)}\|^2) \\
 &\quad -\mu(1-\beta)(1-\alpha)h^2\|v^{(k+1)}-z^{(k)}\|^2 + \mu\beta\alpha h^2\|v^{(k)}-z^{(k)}\|^2 \\
 &\quad +\frac{L}{2}(1-\beta)h\|x^{(k+1)}-z^{(k)}\|^2 + \frac{L}{2}\beta\alpha^2 h^3\|v^{(k)}-z^{(k)}\|^2 \\
 &\quad +\frac{\mu}{2}\beta(1-\beta)h\|v^{(k+1)}-v^{(k)}\|^2 -\frac{\mu}{2}h\|v^{(k+1)}-z^{(k)}\|^2 +\frac{\mu}{2}h\beta^2\|v^{(k+1)}-v^{(k)}\|^2 \\
 &\quad +\frac{L}{2}\|x^{(k+1)}-z^{(k)}\|^2 -\frac{\mu}{2}\|z^{(k)}-x^{(k)}\|^2 -\frac{\mu}{2}\|v^{(k+1)}-v^{(k)}\|^2.
 \end{aligned}$$

By considering

$$\begin{aligned}
 &\|x^{(k+1)}-z^{(k)}\|^2 \\
 &=h^2\|(1-\alpha)(v^{(k+1)}-z^{(k)})+\alpha(v^{(k+1)}-v^{(k)})\|^2 \\
 &=h^2((1-\alpha)\|v^{(k+1)}-z^{(k)}\|^2 +\alpha\|v^{(k+1)}-v^{(k)}\|^2 -\alpha(1-\alpha)\|v^{(k)}-z^{(k)}\|^2),
 \end{aligned}$$

we see that the coefficient of $\|v^{(k+1)}-v^{(k)}\|^2$ is

$$\begin{aligned}
 &-\frac{\mu}{2}(1-\beta)(1-\alpha)(1-\beta h)h -\frac{\mu}{2}(1-\beta)\alpha h(2+h(1-2\beta)) -\frac{\mu}{2}\beta\alpha(1+(1-\beta)h)h \\
 &+\frac{\mu}{2}\beta(1-\beta)h +\frac{\mu}{2}h\beta^2 -\frac{\mu}{2} +\frac{L}{2}((1-\beta)h+1)h^2\alpha \\
 &= -\frac{\mu}{2}((1-\beta)h+1)((\alpha-\beta)h+1) +\frac{L}{2}h^2((1-\beta)h+1)\alpha.
 \end{aligned}$$

Thus, we have a condition on h :

$$\frac{L}{\mu}\alpha h^2 -(\alpha-\beta)h -1 \leq 0. \tag{A9}$$

The coefficient of $\|v^{(k+1)}-z^{(k)}\|^2$ is

$$\begin{aligned}
 &-\frac{\mu}{2}(1-\beta)(1-\alpha)(1-\beta h)h -\frac{\mu}{2}(1-\beta)\alpha h^2 +\frac{\mu}{2}\beta\alpha(1+(1-\beta)h)h \\
 &-\mu(1-\beta)(1-\alpha)h^2 -\frac{\mu}{2}h +\frac{L}{2}((1-\beta)h+1)h^2(1-\alpha) \\
 &= -\frac{\mu}{2}h(2-\alpha-\beta)(1+(1-\beta)h) +\frac{L}{2}((1-\beta)h+1)h^2(1-\alpha),
 \end{aligned}$$

from which we have the condition

$$\frac{L}{\mu}(1 - \alpha)h - (2 - \alpha - \beta) \leq 0. \tag{A10}$$

The coefficient of $\|v^{(k)} - z^{(k)}\|^2$ is

$$\begin{aligned} & \frac{\mu}{2}(1 - \beta)(1 - \alpha)(1 - \beta h)h + \frac{\mu}{2}(1 - \beta)\alpha h^2 - \frac{\mu}{2}\beta\alpha(1 + (1 - \beta)h)h \\ & + \mu\beta\alpha h^2 - \frac{\mu}{2}\alpha^2 h^2 - \frac{L}{2}((1 - \beta)h + 1)h^2\alpha(1 - \alpha) + \frac{L}{2}\beta\alpha^2 h^3 \\ & = \frac{\mu}{2}h((\beta^2 - (1 - \alpha)(\beta - \alpha))h + 1 - \alpha - \beta) - \frac{L}{2}h^2((1 - \alpha - \beta)h + 1 - \alpha)\alpha, \end{aligned}$$

from which we have the condition

$$-\frac{L}{\mu}h((1 - \alpha - \beta)h + 1 - \alpha)\alpha + (\beta^2 + (1 - \alpha)(\beta - \alpha))h + 1 - \alpha - \beta \leq 0. \tag{A11}$$

Let $\kappa = L/\mu$. When $1 - \alpha = 1/(2\sqrt{\kappa})$ and $\beta = 1/2$, the condition (A10) becomes

$$h \leq \frac{2 - \alpha - \beta}{(1 - \alpha)\kappa} = \frac{1 + 1/\sqrt{\kappa}}{\sqrt{\kappa}}.$$

This is satisfied by $h \leq 1/\sqrt{\kappa}$. By the same choice of α and β , the condition (A9) becomes

$$\kappa \left(1 - \frac{1}{2\sqrt{\kappa}}\right)h^2 - \left(\frac{1}{2} - \frac{1}{2\sqrt{\kappa}}\right)h - 1 \leq 0.$$

When $h = 1/\sqrt{\kappa}$, the left-hand side becomes $-1/\sqrt{\kappa} + 1/(2\kappa)$, and the inequality holds. From the convexity, (A9) is satisfied for any $h \leq 1/\sqrt{\kappa}$. The third condition (A11) becomes

$$\begin{aligned} & -\kappa \left(\frac{1}{2\sqrt{\kappa}} - \frac{1}{2}\right) \left(1 - \frac{1}{2\sqrt{\kappa}}\right)h^2 \\ & + \left(-\kappa \frac{1}{2\sqrt{\kappa}} \left(1 - \frac{1}{2\sqrt{\kappa}}\right) + \frac{1}{4} + \frac{1}{2\sqrt{\kappa}} \left(\frac{1}{2\sqrt{\kappa}} - \frac{1}{2}\right)\right)h + \left(\frac{1}{2\sqrt{\kappa}} - \frac{1}{2}\right) \\ & = \left(\frac{1}{2\sqrt{\kappa}} - \frac{1}{2}\right) \left(-\left(1 - \frac{1}{2\sqrt{\kappa}}\right)\kappa h^2 + \left(\frac{1}{2\kappa} + 1\right)\sqrt{\kappa}h + 1\right). \end{aligned}$$

When $h = 1/\sqrt{\kappa}$, the second factor is nonnegative. By the concavity, the second factor is nonnegative for all $h \leq 1/\sqrt{\kappa}$. Since the first factor is nonpositive, (A11) is also satisfied by the above choices of parameters.

In summary, when $\alpha = 1 - 1/(2\sqrt{\kappa})$ and $\beta = 1/2$, we obtain the following convergence rate under the condition $h \leq 1/\sqrt{\kappa}$:

$$E^{(k)} \leq \left(\frac{1 - h\beta}{1 + h(1 - \beta)} \right)^k E^{(0)} = \left(\frac{1 - h/2}{1 + h/2} \right)^k E^{(0)}. \tag{A12}$$

This is better than that in the previous section.

However, the resulting scheme explicitly contains κ in the scheme (recall $\alpha = 1 - 1/(2\sqrt{\kappa})$), which is not desirable (we need to know the condition number of the objective function before we run the scheme.) Let us eliminate this below.

Letting $\alpha = 1 - h/2$ and $\beta = 1/2$, the left-hand side of the condition (A9) becomes

$$\frac{L}{\mu} \left(1 - \frac{h}{2} \right) h^2 - \left(\frac{1}{2} - \frac{h}{2} \right) h - 1 = \left(1 - \frac{h}{2} \right) (\kappa h^2 - h - 1).$$

This is nonpositive for $h \leq 1/\sqrt{\kappa}$. The left-hand side of the condition (A10) becomes

$$\kappa \frac{h^2}{2} - \frac{h}{2} - \frac{1}{2},$$

which is also nonpositive for $h \leq 1/\sqrt{\kappa}$. Finally, the left-hand side of the condition (A11) becomes

$$\begin{aligned} & -\kappa \left(\frac{h}{2} - \frac{1}{2} \right) \left(1 - \frac{h}{2} \right) h^2 + \left(-\kappa \frac{h}{2} \left(1 - \frac{h}{2} \right) + \frac{1}{4} - \frac{h}{2} \left(\frac{h}{2} - \frac{1}{2} \right) \right) h + \left(\frac{h}{2} - \frac{1}{2} \right) \\ & = -\kappa \frac{h^3}{2} \left(1 - \frac{h}{2} \right) - \frac{h^3}{4} + \frac{h^2}{4} + \frac{3h}{4} - \frac{1}{2} \\ & = \frac{1}{4} (2 - h) (\kappa h^3 - h^2 - h + 1), \end{aligned}$$

which is nonpositive for all $h \leq 2$. This can be seen by $\kappa h^3 - h^2 - h + 1 \geq h^3 - h^2 - h + 1 \geq (h - 1)^2(h + 1) \geq 0$ which holds for all $h \geq 0$.

Therefore, by setting $\alpha = 1 - h/2$, and $\beta = 1/2$, we have the rate (A12) for $h \leq 1/\sqrt{\kappa}$. By replacing h with $\sqrt{\mu}h$, we obtain the scheme corresponding to (8).

Acknowledgements The authors are thankful for the anonymous reviewers who gave the authors many insightful comments, which were so helpful to improve the paper.

Author contribution K.U. drafted the manuscript, conducted numerical experiments, and prepared figures. S.S. and T.M. reviewed the manuscript and revised it. All authors contributed to the conception and approved the final version for submission.

Funding Open Access funding provided by The University of Tokyo. The first author is supported by Japan Science and Technology Agency Support for Pioneering Research Initiated by the Next Generation (JST SPRING) (No. JPMJSP2108). The second author is supported by a Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (B) (No. 20H01822) and a JSPS Grant-in-Aid for Early-Career Scientists (No. 22K13955). The third author is supported by a JSPS Grant-in-Aid for

Scientific Research (B) (No. 20H01822 and No. 21H03452) and a JSPS Grant-in-Aid for Challenging Research (Exploratory) (No. 20K21786).

Availability of supporting data Not applicable

Declarations

Ethical approval Not applicable

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Su, W., Boyd, S., Candès, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In: *Advances in Neural Information Processing Systems*, vol. 27 (2014)
2. Wilson, A.C., Recht, B., Jordan, M.I.: A Lyapunov analysis of accelerated methods in optimization. *J. Mach. Learn. Res.* **22**(113), 1–34 (2021)
3. Suh, J.J., Roh, G., Ryu, E.K.: Continuous-time analysis of accelerated gradient methods via conservation laws in dilated coordinate systems. In: *Proceedings of the 39th International Conference on Machine Learning* (2022)
4. Krichene, W., Bayen, A., Bartlett, P.L.: Accelerated mirror descent in continuous and discrete time. In: *Advances in Neural Information Processing Systems*, vol. 28 (2015)
5. Attouch, H., Chbani, Z., Fadili, J., Riahi, H.: First-order optimization algorithms via inertial systems with Hessian driven damping. *Math. Program.* **193**(1, Ser. A), 113–155 (2022)
6. Kim, J., Yang, I.: Unifying Nesterov's accelerated gradient methods for convex and strongly convex objective functions. In: *Proceedings of the 40th International Conference on Machine Learning* (2023)
7. Ushiyama, K., Sato, S., Matsuo, T.: A unified discretization framework for differential equation approach with Lyapunov arguments for convex optimization. In: *Advances in Neural Information Processing Systems* (2023)
8. Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR.* **269**(3), 543–547 (1983)
9. Nesterov, Y.: *Lectures on convex optimization*. Springer, Cham (2018)
10. Gonzalez, O.: Time integration and discrete Hamiltonian systems. *J. Nonlinear Sci.* **6**(5), 449–467 (1996)
11. Itoh, T., Abe, K.: Hamiltonian-conserving discrete canonical equations based on variational difference quotients. *J. Comput. Phys.* **76**(1), 85–102 (1988)
12. Quispel, G.R.W., McLaren, D.I.: A new class of energy-preserving numerical integration methods. *J. Phys. A.* **41**(4), 045206–7 (2008)
13. Mouce, C., Taylor, A., Bach, F.: A systematic approach to Lyapunov analyses of continuous-time models in convex optimization. *SIAM J. Optim.* **33**(3), 1558–1586 (2023)
14. Hairer, E., Lubich, C., Wanner, G.: *Geometric numerical integration*, 2nd edn. Springer, Berlin (2006)
15. Van Scoy, B., Freeman, R.A., Lynch, K.M.: The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Syst. Lett.* **2**(1), 49–54 (2018)

16. Hairer, E., Nørsett, S.P., Wanner, G.: Solving ordinary differential equations. I, 2nd edn. Springer, Berlin (1993). Nonstiff problems

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.