

---

# Time-Reversed Dissipation Induces Duality Between Minimizing Gradient Norm and Function Value

---

**Jaeyeon Kim**  
Seoul National University  
k jy011102@snu.ac.kr

**Asuman Ozdaglar**  
MIT EECS  
asuman@mit.edu

**Chanwoo Park**  
MIT EECS  
cpark97@mit.edu

**Ernest K. Ryu**  
Seoul National University  
ernestryu@snu.ac.kr

## Abstract

In convex optimization, first-order optimization methods efficiently minimizing function values have been a central subject study since Nesterov’s seminal work of 1983. Recently, however, Kim and Fessler’s OGM-G and Lee et al.’s FISTA-G have been presented as alternatives that efficiently minimize the gradient magnitude instead. In this paper, we present H-duality, which represents a surprising one-to-one correspondence between methods efficiently minimizing function values and methods efficiently minimizing gradient magnitude. In continuous-time formulations, H-duality corresponds to reversing the time dependence of the dissipation/friction term. To the best of our knowledge, H-duality is different from Lagrange/Fenchel duality and is distinct from any previously known duality or symmetry relations. Using H-duality, we obtain a clearer understanding of the symmetry between Nesterov’s method and OGM-G, derive a new class of methods efficiently reducing gradient magnitudes of smooth convex functions, and find a new composite minimization method that is simpler and faster than FISTA-G.

## 1 Introduction

Since Nesterov’s seminal work of 1983 [37], accelerated first-order optimization methods that efficiently reduce *function values* have been central to the theory and practice of large-scale optimization and machine learning. In 2012, however, Nesterov initiated the study of first-order methods that efficiently reduce *gradient magnitudes* of convex functions [41]. In convex optimization, making the function value exactly optimal is equivalent to making the gradient exactly zero, but reducing the function-value suboptimality below a threshold is not equivalent to reducing the gradient magnitude below a threshold. This line of research showed that accelerated methods for reducing function values, such as Nesterov’s FGM [37], the more modern OGM [26], and the accelerated composite optimization method FISTA [11] are not optimal for reducing gradient magnitude, and new optimal alternatives, such as OGM-G [29] and FISTA-G [31], were presented.

These new accelerated methods for reducing gradient magnitudes are understood far less than those for minimizing function values. However, an interesting observation of symmetry, described in Section 2, was made between these two types of methods, and it was conjectured that this symmetry might be a key to understanding the acceleration mechanism for efficiently reducing gradient magnitude.

**Contribution.** We present a surprising one-to-one correspondence between methods efficiently minimizing function values and methods efficiently minimizing gradient magnitude. We call this correspondence H-duality and formally establish a duality theory in both discrete- and continuous-time dynamics. Using H-duality, we obtain a clearer understanding of the symmetry between FGM/OGM and OGM-G, derive a new class of methods efficiently reducing gradient magnitudes, and find a new composite minimization method that is simpler and faster than FISTA-G, the prior state-of-the-art in efficiently reducing gradient magnitude in the composite minimization setup.

## 1.1 Preliminaries and Notation

Given  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ , write  $f_\star = \inf_{x \in \mathbb{R}^d} f(x) \in (-\infty, \infty)$  for the minimum value and  $x_\star \in \operatorname{argmin}_{x \in \mathbb{R}^d} f(x) \in \mathbb{R}^d$  for a minimizer, if one exists. Throughout this paper, we assume  $f_\star \neq -\infty$ , but we do not always assume a minimizer  $x_\star$  exists. Given a differentiable  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  and a pre-specified value of  $L > 0$ , we define the notation

$$\begin{aligned} [x, y] &:= f(y) - f(x) + \langle \nabla f(y), x - y \rangle \\ \llbracket x, y \rrbracket &:= f(y) - f(x) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \\ \llbracket x, \star \rrbracket &:= f_\star - f(x) + \frac{1}{2L} \|\nabla f(x)\|^2 \end{aligned}$$

for  $x, y \in \mathbb{R}^d$ . A differentiable function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  is convex if the convexity inequality  $[x, y] \leq 0$  holds for all  $x, y \in \mathbb{R}^d$ . For  $L > 0$ , a function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  is  $L$ -smooth convex if it is differentiable and the cocoercivity inequality  $\llbracket x, y \rrbracket \leq 0$  holds for all  $x, y \in \mathbb{R}^d$  [42]. If  $f$  has a minimizer  $x_\star$ , then  $\llbracket x, \star \rrbracket = \llbracket x, x_\star \rrbracket$ , but the notation  $\llbracket x, \star \rrbracket$  is well defined even when a minimizer  $x_\star$  does not exist. If  $f$  is  $L$ -smooth convex, then  $\llbracket x, \star \rrbracket \leq 0$  holds for all  $x \in \mathbb{R}^d$  [42].

Throughout this paper, we consider the duality between the following two problems.

- (P1) Efficiently reduce  $f(x_N) - f_\star$  assuming  $x_\star$  exists and  $\|x_0 - x_\star\| \leq R$ .
- (P2) Efficiently reduce  $\frac{1}{2L} \|\nabla f(y_N)\|^2$  assuming  $f_\star > -\infty$  and  $f(y_0) - f_\star \leq R$ .

Here,  $R \in (0, \infty)$  is a parameter,  $x_0$  and  $y_0$  denote initial points of methods for (P1) and (P2), and  $x_N$  and  $y_N$  denote outputs of methods for (P1) and (P2).

Finally, the standard gradient descent (GD) with stepsize  $h$  is

$$x_{i+1} = x_i - \frac{h}{L} \nabla f(x_i), \quad i = 0, 1, \dots \quad (\text{GD})$$

## 1.2 Prior works

Classically, the goal of optimization methods is to reduce the function value efficiently. In the smooth convex setup, Nesterov's fast gradient method (FGM) [37] achieves an accelerated  $\mathcal{O}(1/N^2)$ -rate, and the optimized gradient method (OGM) [26] improves this rate by a factor of 2, which is, in fact, exactly optimal [18].

On the other hand, Nesterov initiated the study of methods for reducing the gradient magnitude of convex functions [41] as such methods help us understand non-convex optimization better and design faster non-convex machine learning methods. For smooth convex functions, (GD) achieves a  $\mathcal{O}((f(x_0) - f_\star)/N)$ -rate on the squared gradient magnitude [34, Proposition 3.3.1], while (OGM-G) achieves an accelerated  $\mathcal{O}((f(x_0) - f_\star)/N^2)$ -rate [29], which matches a lower bound and is therefore optimal [35, 36]. Interestingly, (OGM) and (OGM-G) exhibit an interesting hint of symmetry, as we detail in Section 2, and the goal of this work is to derive a more general duality principle from this observation.

In the composite optimization setup, iterative shrinkage-thresholding algorithm (ISTA) [13, 45, 16, 14] achieves a  $\mathcal{O}(\|x_0 - x_\star\|^2/N)$ -rate on function-value suboptimality, while the fast iterative shrinkage-thresholding algorithm (FISTA) [11] achieves an accelerated  $\mathcal{O}(\|x_0 - x_\star\|/N^2)$ -rate. On the squared gradient mapping norm, FISTA-G achieves  $\mathcal{O}((F(x_0) - F_\star)/N^2)$ -rate [31], which is optimal [35, 36]. Analysis of an accelerated method often uses the estimate sequence technique [38, 8, 39, 9, 40, 32] or a Lyapunov analysis [37, 11, 50, 10, 53, 1, 5, 6, 7, 43]. In this work, we focus on the Lyapunov analysis technique, as it is simpler and more amenable to a continuous-time view.

The notion of duality is fundamental in many branches of mathematics, including optimization. Lagrange duality [46, 47, 12], Wolfe duality [57, 15, 49, 33], and Fenchel–Rockafeller duality [20, 46] are related (arguably equivalent) notions that consider a pairing of primal and dual optimization problems. The recent gauge duality [21, 22, 3, 58] and radial duality [24, 23] are alternative notions of duality for optimization problems. Attouch–Théra duality [4, 48] generalizes Fenchel–Rockafeller to the setup of monotone inclusion problems. In this work, we present H-duality, which is a notion of duality for optimization *algorithms*, and it is, to the best of our knowledge, distinct from any previously known duality or symmetry relations.

## 2 H-duality

In this section, we will introduce H-duality, state the main H-duality theorem, and provide applications. Let  $N \geq 1$  be a pre-specified iteration count. Let  $\{h_{k,i}\}_{0 \leq i < k \leq N}$  be an array of (scalar) stepsizes and identify it with a lower triangular matrix  $H \in \mathbb{R}^{N \times N}$  via  $H_{k+1,i+1} = h_{k+1,i}$  if  $0 \leq i \leq k \leq N-1$  and  $H_{k,i} = 0$  otherwise. An  $N$ -step Fixed Step First Order Method (FSFOM) with  $H$  is

$$x_{k+1} = x_k - \frac{1}{L} \sum_{i=0}^k h_{k+1,i} \nabla f(x_i), \quad \forall k = 0, \dots, N-1 \quad (1)$$

for any initial point  $x_0 \in \mathbb{R}^d$  and differentiable  $f$ . For  $H \in \mathbb{R}^{N \times N}$ , define its *anti-transpose*  $H^A \in \mathbb{R}^{N \times N}$  with  $H_{i,j}^A = H_{N-j+1, N-i+1}$  for  $i, j = 1, \dots, N$ . We call [FSFOM with  $H^A$ ] the **H-dual** of [FSFOM with  $H$ ].

### 2.1 Symmetry between OGM and OGM-G

Let  $f$  be an  $L$ -smooth convex function. Define the notation  $z^+ = z - \frac{1}{L} \nabla f(z)$  for  $z \in \mathbb{R}^d$ . The accelerated methods OGM [19, 26] and OGM-G [29] are

$$x_{k+1} = x_k^+ + \frac{\theta_k - 1}{\theta_{k+1}} (x_k^+ - x_{k-1}^+) + \frac{\theta_k}{\theta_{k+1}} (x_k^+ - x_k) \quad (\text{OGM})$$

$$y_{k+1} = y_k^+ + \frac{(\theta_{N-k} - 1)(2\theta_{N-k-1} - 1)}{\theta_{N-k}(2\theta_{N-k} - 1)} (y_k^+ - y_{k-1}^+) + \frac{2\theta_{N-k-1} - 1}{2\theta_{N-k} - 1} (y_k^+ - y_k) \quad (\text{OGM-G})$$

for  $k = 0, \dots, N-1$ , where  $\{\theta_i\}_{i=0}^N$  are defined as  $\theta_0 = 1$ ,  $\theta_{i+1}^2 - \theta_i^2 = \theta_i^2$  for  $0 \leq i \leq N-2$ , and  $\theta_N^2 - \theta_{N-1}^2 = 2\theta_{N-1}^2$ .<sup>1</sup> (OGM) and (OGM-G) are two representative accelerated methods for the setups (P1) and (P2), respectively. As a surface-level symmetry, the methods both access the  $\{\theta_i\}_{i=0}^N$  sequence, but (OGM-G) does so in a reversed ordering [29]. There turns out to be a deeper-level symmetry: (OGM) and (OGM-G) are H-duals of each other, i.e.,  $H_{\text{OGM}}^A = H_{\text{OGM-G}}$ . The proof structures of (OGM) and (OGM-G) also exhibit symmetry. We can analyze (OGM) with the Lyapunov function

$$\mathcal{U}_k = \frac{L}{2} \|x_0 - x_\star\|^2 + \sum_{i=0}^{k-1} u_i \llbracket x_i, x_{i+1} \rrbracket + \sum_{i=0}^k (u_i - u_{i-1}) \llbracket x_\star, x_i \rrbracket \quad (2)$$

for  $-1 \leq k \leq N$  with  $\{u_i\}_{i=0}^N = (2\theta_0^2, \dots, 2\theta_{N-1}^2, \theta_N^2)$  and  $u_{-1} = 0$ . Since  $\llbracket \cdot, \cdot \rrbracket \leq 0$  and  $\{u_i\}_{i=0}^N$  is a positive monotonically increasing sequence,  $\{\mathcal{U}_k\}_{k=-1}^N$  is dissipative, i.e.,  $\mathcal{U}_N \leq \mathcal{U}_{N-1} \leq \dots \leq \mathcal{U}_0 \leq \mathcal{U}_{-1}$ . So

$$\theta_N^2 (f(x_N) - f_\star) \leq \theta_N^2 (f(x_N) - f_\star) + \frac{L}{2} \|x^\star - x_0 + z\|^2 \stackrel{(\bullet)}{=} \mathcal{U}_N \leq \mathcal{U}_{-1} = \frac{L \|x_0 - x^\star\|^2}{2},$$

where  $z = \sum_{i=0}^N \frac{u_i - u_{i-1}}{L} \nabla f(x_i)$ . The justification of  $(\bullet)$  is the main technical challenge of this analysis, and it is provided in Appendix B.2. Dividing both sides by  $\theta_N^2$ , we conclude the rate

$$f(x_N) - f_\star \leq \frac{1}{\theta_N^2} \frac{L}{2} \|x_0 - x^\star\|^2.$$

Likewise, we can analyze (OGM-G) with the Lyapunov function

$$\mathcal{V}_k = v_0 (f(y_0) - f_\star + \llbracket y_N, \star \rrbracket) + \sum_{i=0}^{k-1} v_{i+1} \llbracket y_i, y_{i+1} \rrbracket + \sum_{i=0}^{k-1} (v_{i+1} - v_i) \llbracket y_N, y_i \rrbracket \quad (3)$$

for  $0 \leq k \leq N$  with  $\{v_i\}_{i=0}^N = \left(\frac{1}{\theta_N^2}, \frac{1}{2\theta_{N-1}^2}, \dots, \frac{1}{2\theta_0^2}\right)$ . Similarly,  $\{\mathcal{V}_k\}_{k=0}^N$  is dissipative, so

$$\frac{1}{2L} \|\nabla f(y_N)\|^2 \stackrel{(\circ)}{=} \mathcal{V}_N \leq \mathcal{V}_0 = \frac{1}{\theta_N^2} (f(y_0) - f_\star) + \frac{1}{\theta_N^2} \llbracket y_N, \star \rrbracket \leq \frac{1}{\theta_N^2} (f(y_0) - f_\star).$$

<sup>1</sup>Throughout this paper, we use the convention of denoting iterates of a given ‘‘primal’’ FSFOM as  $x_k$  while denoting the iterates of the H-dual FSFOM as  $y_k$ .

Again, the justification of  $(\circ)$  is the main technical challenge of this analysis, and it is provided in Appendix B.2. The crucial observations are **(i)**  $u_i = 1/v_{N-i}$  for  $0 \leq i \leq N$  and **(ii)** the convergence rates share the identical factor  $1/\theta_N^2 = 1/u_N = v_0$ . Interestingly, a similar symmetry relation holds between method pairs [(OBL-F<sub>b</sub>), (OBL-G<sub>b</sub>)] [44] and [(GD), (GD)], which we discuss later in Section 2.4.

## 2.2 H-duality theorem

The symmetry observed in Section 2.1 is, in fact, not a coincidence. Suppose we have  $N$ -step FSFOMs with  $H$  and  $H^A$ . We denote their iterates as  $\{x_i\}_{i=0}^N$  and  $\{y_i\}_{i=0}^N$ . For clarity, note that  $\{u_i\}_{i=0}^N$  are free variables and can be appropriately chosen for the convergence rate analysis. For the FSFOM with  $H$ , define  $\{\mathcal{U}_k\}_{k=-1}^N$  with the general form (2) with  $u_{-1} = 0$ . If  $0 = u_{-1} \leq u_0 \leq u_1 \leq \dots \leq u_N$ , then  $\{\mathcal{U}_k\}_{k=-1}^N$  is monotonically nonincreasing (dissipative). Assume we can show

$$u_N(f(x_N) - f_\star) \leq \mathcal{U}_N \quad (\forall x_0, x_\star, \nabla f(x_0), \dots, \nabla f(x_N) \in \mathbb{R}^d). \quad (\text{C1})$$

To clarify, since  $\{x_i\}_{i=0}^N$  lies within  $\text{span}\{x_0, \nabla f(x_0), \dots, \nabla f(x_N)\}$ , the  $\mathcal{U}_N$  depends on  $(x_0, x_\star, \{\nabla f(x_i)\}_{i=0}^N, \{u_i\}_{i=0}^N, H)$ . If (C1) holds, the FSFOM with  $H$  exhibits the convergence rate

$$u_N(f(x_N) - f_\star) \leq \mathcal{U}_N \leq \dots \leq \mathcal{U}_{-1} = \frac{L}{2} \|x_0 - x_\star\|^2. \quad (4)$$

For the FSFOM with  $H^A$ , define  $\{\mathcal{V}_k\}_{k=0}^N$  with the general form (3). Also, note that  $\{v_i\}_{i=0}^N$  are free variables and can be appropriately chosen for the convergence rate analysis. If  $0 \leq v_0 \leq v_1 \leq \dots \leq v_N$ , then  $\{\mathcal{V}_k\}_{k=0}^N$  is monotonically nonincreasing (dissipative). Assume we can show

$$\frac{1}{2L} \|\nabla f(y_N)\|^2 \leq \mathcal{V}_N \quad (\forall y_0, \nabla f(y_0), \dots, \nabla f(y_N) \in \mathbb{R}^d, f_\star \in \mathbb{R}). \quad (\text{C2})$$

To clarify, since  $\{y_i\}_{i=0}^N$  lies within  $\text{span}\{y_0, \nabla f(y_0), \dots, \nabla f(y_N)\}$ , the  $\mathcal{V}_N$  depends on  $(y_0, \{\nabla f(y_i)\}_{i=0}^N, f_\star, \{v_i\}_{i=0}^N, H^A)$ . If (C2) holds, the FSFOM with  $H^A$  exhibits the convergence rate

$$\frac{1}{2L} \|\nabla f(y_N)\|^2 \leq \mathcal{V}_N \leq \dots \leq \mathcal{V}_0 = v_0(f(y_0) - f_\star) + v_0 \llbracket y_N, \star \rrbracket \leq v_0(f(y_0) - f_\star). \quad (5)$$

We now state our main H-duality theorem, which establishes a correspondence between the two types of bounds for the FSFOMs induced by  $H$  and  $H^A$ .

**Theorem 1.** Consider sequences of positive real numbers  $\{u_i\}_{i=0}^N$  and  $\{v_i\}_{i=0}^N$  related through  $v_i = \frac{1}{u_{N-i}}$  for  $i = 0, \dots, N$ . Let  $H \in \mathbb{R}^{N \times N}$  be lower triangular. Then,

$$[(\text{C1}) \text{ is satisfied with } \{u_i\}_{i=0}^N \text{ and } H] \quad \Leftrightarrow \quad [(\text{C2}) \text{ is satisfied with } \{v_i\}_{i=0}^N \text{ and } H^A].$$

Theorem 1 provides a sufficient condition that ensures an FSFOM with  $H$  with a convergence guarantee on  $(f(x_N) - f_\star)$  can be H-dualized to obtain an FSFOM with  $H^A$  with a convergence guarantee on  $\|\nabla f(y_N)\|^2$ . To the best of our knowledge, this is the first result establishing a symmetrical relationship between (P1) and (P2). Section 2.3 provides a proof outline of Theorem 1.

## 2.3 Proof outline of Theorem 1

Define

$$\begin{aligned} \mathbf{U} &:= \mathcal{U}_N - u_N(f(x_N) - f_\star) - \frac{L}{2} \left\| x_\star - x_0 + \frac{1}{L} \sum_{i=0}^N (u_i - u_{i-1}) \nabla f(x_i) \right\|^2 \\ \mathbf{V} &:= \mathcal{V}_N - \frac{1}{2} \|\nabla f(y_N)\|^2. \end{aligned}$$

Expanding  $\mathbf{U}$  and  $\mathbf{V}$  reveals that all function value terms are eliminated and only quadratic terms of  $\{\nabla f(x_i)\}_{i=0}^N$  and  $\{\nabla f(y_i)\}_{i=0}^N$  remain. Now, (C1) and (C2) are equivalent to the conditions

$$\left[ \mathbf{U} \geq 0, \quad \forall (\nabla f(x_0), \dots, \nabla f(x_N) \in \mathbb{R}^d) \right], \quad \left[ \mathbf{V} \geq 0 \quad \forall (\nabla f(y_0), \dots, \nabla f(y_N) \in \mathbb{R}^d) \right],$$

respectively. Next, define  $g_x = [|\nabla f(x_0)|, |\nabla f(x_1)|, \dots, |\nabla f(x_N)|] \in \mathbb{R}^{d \times (N+1)}$  and  $g_y = [|\nabla f(y_0)|, |\nabla f(y_1)|, \dots, |\nabla f(y_N)|] \in \mathbb{R}^{d \times (N+1)}$ . We show that there is  $\mathcal{S}(H, u)$  and  $\mathcal{T}(H^A, v) \in \mathbb{S}^{N+1}$  such that

$$\mathbf{U} = \text{Tr}(g_x \mathcal{S}(H, u) g_x^\top), \quad \mathbf{V} = \text{Tr}(g_y \mathcal{T}(H^A, v) g_y^\top).$$

Next, we find an explicit invertible matrix  $M(u) \in \mathbb{R}^{(N+1) \times (N+1)}$  such that  $\mathcal{S}(H, u) = \mathcal{M}(u)^\top \mathcal{T}(H^A, v) \mathcal{M}(u)$ . Therefore,

$$\text{Tr}(g_x \mathcal{S}(H, u) g_x^\top) = \text{Tr}(g_y \mathcal{T}(H^A, v) g_y^\top)$$

with  $g_y = g_x \mathcal{M}(u)^\top$  and we conclude the proof. This technique of considering the quadratic forms of Lyapunov functions as a trace of matrices is inspired by the ideas from the Performance Estimation Problem (PEP) literature [19, 55]. The full proof is given in Appendix A.

## 2.4 Verifying conditions for H-duality theorem

In this section, we illustrate how to verify conditions (C1) and (C2) through examples. Detailed calculations are deferred to Appendices B.1 and B.2.

**Example 1.** For (OGM) and (OGM-G), the choice

$$\{u_i\}_{i=0}^N = (2\theta_0^2, \dots, 2\theta_{N-1}^2, \theta_N^2), \quad \{v_i\}_{i=0}^N = \left( \frac{1}{\theta_N^2}, \frac{1}{2\theta_{N-1}^2}, \dots, \frac{1}{2\theta_0^2} \right)$$

leads to

$$\mathbf{U} = 0, \quad \mathbf{V} = 0.$$

Therefore, (C1) and (C2) hold.

**Example 2.** Again, define  $z^+ = z - \frac{1}{L} \nabla f(z)$  for  $z \in \mathbb{R}^d$ . Consider the FSFSOMs [44]

$$\begin{aligned} x_{k+1} &= x_k^+ + \frac{k}{k+3} (x_k^+ - x_{k-1}^+) + \frac{k}{k+3} (x_k^+ - x_k) \quad k = 0, \dots, N-2 \\ x_N &= x_{N-1}^+ + \frac{N-1}{2(\gamma+1)} (x_{N-1}^+ - x_{N-2}^+) + \frac{N-1}{2(\gamma+1)} (x_{N-1}^+ - x_{N-1}) \end{aligned} \quad (\text{OBL-F}_b)$$

and

$$\begin{aligned} y_1 &= y_0^+ + \frac{N-1}{2(\gamma+1)} (y_0^+ - y_{-1}^+) + \frac{N-1}{2(\gamma+1)} (y_0^+ - y_0) \\ y_{k+1} &= y_k^+ + \frac{N-k-1}{N-k+2} (y_k^+ - y_{k-1}^+) + \frac{N-k-1}{N-k+2} (y_k^+ - y_k) \quad k = 1, \dots, N-1 \end{aligned} \quad (\text{OBL-G}_b)$$

where  $y_{-1}^+ = y_0$ ,  $x_{-1}^+ = x_0$  and  $\gamma = \sqrt{N(N+1)}/2$ . It turns out that (OBL-F<sub>b</sub>) and (OBL-G<sub>b</sub>) are H-duals of each other. The choice

$$\{u_i\}_{i=0}^N = \left( \frac{1.2}{2}, \dots, \frac{N(N+1)}{2}, \gamma^2 + \gamma \right), \quad \{v_i\}_{i=0}^N = \left( \frac{1}{\gamma^2 + \gamma}, \frac{2}{N(N+1)}, \dots, \frac{2}{1.2} \right)$$

leads to

$$\mathbf{U} = \sum_{i=0}^N \frac{u_i - u_{i-1}}{2L} \|\nabla f(x_i)\|^2, \quad \mathbf{V} = \frac{v_0}{2L} \|\nabla f(y_N)\|^2 + \sum_{i=0}^{N-1} \frac{v_{i+1} - v_i}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2$$

where  $u_{-1} = 0$ . Since  $\mathbf{U}$  and  $\mathbf{V}$  are expressed as a sum of squares, (C1) and (C2) hold.

**Example 3.** Interestingly, (GD) is a self-dual FSFOM in the H-dual sense. For the case  $h = 1$ , the choice

$$\{u_i\}_{i=0}^N = \left( \dots, \frac{(2N+1)(i+1)}{2N-i}, \dots, 2N+1 \right), \quad \{v_i\}_{i=0}^N = \left( \frac{1}{2N+1}, \dots, \frac{N+i}{(2N+1)(N-i+1)}, \dots \right)$$

leads to

$$\mathbf{U} = \sum_{0 \leq i, j \leq N} \frac{s_{ij}}{L} \langle \nabla f(x_i), \nabla f(x_j) \rangle, \quad \mathbf{V} = \sum_{0 \leq i, j \leq N} \frac{t_{ij}}{L} \langle \nabla f(y_i), \nabla f(y_j) \rangle$$

for some  $\{s_{ij}\}$  and  $\{t_{ij}\}$  stated precisely in Appendix B.2.  $\mathbf{V} \geq 0$  can be established by showing that the  $\{t_{ij}\}$  forms a diagonally dominant and hence positive semidefinite matrix [29].  $\mathbf{U} \geq 0$  can be established with a more elaborate argument [19], but that is not necessary;  $\mathbf{V} \geq 0$  implies (C2), and, by Theorem 1, this implies (C1).

## 2.5 Applications of the H-duality theorem

**A family of gradient reduction methods.** Parameterized families of accelerated FSFOMs for reducing function values have been presented throughout the extensive prior literature. Such families generalize Nesterov's method and elucidate the essential algorithmic component that enables acceleration. For reducing gradient magnitude, however, there are only four accelerated FSFOMs (OGM-G), (OBL-G<sub>b</sub>), and M-OGM-G [60], and [17, Lemma 2.6]. Here, we construct a simple parameterized family of accelerated FSFOMs for reducing gradient magnitude by H-dualizing an accelerated FSFOM family for reducing function values.

Let  $\{t_i\}_{i=0}^N$  and  $\{T_i\}_{i=0}^N$  be sequences positive real numbers satisfying  $t_i^2 \leq 2T_i = 2\sum_{j=0}^i t_j$  for  $0 \leq i \leq N-1$  and  $t_N^2 \leq T_N = \sum_{j=0}^N t_j$ . Consider a family of FSFOMs

$$x_{k+1} = x_k^+ + \frac{(T_k - t_k)t_{k+1}}{t_k T_{k+1}} (x_k^+ - x_{k-1}^+) + \frac{(t_k^2 - T_k)t_{k+1}}{t_k T_{k+1}} (x_k^+ - x_k) \quad (6)$$

for  $k = 0, 1, \dots, N-1$ , where  $x_{-1}^+ = x_0$ . This family coincides with the GOGM of [28], and it exhibits the rate [28, Theorem 5]

$$f(x_N) - f_* \leq \frac{1}{T_N} \frac{L}{2} \|x_0 - x_*\|^2,$$

which can be established from (2) with  $u_i = T_i$  for  $0 \leq i \leq N$ .

**Corollary 1.** The H-dual of (6) is

$$y_{k+1} = y_k^+ + \frac{T_{N-k-1}(t_{N-k-1} - 1)}{T_{N-k}(t_{N-k} - 1)} (y_k^+ - y_{k-1}^+) + \frac{(t_{N-k}^2 - T_{N-k})(t_{N-k-1} - 1)}{T_{N-k}(t_{N-k} - 1)} (y_k^+ - y_k)$$

for  $k = 0, \dots, N-1$ , where  $y_{-1}^+ = y_0$ , and it exhibits the rate

$$\frac{1}{2L} \|\nabla f(y_N)\|^2 \leq \frac{1}{T_N} (f(y_0) - f_*).$$

*Proof outline.* By Theorem 1, (C2) holds with  $v_i = 1/T_{N-i}$  for  $0 \leq i \leq N$ . We then use (5).  $\square$

When  $T_i = t_i^2$  for  $0 \leq i \leq N$ , the FSFOM (6) reduces to Nesterov's FGM [37] and its H-dual is, to the best of our knowledge, a new method without a name. If  $t_i^2 = 2T_i$  for  $0 \leq i \leq N-1$  and  $t_N^2 = T_N$ , (6) reduces to (OGM) and its H-dual is (OGM-G). If  $t_i = i+1$  for  $0 \leq i \leq N-1$  and  $t_N = \sqrt{N(N+1)}/2$ , (6) reduces to (OBL-F<sub>b</sub>) and its H-dual is (OBL-G<sub>b</sub>).

**Gradient magnitude rate of (GD).** For gradient descent (GD) with stepsize  $h$ , the  $H$  matrix is the identity matrix scaled by  $h$ , and the H-dual is (GD) itself, i.e., (GD) is self-dual. For  $0 < h \leq 1$ , the rate  $f(x_N) - f_* \leq \frac{1}{2Nh+1} \frac{L}{2} \|x_0 - x_*\|^2$ , originally due to [19], can be established from (2) with  $\{u_i\}_{i=0}^N = \left( \dots, \frac{(2Nh+1)(i+1)}{2^{N-i}}, \dots, 2Nh+1 \right)$ . Applying Theorem 1 leads to the following.

**Corollary 2.** Consider (GD) with  $0 < h \leq 1$  applied to an  $L$ -smooth convex  $f$ . For  $N \geq 1$ ,

$$\frac{1}{2L} \|\nabla f(x_N)\|^2 \leq \min \left( \frac{f(x_0) - f_*}{2Nh+1}, \frac{L \|x_0 - x_*\|^2}{2(2\lfloor \frac{N}{2} \rfloor h + 1)(2\lceil \frac{N}{2} \rceil h + 1)} \right).$$

To the best of our knowledge, Corollary 2 is the tightest rate on gradient magnitude for (GD) for the general step size  $0 < h < 1$ , and it matches [53, Theorem 3] for  $h = 1$ .

**Resolving conjectures of  $\mathcal{A}^*$ -optimality of (OGM-G) and (OBL-F<sub>b</sub>).** The prior work of [44] defines the notion of  $\mathcal{A}^*$ -optimality, a certain restricted sense of optimality of FSFOMs, and shows that (OGM) and (OBL-F<sub>b</sub>) are  $\mathcal{A}^*$ -optimal under a certain set of relaxed inequalities. On the other hand,  $\mathcal{A}^*$ -optimality of (OGM-G) and (OBL-G<sub>b</sub>) are presented as conjectures. Combining Theorem 1 and the  $\mathcal{A}^*$ -optimality of (OGM) and (OBL-F<sub>b</sub>) resolves these conjectures; (OGM-G) and (OBL-G<sub>b</sub>) are  $\mathcal{A}^*$ -optimal.

## 2.6 Intuition behind energy functions: Lagrangian Formulation

One might ask where the energy functions (2) and (3) came from. In this section, we provide an intuitive explanation of these energy functions using the QCQP and its Lagrangian. Consider an FSFOM (1) given with a lower triangular matrix  $H \in \mathbb{R}^{N \times N}$  with function  $f$ , resulting in the sequence  $\{x_i\}_{i=0}^N$ . To analyze the convergence rate of the function value, we formulate the following optimization problem:

$$\begin{aligned} & \sup_f f(x_N) - f_\star \\ & \text{subject to } f: \mathbb{R}^n \rightarrow \mathbb{R} \text{ is } L\text{-smooth convex, } \|x_0 - x_\star\|^2 \leq R^2. \end{aligned}$$

However, this optimization problem is not solvable, as  $f$  is a functional variable. To address this, [19, 55] demonstrated its equivalence to a QCQP:

$$\begin{aligned} & \sup_f f(x_N) - f_\star \\ & \text{subject to } \llbracket x_i, x_j \rrbracket \leq 0 \quad [i, j] \in [-1, \dots, N]^2, \quad \|x_0 - x_\star\|^2 \leq R^2 \end{aligned}$$

where  $x_{-1} := x_\star$ . To clarify, the optimization variables are  $\{\nabla f(x_i), f(x_i)\}_{i=0}^N$ ,  $f_\star$ ,  $x_0$ , and  $x_\star$  since  $\{x_i\}_{i=0}^N$  lies within  $\text{span}\{x_0, \nabla f(x_0), \dots, \nabla f(x_N)\}$ . We consider a relaxed optimization problem as [26]:

$$\begin{aligned} & \sup_f f(x_N) - f_\star \\ & \text{subject to } \llbracket x_i, x_{i+1} \rrbracket \leq 0, \quad i = 0, 1, \dots, N-1, \quad \llbracket x_\star, x_i \rrbracket \leq 0, \quad i = 0, 1, \dots, N, \\ & \quad \quad \quad \|x_0 - x_\star\|^2 \leq R^2 \end{aligned}$$

Now, consider the Lagrangian function and the convex dual.

$$\mathcal{L}_1(f, \{a_i\}_{i=0}^{N-1}, \{b_i\}_{i=0}^N, \alpha) = -f(x_N) + f_\star + \sum_{i=0}^{N-1} a_i \llbracket x_i, x_{i+1} \rrbracket + \sum_{i=0}^N b_i \llbracket x_\star, x_i \rrbracket + \alpha \|x_0 - x_\star\|^2 - \alpha R^2$$

where  $\{a_i\}_{i=0}^{N-1}$ ,  $\{b_i\}_{i=0}^N$ , and  $\alpha$  are dual variables which are nonnegative. Considering  $a_{-1} = 0$  and  $a_N = 1$ , the infimum of  $\mathcal{L}_1$  equals  $-\infty$  unless  $b_i = a_i - a_{i-1}$  for  $0 \leq i \leq N$ . Therefore, by introducing  $u_N = \frac{L}{2\alpha}$  and  $u_i = a_i u_N$  for  $0 \leq i \leq N$ , the convex dual problem can be simplified as follows:

$$\begin{aligned} & \inf_{\{u_i\}_{i=0}^N} -\frac{LR^2}{2u_N} \\ & \text{s.t. } \inf_{x_0, x_\star, \{\nabla f(x_i)\}_{i=0}^N} -u_N(f(x_N) - f_\star) + \frac{L}{2} \|x_0 - x_\star\|^2 + \sum_{i=0}^{N-1} u_i \llbracket x_i, x_{i+1} \rrbracket + \sum_{i=0}^N (u_i - u_{i-1}) \llbracket x_\star, x_i \rrbracket \geq 0 \\ & \quad \quad \quad \{u_i\}_{i=0}^N \text{ are nonnegative and nondecreasing.} \end{aligned}$$

If the above two constraints holds for  $\{u_i\}_{i=0}^N$ , we have  $f(x_N) - f_\star \leq \frac{L}{2u_N} R^2$ . This understanding motivates the introduction of (2) and (C1). We can perform a similar analysis on the gradient norm minimization problem with the relaxed optimization problem as follows:

$$\begin{aligned} & \sup_f \frac{1}{2L} \|\nabla f(y_N)\|^2 \\ & \text{subject to } \llbracket y_i, y_{i+1} \rrbracket \leq 0, \quad i = 0, 1, \dots, N-1, \quad \llbracket y_N, y_i \rrbracket \leq 0, \quad i = 0, 1, \dots, N-1, \\ & \quad \quad \quad \llbracket y_N, \star \rrbracket \leq 0, \quad f(y_0) - f_\star \leq R. \end{aligned}$$

Finally, we note that although (2) and (3) both originate from the relaxed optimization problems, they have been commonly employed to achieve the convergence analysis. The function value convergence rate of OGM[26], FGM[37], G-OGM[28], GD[19], OBL-F<sub>b</sub>[44] can be proved by using (2) with appropriate  $\{u_i\}_{i=0}^N$ . The gradient norm convergence rate of OGM-G[29], OBL-G<sub>b</sub>[44], M-OGM-G[60], and [17, Lemma 2.6] can be proved by using (3) with appropriate  $\{v_i\}_{i=0}^N$ . We also note that recent works [2, 25] do not employ (2) to achieve the convergence rate, particularly for gradient descent with varying step sizes.

### 3 H-duality in continuous time

We now establish a continuous-time analog of the H-duality theorem. As the continuous-time result and, especially, its proof is much simpler than its discrete-time counterpart, the results of this section serve as a vehicle to convey the key ideas more clearly. Let  $T > 0$  be a pre-specified terminal time. Let  $H(t, s)$  be an appropriately integrable<sup>2</sup> real-valued kernel with domain  $\{(t, s) \mid 0 < s < t < T\}$ . We define a Continuous-time Fixed Step First Order Method (C-FSFOM) with  $H$  as

$$X(0) = x_0, \quad \dot{X}(t) = - \int_0^t H(t, s) \nabla f(X(s)) ds, \quad \forall t \in (0, T) \quad (7)$$

for any initial point  $x_0 \in \mathbb{R}^d$  and differentiable  $f$ . Note, the Euler discretization of C-FSFOMs (7) corresponds to FSFOMs (1). The notion of C-FSFOMs has been considered previously in [30].

Given a kernel  $H(t, s)$ , analogously define its *anti-transpose* as  $H^A(t, s) = H(T - s, T - t)$ . We call [C-FSFOM with  $H^A$ ] the **H-dual** of [C-FSFOM with  $H$ ]. In the special case  $H(t, s) = e^{\gamma(s) - \gamma(t)}$  for some function  $\gamma(\cdot)$ , the C-FSFOMs with  $H$  and its H-dual have the form

$$\begin{aligned} \ddot{X}(t) + \dot{\gamma}(t)\dot{X}(t) + \nabla f(X(t)) &= 0 && \text{(C-FSFOM with } H(t, s) = e^{\gamma(s) - \gamma(t)}) \\ \ddot{Y}(t) + \dot{\gamma}(T - t)\dot{Y}(t) + \nabla f(Y(t)) &= 0 && \text{(C-FSFOM with } H^A(t, s)) \end{aligned}$$

Interestingly, friction terms with  $\gamma'$  have time-reversed dependence between the H-duals, and this is why we refer to this phenomenon as time-reversed dissipation.

#### 3.1 Continuous-time H-duality theorem

For the C-FSFOM with  $H$ , define the energy function

$$\mathcal{U}(t) = \frac{1}{2} \|X(t) - x_\star\|^2 + \int_0^t u'(s) [x_\star, X(s)] ds \quad (8)$$

for  $t \in [0, T]$  with differentiable  $u: (0, T) \rightarrow \mathbb{R}$ . If  $u'(\cdot) \geq 0$ , then  $\{\mathcal{U}(t)\}_{t \in [0, T]}$  is dissipative. Assume we can show

$$u(T) (f(X(T)) - f_\star) \leq \mathcal{U}(T) \quad (\forall X(0), x_\star, \{\nabla f(X(s))\}_{s \in [0, T]} \in \mathbb{R}^d). \quad (C3)$$

Then, the C-FSFOM with  $H$  exhibits the convergence rate

$$u(T) (f(X(T)) - f_\star) \leq \mathcal{U}(T) \leq \mathcal{U}(0) = \frac{1}{2} \|X(0) - x_\star\|^2.$$

For the C-FSFOM with  $H^A$ , define the energy function

$$\mathcal{V}(t) = v(0) (f(Y(0)) - f(Y(T))) + \int_0^t v'(s) [Y(T), Y(s)] ds \quad (9)$$

for  $t \in [0, T]$  with differentiable  $v: (0, T) \rightarrow \mathbb{R}$ . If  $v'(\cdot) \geq 0$ , then  $\{\mathcal{V}(t)\}_{t \in [0, T]}$  is dissipative. Assume we can show

$$\frac{1}{2} \|\nabla f(Y(T))\|^2 \leq \mathcal{V}(T) \quad (\forall Y(0), \{\nabla f(Y(s))\}_{s \in [0, T]} \in \mathbb{R}^d). \quad (C4)$$

Then, the C-FSFOM with  $H^A$  exhibits the convergence rate

$$\frac{1}{2} \|\nabla f(Y(T))\|^2 \leq \mathcal{V}(T) \leq \mathcal{V}(0) = v(0) (f(Y(0)) - f(Y(T))) \leq v(0) (f(Y(0)) - f_\star).$$

**Theorem 2** (informal). Consider differentiable functions  $u, v: (0, T) \rightarrow \mathbb{R}$  related through  $v(t) = \frac{1}{u(T-t)}$  for  $t \in [0, T]$ . Assume certain regularity conditions (specified in Appendix C.2). Then,

$$[(C3) \text{ is satisfied with } u(\cdot) \text{ and } H] \Leftrightarrow [(C4) \text{ is satisfied with } v(\cdot) \text{ and } H^A].$$

The formal statement of 2 and its proof are given in Appendix C.2. Loosely speaking, we can consider Theorem 2 as the limit of Theorem 1 with  $N \rightarrow \infty$ .

<sup>2</sup>In this paper, we avoid analytical and measure-theoretic details and focus on convergence (rather than existence) results of the continuous-time dynamics.

### 3.2 Verifying conditions for H-duality theorem

As an illustrative example, consider the case  $H(t, s) = \frac{s^r}{t^r}$  for  $r \geq 3$  which corresponds to an ODE studied in the prior work [50, 51]. For the C-FSFOM with  $H$ , the choice  $u(t) = \frac{t^2}{2(r-1)}$  for the dissipative energy function  $\{\mathcal{U}(t)\}_{t=0}^T$  of (8) leads to

$$\mathcal{U}(T) - u(T) (f(X(T)) - f_*) = \frac{\|T\dot{X}(T) + 2(X(T) - x_*)\|^2 + 2(r-3)\|X(T) - x_*\|^2}{4(r-1)} + \int_0^T \frac{(r-3)s}{2(r-1)} \|\dot{X}(s)\|^2 ds.$$

For the C-FSFOM with  $H^A$ , the choice  $v(t) = \frac{1}{u(T-t)} = \frac{2(r-1)}{(T-t)^2}$  for the dissipative energy function  $\{\mathcal{V}(t)\}_{t=0}^T$  of (9) leads to

$$\mathcal{V}(T) - \frac{1}{2} \|\nabla f(Y(T))\|^2 = \frac{2(r-1)(r-3)\|Y(0) - Y(T)\|^2}{T^4} + \int_0^T \frac{2(r-1)(r-3)\|(T-s)\dot{Y}(s) + 2(Y(s) - Y(T))\|^2}{(T-s)^5} ds.$$

Since the right-hand sides are expressed as sums/integrals of squares, they are nonnegative, so (C3) and (C4) hold. (By Theorem 2, verifying (C3) implies (C4) and vice versa.) The detailed calculations are provided in Appendix C.1.

### 3.3 Applications of continuous-time H-duality theorem

The C-FSFOM (7) with  $H(t, s) = \frac{Cp^2 s^{2p-1}}{t^{p+1}}$  recovers

$$\ddot{X}(t) + \frac{p+1}{t} \dot{X}(t) + Cp^2 t^{p-2} \nabla f(X(t)) = 0,$$

an ODE considered in [56]. The rate  $f(X(T)) - f_* \leq \frac{1}{2CT^p} \|X(0) - x_*\|^2$  can be established from (8) with  $u(t) = Ct^p$ . The C-FSFOM with  $H^A$  can be expressed as the ODE

$$\ddot{Y}(t) + \frac{2p-1}{T-t} \dot{Y}(t) + Cp^2 (T-t)^{p-2} \nabla f(Y(t)) = 0. \quad (10)$$

By Theorem 2, using (9) with  $v(t) = \frac{1}{C(T-t)^p}$  leads to the rate

$$\frac{1}{2} \|\nabla f(Y(T))\|^2 \leq \frac{1}{CT^p} (f(Y(0)) - f_*).$$

Note that the continuous-time models of (OGM) and (OGM-G), considered in [51], are special cases of this setup with  $p = 2$  and  $C = 1/2$ . The detailed derivation and well-definedness of the ODE are presented in Appendix C.3.

## 4 New method efficiently reducing gradient mapping norm: (SFG)

In this section, we introduce a novel algorithm obtained using the insights of Theorem 1. Consider minimizing  $F(x) := f(x) + g(x)$ , where  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  is  $L$ -smooth convex with  $0 < L < \infty$  and  $g: \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  is a closed convex proper function. Write  $F_* = \inf_{x \in \mathbb{R}^n} F(x)$  for the minimum value. For  $\alpha > 0$ , define the  $\alpha$ -proximal gradient step as

$$y^{\oplus, \alpha} = \operatorname{argmin}_{z \in \mathbb{R}^n} \left( f(y) + \langle \nabla f(y), z - y \rangle + g(z) + \frac{\alpha L}{2} \|z - y\|^2 \right) = \operatorname{Prox}_{\frac{g}{\alpha L}} \left( y - \frac{1}{\alpha L} \nabla f(y) \right).$$

Consider FSFOMs defined by a lower triangular matrix  $H = \{h_{k,i}\}_{0 \leq i < k \leq N}$  as follows:

$$x_{k+1} = x_k - \sum_{i=0}^k \alpha h_{k+1,i} (x_i - x_i^{\oplus, \alpha}), \quad \forall k = 0, \dots, N-1.$$

When  $g = 0$ , this reduces to (1). FISTA [11], FISTA-G [31] and GFPGM [27] are instances of this FSFOM with  $\alpha = 1$ . In this section, we present a new method for efficiently reducing the gradient mapping norm. This method is faster than the prior state-of-the-art FISTA-G [31] by a constant factor of 5.28 while having substantially simpler coefficients.

**Theorem 3.** Consider the method

$$\begin{aligned}
 y_{k+1} &= y_k^{\oplus,4} + \frac{(N-k+1)(2N-2k-1)}{(N-k+3)(2N-2k+1)} \left( y_k^{\oplus,4} - y_{k-1}^{\oplus,4} \right) + \frac{(4N-4k-1)(2N-2k-1)}{6(N-k+3)(2N-2k+1)} \left( y_k^{\oplus,4} - y_k \right) \\
 y_N &= y_{N-1}^{\oplus,4} + \frac{3}{10} \left( y_{N-1}^{\oplus,4} - y_{N-2}^{\oplus,4} \right) + \frac{3}{40} \left( y_{N-1}^{\oplus,4} - y_{N-1} \right)
 \end{aligned} \tag{SFG}$$

for  $k = 0, \dots, N - 2$ , where  $y_{-1}^{\oplus,4} = y_0$ . This method exhibits the rate

$$\min_{v \in \partial F(y_N^{\oplus,4})} \|v\|^2 \leq 25L^2 \left\| y_N - y_N^{\oplus,4} \right\|^2 \leq \frac{50L}{(N+2)(N+3)} (F(y_0) - F_*) .$$

We call this method *Super FISTA-G* (SFG), and in Appendix D.3, we present a further general parameterized family (SFG-family). To derive (SFG-family), we start with the parameterized family GFPGM [27], which exhibits an accelerated rate on function values, and expresses it as FSFOMs with  $H$ . We then obtain the FSFOMs with  $H^A + C$ , where  $C$  is a lower triangular matrix satisfying certain constraints. We find that the appropriate H-dual for the composite setup is given by this  $H^A + C$ , rather than  $H^A$ . We provide the proof of Theorem 3 in Appendix D.2.

(SFG) is an instance of (SFG-family) with simple rational coefficients. Among the family, the optimal choice has complicated coefficients, but its rate has a leading coefficient of 46, which is slightly smaller than the 50 of (SFG). We provide the details Appendix D.4.

## 5 Conclusion

In this work, we defined the notion of H-duality and formally established that the H-dual of an optimization method designed to efficiently reduce function values is another method that efficiently reduces gradient magnitude. For optimization algorithms, the notion of equivalence, whether informal or formal [59], is intuitive and standard. For optimization problems, the notion of equivalence is also standard, but the beauty of convex optimization is arguably derived from the elegant duality of optimization problems. In fact, there are many notions of duality for spaces, problems, operators, functions, sets, etc. However, the notion of duality for algorithms is something we, the authors, are unfamiliar with within the context of optimization, applied mathematics, and computer science. In our view, the significance of this work is establishing the first instance of a duality of algorithms. The idea that an optimization algorithm is an abstract mathematical object that we can take the dual of opens the door to many interesting questions. In particular, exploring for what type of algorithms the H-dual or a similar notion of duality makes sense is an interesting direction for future work.

## Acknowledgments and Disclosure of Funding

We thank Soonwon Choi, G. Bruno De Luca, and Eva Silverstein for sharing their insights as physicists. We thank Jaewook J. Suh for reviewing the manuscript and providing valuable feedback. We are grateful to Jungbin Kim and Jeongwhan Lee for their valuable discussions at the outset of this work. J.K. and E.K.R. were supported by the National the Samsung Science and Technology Foundation (Project Number SSTF-BA2101-02). C.P. acknowledges support from the Xianhong Wu Fellowship, the Korea Foundation for Advanced Studies, and the Siebel Scholarship. A.O. acknowledges support from the MIT-IBM Watson AI Lab grant 027397-00196, Trustworthy AI grant 029436-00103, and the MIT-DSTA grant 031017-00016. Finally, we also thank anonymous reviewers for giving thoughtful comments.

## References

- [1] K. Ahn and S. Sra. From Nesterov’s estimate sequence to Riemannian acceleration. *COLT*, 2020.
- [2] J. M. Altschuler and P. A. Parrilo. Acceleration by stepsize hedging II: Silver stepsize schedule for smooth convex optimization. *arXiv:2309.16530*, 2023.
- [3] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and K. J. MacPhee. Foundations of gauge and perspective duality. *SIAM Journal on Optimization*, 28(3):2406–2434, 2018.

- [4] H. Attouch and M. Théra. A general duality principle for the sum of two operators. *Journal of Convex Analysis*, 3:1–24, 01 1996.
- [5] J.-F. Aujol and C. Dossal. Optimal rate of convergence of an ODE associated to the fast gradient descent schemes for  $b > 0$ . *HAL Archives Ouvertes*, 2017.
- [6] J.-F. Aujol, C. Dossal, G. Fort, and É. Moulines. Rates of convergence of perturbed FISTA-based algorithms. *HAL Archives Ouvertes*, 2019.
- [7] J.-F. Aujol, C. Dossal, and A. Rondepierre. Optimal convergence rates for Nesterov acceleration. *SIAM Journal on Optimization*, 29(4):3131–3153, 2019.
- [8] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.
- [9] M. Baes. Estimate sequence methods: extensions and approximations. 2009.
- [10] N. Bansal and A. Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(4):1–32, 2019.
- [11] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- [12] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [13] R. E. Bruck. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 61(1):159–164, 1977.
- [14] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2005.
- [15] G. B. Dantzig, E. Eisenberg, and R. W. Cottle. Symmetric dual nonlinear programs. *Pacific Journal of Mathematics*, 15(3):809–812, 1965.
- [16] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.
- [17] J. Diakonikolas and P. Wang. Potential function-based framework for making the gradients small in convex and min-max optimization. *SIAM Journal on Optimization*, 32(3):1668–1697, 2021.
- [18] Y. Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.
- [19] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014.
- [20] W. Fenchel. On conjugate convex functions. *Canadian Journal of Mathematics*, 1(1):73–77, 1949.
- [21] R. M. Freund. Dual gauge programs, with applications to quadratic programming and the minimum-norm problem. *Mathematical Programming*, 38(1):47–67, 1987.
- [22] M. P. Friedlander, I. Macê do, and T. K. Pong. Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014.
- [23] B. Grimmer. Radial duality part II: Applications and algorithms. *arXiv:2104.11185*, 2021.
- [24] B. Grimmer. Radial duality part I: Foundations. *arXiv:2104.11179*, 2022.
- [25] B. Grimmer. Provably faster gradient descent via long steps. *arXiv:2307.06324*, 2023.
- [26] D. Kim and J. A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1):81–107, 2016.

- [27] D. Kim and J. A. Fessler. Another look at the fast iterative shrinkage/thresholding algorithm (FISTA). *SIAM Journal on Optimization*, 28(1):223–250, 2018.
- [28] D. Kim and J. A. Fessler. Generalizing the optimized gradient method for smooth convex minimization. *SIAM Journal on Optimization*, 28(2):1920–1950, 2018.
- [29] D. Kim and J. A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, 188(1):192–219, 2021.
- [30] J. Kim and I. Yang. Unifying nesterov’s accelerated gradient methods for convex and strongly convex objective functions: From continuous-time dynamics to discrete-time algorithms. *ICML*, 2023.
- [31] J. Lee, C. Park, and E. K. Ryu. A geometric structure of acceleration and its role in making gradients small fast. *NeurIPS*, 2021.
- [32] B. Li, M. Coutiño, and G. B. Giannakis. Revisit of estimate sequence for accelerated gradient methods. *ICASSP*, 2020.
- [33] O. L. Mangasarian and J. Ponstein. Minmax and duality in nonlinear programming. *Journal of Mathematical Analysis and Applications*, 11:504–518, 1965.
- [34] A. S. Nemirovski. Optimization II: Numerical methods for nonlinear continuous optimization., 1999.
- [35] A. S. Nemirovsky. On optimality of Krylov’s information when solving linear operator equations. *Journal of Complexity*, 7(2):121–130, 1991.
- [36] A. S. Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.
- [37] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence  $\mathcal{O}(1/k^2)$ . *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.
- [38] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.
- [39] Y. Nesterov. Accelerating the cubic regularization of Newton’s method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.
- [40] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.
- [41] Y. Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012.
- [42] Y. Nesterov. *Lectures on Convex Optimization*. Springer, 2nd edition, 2018.
- [43] C. Park, J. Park, and E. K. Ryu. Factor- $\sqrt{2}$  acceleration of accelerated gradient methods. *Applied Mathematics and Optimization*, 2023.
- [44] C. Park and E. K. Ryu. Optimal first-order algorithms as a function of inequalities. *arXiv:2110.11035*, 2021.
- [45] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.
- [46] R. T. Rockafellar. *Convex Analysis*. Princeton university press, 1970.
- [47] R. T. Rockafellar. *Conjugate Duality and Optimization*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1974.
- [48] E. K. Ryu and W. Yin. *Large-Scale Convex Optimization*. Cambridge University Press, 2022.

- [49] J. Stoer. Duality in nonlinear programming and the minimax theorem. *Numerische Mathematik*, 5(1):371–379, 1963.
- [50] W. Su, S. Boyd, and E. Candes. A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights. *NeurIPS*, 2014.
- [51] J. J. Suh, G. Roh, and E. K. Ryu. Continuous-time analysis of accelerated gradient methods via conservation laws in dilated coordinate systems. *ICML*, 2022.
- [52] A. B. Taylor. Convex interpolation and performance estimation of first-order methods for convex optimization. 2017.
- [53] A. B. Taylor and F. Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. *COLT*, 2019.
- [54] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3):1283–1313, 2017.
- [55] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017.
- [56] A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences of the United States of America*, 113(47):E7351—E7358, 2016.
- [57] P. Wolfe. A duality theorem for non-linear programming. *Quarterly of applied mathematics*, 19(3):239–244, 1961.
- [58] S. Yamanaka and N. Yamashita. Duality of optimization problems with gauge functions. *Optimization*, pages 1–31, 2022.
- [59] S. Zhao, L. Lessard, and M. Udell. An automatic system to detect equivalence between iterative algorithms. *arXiv:2105.04684*, 2021.
- [60] K. Zhou, L. Tian, A. M.-C. So, and J. Cheng. Practical schemes for finding near-stationary points of convex finite-sums. *AISTATS*, 2022.

## A Proof of Theorem 1

**Reformulate (C1) and (C2) into U and V.** In this paragraph, we will show

$$(C1) \Leftrightarrow \left[ \mathbf{U} \geq 0, \quad (\forall \nabla f(x_0), \dots, \nabla f(x_N) \in \mathbb{R}^d) \right]$$

and

$$(C2) \Leftrightarrow \left[ \mathbf{V} \geq 0, \quad (\forall \nabla f(y_0), \dots, \nabla f(y_N) \in \mathbb{R}^d) \right].$$

Recall the definition of  $\mathbf{U}$  and  $\mathbf{V}$ .

$$\mathbf{U}: = \mathcal{U}_N - u_N(f(x_N) - f_\star) - \frac{L}{2} \left\| x_\star - x_0 + \frac{1}{L} \sum_{i=0}^N (u_i - u_{i-1}) \nabla f(x_i) \right\|^2 \quad (11)$$

$$\mathbf{V}: = \mathcal{V}_N - \frac{1}{2} \|\nabla f(y_N)\|^2. \quad (12)$$

First we calculate  $\mathcal{U}_N - u_N(f(x_N) - f_\star)$ .

$$\begin{aligned} & \mathcal{U}_N - u_N(f(x_N) - f_\star) \\ &= \frac{L}{2} \|x_0 - x_\star\|^2 + \sum_{i=0}^N (u_i - u_{i-1}) \left( f(x_i) - f_\star + \langle \nabla f(x_i), x_\star - x_i \rangle + \frac{1}{2L} \|\nabla f(x_i)\|^2 \right) \\ & \quad + \sum_{i=0}^{N-1} u_i \left( f(x_{i+1}) - f(x_i) + \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle + \frac{1}{2L} \|\nabla f(x_i) - \nabla f(x_{i+1})\|^2 \right) - u_N(f(x_N) - f_\star) \\ &\stackrel{(\circ)}{=} \frac{L}{2} \|x_0 - x_\star\|^2 + \sum_{i=0}^N (u_i - u_{i-1}) \left( \langle \nabla f(x_i), x_\star - x_i \rangle + \frac{1}{2L} \|\nabla f(x_i)\|^2 \right) \\ & \quad + \sum_{i=0}^{N-1} u_i \left( \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle + \frac{1}{2L} \|\nabla f(x_i) - \nabla f(x_{i+1})\|^2 \right) \\ &= \frac{L}{2} \left\| x_0 - x_\star + \sum_{i=0}^N \frac{u_i - u_{i-1}}{L} \nabla f(x_i) \right\|^2 - \frac{1}{2L} \left\| \sum_{i=0}^N (u_i - u_{i-1}) \nabla f(x_i) \right\|^2 \\ & \quad + \sum_{i=0}^N (u_i - u_{i-1}) \left( \langle \nabla f(x_i), x_0 - x_i \rangle + \frac{1}{2L} \|\nabla f(x_i)\|^2 \right) \\ & \quad + \sum_{i=0}^{N-1} u_i \left( \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle + \frac{1}{2L} \|\nabla f(x_i) - \nabla f(x_{i+1})\|^2 \right). \end{aligned}$$

Note that all function value terms are deleted at  $\circ$ . Therefore,

$$\begin{aligned} \mathbf{U} &= -\frac{1}{2L} \left\| \sum_{i=0}^N (u_i - u_{i-1}) \nabla f(x_i) \right\|^2 + \sum_{i=0}^N (u_i - u_{i-1}) \left( \langle \nabla f(x_i), x_0 - x_i \rangle + \frac{1}{2L} \|\nabla f(x_i)\|^2 \right) \\ & \quad + \sum_{i=0}^{N-1} u_i \left( \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle + \frac{1}{2L} \|\nabla f(x_i) - \nabla f(x_{i+1})\|^2 \right) \end{aligned} \quad (13)$$

and

$$\begin{aligned} & \mathcal{U}_N - u_N(f(x_N) - f_\star) \\ &= \mathbf{U} + \frac{L}{2} \left\| x_0 - x_\star + \sum_{i=0}^N \frac{u_i - u_{i-1}}{L} \nabla f(x_i) \right\|^2. \end{aligned}$$

Since  $(x_0 - x_i), (x_i - x_{i+1}) \in \text{span}\{\nabla f(x_0), \dots, \nabla f(x_N)\}$ , the value of  $\mathbf{U}$  is independent with  $x_0, x_*$ . Thus the only term that depends on  $x_0$  and  $x_*$  is  $\frac{L}{2} \left\| x_0 - x_* + \sum_{i=0}^N \frac{u_i - u_{i-1}}{L} \nabla f(x_i) \right\|^2$ . Next, since  $x_0, x_*$  can have any value, we can take  $x_0 - x_* = \frac{u_i - u_{i-1}}{L} \nabla f(x_i)$ . Thus it gives the fact that (C1) is equivalent to  $\left[ \mathbf{U} \geq 0, \quad (\forall \nabla f(x_0), \dots, \nabla f(x_N) \in \mathbb{R}^d) \right]$ .

Now we calculate  $\mathcal{V}_N - \frac{1}{2} \|\nabla f(y_N)\|^2$ .

$$\begin{aligned}
& \mathcal{V}_N - \frac{1}{2} \|\nabla f(y_N)\|^2 \\
&= v_0 \left( f_* - f(y_N) + \frac{1}{2L} \|\nabla f(y_N)\|^2 \right) + v_0 (f(y_0) - f_*) \\
&\quad + \sum_{i=0}^{N-1} v_{i+1} \left( f(y_{i+1}) - f(y_i) + \langle \nabla f(y_{i+1}), y_i - y_{i+1} \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_{i+1})\|^2 \right) \\
&\quad + \sum_{i=0}^{N-1} (v_{i+1} - v_i) \left( f(y_i) - f(y_N) + \langle \nabla f(y_i), y_N - y_i \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2 \right) - \frac{1}{2L} \|\nabla f(y_N)\|^2 \\
&\stackrel{(\circ)}{=} \frac{v_0}{2L} \|\nabla f(y_N)\|^2 + \sum_{i=0}^{N-1} v_{i+1} \left( \langle \nabla f(y_{i+1}), y_i - y_{i+1} \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_{i+1})\|^2 \right) \\
&\quad + \sum_{i=0}^{N-1} (v_{i+1} - v_i) \left( \langle \nabla f(y_i), y_N - y_i \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2 \right) - \frac{1}{2L} \|\nabla f(y_N)\|^2.
\end{aligned}$$

Note that all function values are deleted at  $(\circ)$ . By the calculation result,

$$\begin{aligned}
\mathbf{V} &= \frac{v_0}{2L} \|\nabla f(y_N)\|^2 + \sum_{i=0}^{N-1} v_{i+1} \left( \langle \nabla f(y_{i+1}), y_i - y_{i+1} \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_{i+1})\|^2 \right) \\
&\quad + \sum_{i=0}^{N-1} (v_{i+1} - v_i) \left( \langle \nabla f(y_i), y_N - y_i \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2 \right) - \frac{1}{2L} \|\nabla f(y_N)\|^2.
\end{aligned} \tag{14}$$

(C2) is equivalent to  $\left[ \mathbf{V} \geq 0, \quad (\forall \nabla f(y_0), \dots, \nabla f(y_N) \in \mathbb{R}^d) \right]$ . To establish Theorem 1, demonstrating

$$\left[ \mathbf{U} \geq 0, \quad (\forall \nabla f(x_0), \dots, \nabla f(x_N) \in \mathbb{R}^d) \right] \Leftrightarrow \left[ \mathbf{V} \geq 0, \quad (\forall \nabla f(y_0), \dots, \nabla f(y_N) \in \mathbb{R}^d) \right] \tag{15}$$

would suffice.

**Transforming  $\mathbf{U}$  and  $\mathbf{V}$  into a trace.** Define

$$\begin{aligned}
g_x &:= [\nabla f(x_0) | \nabla f(x_1) | \dots | \nabla f(x_N)] \in \mathbb{R}^{d \times (N+1)}, \\
g_y &:= [\nabla f(y_0) | \nabla f(y_1) | \dots | \nabla f(y_N)] \in \mathbb{R}^{d \times (N+1)}.
\end{aligned}$$

In this paragraph, we convert the  $\mathbf{U}$  of (11) and the  $\mathbf{V}$  of (12) into the trace of symmetric matrices. The key idea is: For each  $\langle a, b \rangle$  term where  $a, b \in \text{span}\{\nabla f(x_0), \dots, \nabla f(x_N)\}$ , we can write  $a = g_x \mathbf{a}$ ,  $b = g_x \mathbf{b}$  for some  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{(N+1) \times 1}$ . Then

$$\langle a, b \rangle = \langle g_x \mathbf{a}, g_x \mathbf{b} \rangle = \mathbf{b}^\top g_x^\top g_x \mathbf{a} = \text{Tr}(\mathbf{b}^\top g_x^\top g_x \mathbf{a}) = \text{Tr}(\mathbf{a} \mathbf{b}^\top g_x^\top g_x) = \text{Tr}(g_x \mathbf{a} \mathbf{b}^\top g_x^\top). \tag{16}$$

Also note that  $(x_0 - x_i), (x_i - x_{i+1}) \in \text{span}\{\nabla f(x_0), \dots, \nabla f(x_N)\}$  and  $(y_i - y_N), (y_i - y_{i+1}) \in \text{span}\{\nabla f(y_0), \dots, \nabla f(y_N)\}$ . By using this technique, we observe that there exists  $\mathcal{S}(H, u)$ ,  $\mathcal{T}(H^A, v)$  that satisfy

$$(11) = \text{Tr}(g_x \mathcal{S}(H, u) g_x^\top), \quad (12) = \text{Tr}(g_y \mathcal{T}(H^A, v) g_y^\top).$$

From now, we specifically calculate  $\mathcal{S}(H, u)$  and  $\mathcal{T}(H^A, v)$ . Denote  $\{\mathbf{e}_i\}_{i=0}^N \in \mathbb{R}^{(N+1) \times 1}$  as a unit vector which  $(i+1)$ -th component is 1,  $\mathbf{e}_{-1} = \mathbf{e}_{N+1} = \mathbf{0}$  and define  $\mathcal{H}$  as

$$\mathcal{H} = \begin{bmatrix} 0 & 0 \\ H & 0 \end{bmatrix} \in R^{(N+1) \times (N+1)}.$$

Then, by the definition of  $g_x$  and  $\mathcal{H}$ , we have

$$g_x \mathbf{e}_i = \nabla f(x_i) \quad 0 \leq i \leq N, \quad \frac{1}{L} g_x \mathcal{H}^\top \mathbf{e}_0 = 0, \quad (17)$$

$$\frac{1}{L} g_x \mathcal{H}^\top \mathbf{e}_{i+1} = \frac{1}{L} \sum_{j=0}^i h_{i,j} g_x \mathbf{e}_j = \frac{1}{L} \sum_{j=0}^i h_{i,j} \nabla f(x_j) = x_i - x_{i+1} \quad 0 \leq i \leq N-1. \quad (18)$$

Therefore, we can express (11) with  $\mathcal{H}$ ,  $\{u_i\}_{i=0}^N$ ,  $g_x$  and  $\{\mathbf{e}_i\}_{i=0}^N$  using (17) and (18) as

$$\begin{aligned} (11) &= -\frac{1}{2L} \left\| \sum_{i=0}^N (u_i - u_{i-1}) \nabla f(x_i) \right\|^2 + \sum_{i=0}^N (u_i - u_{i-1}) \left( \langle \nabla f(x_i), x_0 - x_i \rangle + \frac{1}{2L} \|\nabla f(x_i)\|^2 \right) \\ &\quad + \sum_{i=0}^{N-1} u_i \left( \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle + \frac{1}{2L} \|\nabla f(x_i) - \nabla f(x_{i+1})\|^2 \right) \\ &= -\frac{1}{2L} \left\| \sum_{i=0}^N (u_i - u_{i-1}) g_x \mathbf{e}_i \right\|^2 + \sum_{i=0}^N (u_i - u_{i-1}) \left( \left\langle g_x \mathbf{e}_i, \frac{1}{L} g_x \mathcal{H}^\top (\mathbf{e}_0 + \dots + \mathbf{e}_i) \right\rangle + \frac{1}{2L} \|g_x \mathbf{e}_i\|^2 \right) \\ &\quad + \sum_{i=0}^{N-1} u_i \left( \left\langle g_x \mathbf{e}_{i+1}, \frac{1}{L} g_x \mathcal{H}^\top \mathbf{e}_{i+1} \right\rangle + \frac{1}{2L} \|g_x (\mathbf{e}_i - \mathbf{e}_{i+1})\|^2 \right). \end{aligned}$$

Using (16) induces (11) =  $\text{Tr}(g_x \mathcal{S}(H, u) g_x^\top)$  where

$$\begin{aligned} \mathcal{S}(H, u) &= -\frac{1}{2L} \left( \sum_{i=0}^N (u_i - u_{i-1}) \mathbf{e}_i \right) \left( \sum_{i=0}^N (u_i - u_{i-1}) \mathbf{e}_i \right)^\top \\ &\quad + \frac{1}{2L} \mathcal{H}^\top \left[ \sum_{i=0}^N u_i (\mathbf{e}_0 + \dots + \mathbf{e}_i) (\mathbf{e}_i - \mathbf{e}_{i+1})^\top \right] + \frac{1}{2L} \left[ \sum_{i=0}^N u_i (\mathbf{e}_i - \mathbf{e}_{i+1}) (\mathbf{e}_0 + \dots + \mathbf{e}_i)^\top \right] \mathcal{H} \\ &\quad + \frac{1}{2L} \left[ \sum_{i=0}^N u_i ((\mathbf{e}_i - \mathbf{e}_{i+1}) \mathbf{e}_i^\top + \mathbf{e}_i (\mathbf{e}_i - \mathbf{e}_{i+1})^\top) - u_N \mathbf{e}_N \mathbf{e}_N^\top \right]. \end{aligned} \quad (19)$$

Similarly, we calculate (12). Define  $\mathcal{H}^A$  as a anti-transpose matrix of  $\mathcal{H}$ :

$$\mathcal{H}^A = \begin{bmatrix} 0 & 0 \\ H^A & 0 \end{bmatrix} \in R^{(N+1) \times (N+1)}.$$

Then, by the definition of  $g_y$  and  $\mathcal{H}^A$ , we have

$$g_y \mathbf{e}_i = \nabla f(y_i) \quad 0 \leq i \leq N, \quad \frac{1}{L} g_y (\mathcal{H}^A)^\top \mathbf{e}_0 = 0, \quad (20)$$

$$\frac{1}{L} g_y (\mathcal{H}^A)^\top \mathbf{e}_{i+1} = \frac{1}{L} \sum_{j=0}^i h_{i,j} g_y \mathbf{e}_j = \frac{1}{L} \sum_{j=0}^i h_{i,j} \nabla f(y_j) = y_i - y_{i+1} \quad 0 \leq i \leq N-1. \quad (21)$$

Therefore, we can express (12) with  $\mathcal{H}^A$ ,  $\{v_i\}_{i=0}^N$ ,  $g$  and  $\{\mathbf{e}_i\}_{i=0}^N$  using (20) and (21) as  
(12)

$$\begin{aligned}
&= \frac{v_0 - 1}{2L} \|\nabla f(y_N)\|^2 + \sum_{i=0}^{N-1} v_{i+1} \left( \langle \nabla f(y_{i+1}), y_i - y_{i+1} \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_{i+1})\|^2 \right) \\
&\quad + \sum_{i=0}^{N-1} (v_{i+1} - v_i) \left( \langle \nabla f(y_i), y_N - y_i \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2 \right) \\
&= \frac{v_0 - 1}{2L} \|g_y \mathbf{e}_N\|^2 + \sum_{i=0}^{N-1} v_{i+1} \left( \left\langle g_y \mathbf{e}_{i+1}, \frac{1}{L} g_y (\mathcal{H}^A)^\top \mathbf{e}_{i+1} \right\rangle + \frac{1}{2L} \|g_y (\mathbf{e}_i - \mathbf{e}_{i+1})\|^2 \right) \\
&\quad + \sum_{i=0}^{N-1} (v_{i+1} - v_i) \left( - \left\langle g_y \mathbf{e}_i, \frac{1}{L} g_y (\mathcal{H}^A)^\top (\mathbf{e}_{i+1} + \dots + \mathbf{e}_N) \right\rangle + \frac{1}{2L} \|g_y (\mathbf{e}_i - \mathbf{e}_N)\|^2 \right).
\end{aligned}$$

We can write (12) =  $\text{Tr} (g_y \mathcal{T}(H^A, v) g_y^\top)$  where

$$\begin{aligned}
&\mathcal{T}(H^A, v) \\
&= \frac{1}{2L} \left[ \sum_{i=0}^N v_i ((\mathbf{e}_{i-1} - \mathbf{e}_i)(\mathbf{e}_{i-1} - \mathbf{e}_N)^\top + (\mathbf{e}_{i-1} - \mathbf{e}_N)(\mathbf{e}_{i-1} - \mathbf{e}_i)^\top) \right] - \frac{v_0}{2L} \mathbf{e}_0 \mathbf{e}_0^\top - \frac{1}{2L} \mathbf{e}_N \mathbf{e}_N^\top \\
&\quad + \frac{1}{2L} \left[ \sum_{i=0}^N v_i \left( (\mathcal{H}^A)^\top (\mathbf{e}_i + \dots + \mathbf{e}_N)(\mathbf{e}_i - \mathbf{e}_{i-1})^\top + (\mathbf{e}_i - \mathbf{e}_{i-1})(\mathbf{e}_i + \dots + \mathbf{e}_N)^\top \mathcal{H}^A \right) \right] \\
&= \frac{1}{2L} \left[ \sum_{i=0}^N \frac{1}{u_{N-i}} ((\mathbf{e}_{i-1} - \mathbf{e}_i)(\mathbf{e}_{i-1} - \mathbf{e}_N)^\top + (\mathbf{e}_{i-1} - \mathbf{e}_N)(\mathbf{e}_{i-1} - \mathbf{e}_i)^\top) \right] - \frac{1}{2u_N L} \mathbf{e}_0 \mathbf{e}_0^\top - \frac{1}{2L} \mathbf{e}_N \mathbf{e}_N^\top \\
&\quad + \frac{1}{2L} \left[ \sum_{i=0}^N \frac{1}{u_{N-i}} \left( (\mathcal{H}^A)^\top (\mathbf{e}_i + \dots + \mathbf{e}_N)(\mathbf{e}_i - \mathbf{e}_{i-1})^\top + (\mathbf{e}_i - \mathbf{e}_{i-1})(\mathbf{e}_i + \dots + \mathbf{e}_N)^\top \mathcal{H}^A \right) \right].
\end{aligned} \tag{22}$$

**Finding auxiliary matrix  $\mathcal{M}(u)$  that gives the relation between  $\mathcal{S}(H, u)$  and  $\mathcal{T}(H^A, v)$ .** We can show that there exists an invertible  $M(u) \in \mathbb{R}^{(N+1) \times (N+1)}$  such that

$$\mathcal{S}(H, u) = \mathcal{M}(u)^\top \mathcal{T}(H^A, v) \mathcal{M}(u). \tag{23}$$

If we assume the above equation,

$$\text{Tr} (g_x \mathcal{S}(H, u) g_x^\top) = \text{Tr} (g_y \mathcal{T}(H^A, v) g_y^\top) \tag{24}$$

with  $g_y = g_x \mathcal{M}(u)^\top$ . Since  $\mathcal{M}(u)$  is invertible,

$$\{g | g \in \mathbb{R}^{d \times (N+1)}\} = \{g \mathcal{M}(u)^\top | g \in \mathbb{R}^{d \times (N+1)}\}. \tag{25}$$

Also, note that

$$\left[ \mathbf{U} \geq 0 \quad \forall \nabla f(x_0), \dots, \nabla f(x_N) \right] \Leftrightarrow \left[ \text{Tr} (g_x \mathcal{S}(H, u) g_x^\top) \geq 0 \quad \forall g_x \in \mathbb{R}^{d \times (N+1)} \right]$$

and

$$\left[ \mathbf{V} \geq 0 \quad \forall \nabla f(y_0), \dots, \nabla f(y_N) \right] \Leftrightarrow \left[ \text{Tr} (g_y \mathcal{T}(H^A, v) g_y^\top) \geq 0 \quad \forall g_y \in \mathbb{R}^{d \times (N+1)} \right].$$

By combining (24) and (25), we obtain

$$\begin{aligned}
&\left[ \text{Tr} (g_x \mathcal{S}(H, u) g_x^\top) \geq 0 \quad \forall g_x \in \mathbb{R}^{d \times (N+1)} \right] \\
&\Leftrightarrow \left[ \text{Tr} (g_y \mathcal{S}(H, u) g_y^\top) \geq 0 \quad \forall g_x \in \mathbb{R}^{d \times (N+1)}, g_y = g_x \mathcal{M}(u)^\top \right] \\
&\Leftrightarrow \left[ \text{Tr} (g_y \mathcal{T}(H^A, v) g_y^\top) \geq 0 \quad \forall g_y \in \mathbb{R}^{d \times (N+1)} \right].
\end{aligned}$$

To sum up, we obtain (15)

$$\left[ \mathbf{U} \geq 0 \quad \forall \nabla f(x_0), \dots, \nabla f(x_N) \right] \Leftrightarrow \left[ \mathbf{V} \geq 0 \quad \forall \nabla f(y_0), \dots, \nabla f(y_N) \right],$$

which concludes the proof.

**Explicit form of  $\mathcal{M}(u)$  and justification of (23).** Explicit form of  $\mathcal{M}(u)$  is

$$\mathcal{M} = \begin{bmatrix} 0 & \cdots & 0 & 0 & u_N \\ 0 & \cdots & 0 & u_{N-1} & u_N - u_{N-1} \\ 0 & \cdots & u_{N-2} & u_{N-1} - u_{N-2} & u_N - u_{N-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ u_0 & \cdots & u_{N-2} - u_{N-3} & u_{N-1} - u_{N-2} & u_N - u_{N-1} \end{bmatrix} \in \mathbb{R}^{(N+1) \times (N+1)}. \quad (26)$$

Now, we express  $\mathcal{M}(u) = \sum_{0 \leq i, j \leq N} m_{ij}(u) \mathbf{e}_i \mathbf{e}_j^\top$ ,  $\mathcal{S}(H, u) = \sum_{0 \leq i, j \leq N} s_{ij} \mathbf{e}_i \mathbf{e}_j^\top$ , and  $\mathcal{T}(H^A, v) = \sum_{0 \leq i, j \leq N} t_{ij} \mathbf{e}_i \mathbf{e}_j^\top$ . Calculating  $\mathcal{M}^\top(u) \mathcal{T}(H^A, v) \mathcal{M}(u)$  gives

$$\begin{aligned} \mathcal{M}^\top(u) \mathcal{T}(H^A, v) \mathcal{M}(u) &= \left( \sum_{i, j} m_{ij}(u) \mathbf{e}_j \mathbf{e}_i^\top \right) \left( \sum_{i, j} t_{ij} \mathbf{e}_i \mathbf{e}_j^\top \right) \left( \sum_{i, j} m_{ij}(u) \mathbf{e}_i \mathbf{e}_j^\top \right) \\ &= \sum_{i, j} t_{ij} \left( \sum_k m_{ik}(u) \mathbf{e}_k \right) \left( \sum_l m_{jl}(u) \mathbf{e}_l \right)^\top \\ &:= \sum_{i, j} t_{ij} \mathbf{f}_i(u) \mathbf{f}_j(u)^\top. \end{aligned}$$

Thus it is enough to show that  $\sum_{i, j} s_{ij} \mathbf{e}_i \mathbf{e}_j^\top$  and  $\sum_{i, j} t_{ij} \mathbf{f}_i(u) \mathbf{f}_j(u)^\top$  are the same under the basis transformation  $\mathbf{f}_i(u) = \sum_k m_{ik}(u) \mathbf{e}_k$ . From here, we briefly write  $\mathbf{f}_i$  instead  $\mathbf{f}_i(u)$ , and  $\mathbf{f}_{-1} = \mathbf{0}$ . Note that  $u_i(\mathbf{e}_i - \mathbf{e}_{i+1}) = (\mathbf{f}_{N-i} - \mathbf{f}_{N-i-1})$ ,  $0 \leq i \leq N$  by definition of  $\mathcal{M}(u)$ . Therefore, we have

$$\begin{aligned} \frac{1}{L} \mathcal{H}^\top \sum_{i=0}^N u_i (\mathbf{e}_0 + \cdots + \mathbf{e}_i) (\mathbf{e}_i - \mathbf{e}_{i+1})^\top &= \frac{1}{L} \mathcal{H}^\top \sum_{i=0}^N (\mathbf{e}_0 + \cdots + \mathbf{e}_i) (\mathbf{f}_{N-i} - \mathbf{f}_{N-i-1})^\top \\ &= \frac{1}{L} \mathcal{H}^\top \left[ \sum_{i=0}^N e_i \mathbf{f}_{N-i}^\top \right]. \end{aligned}$$

Therefore, we can rewrite (19) as follows:

$$\begin{aligned} \mathcal{S}(H, u) &= -\frac{1}{2L} \mathbf{f}_N \mathbf{f}_N^\top + \frac{1}{2L} \mathcal{H}^\top \left[ \sum_{i=0}^N \mathbf{e}_i \mathbf{f}_{N-i}^\top \right] + \frac{1}{2L} \sum_{i=0}^N \mathbf{f}_{N-i} \mathbf{e}_i^\top \mathcal{H} \\ &\quad + \frac{1}{2L} \sum_{i=0}^N \left[ (\mathbf{f}_{N-i} - \mathbf{f}_{N-i-1}) \mathbf{e}_i^\top + \mathbf{e}_i (\mathbf{f}_{N-i} \mathbf{f}_{N-i-1})^\top \right] - \frac{u_N}{2L} \mathbf{e}_N \mathbf{e}_N^\top \\ &= \underbrace{-\frac{1}{2L} \mathbf{f}_N \mathbf{f}_N^\top}_{\mathbf{A}_1} + \underbrace{\frac{1}{2L} \left[ \sum_{i, j} h_{i, j} \mathbf{e}_j \mathbf{f}_{N-i}^\top \right] + \frac{1}{2L} \left[ \sum_{i, j} h_{i, j} \mathbf{f}_{N-i} \mathbf{e}_j^\top \right]}_{\mathbf{B}_1} \\ &\quad + \underbrace{\frac{1}{2L} \sum_{i=0}^N \left[ (\mathbf{f}_{N-i} - \mathbf{f}_{N-i-1}) \mathbf{e}_i^\top + \mathbf{e}_i (\mathbf{f}_{N-i} - \mathbf{f}_{N-i-1})^\top \right]}_{\mathbf{C}_1} - \underbrace{\frac{u_N}{2L} \mathbf{e}_N \mathbf{e}_N^\top}_{\mathbf{D}_1}. \end{aligned} \quad (27)$$

Similarly, by using

$$\mathbf{e}_{N-i} - \mathbf{e}_{N-i+1} = \frac{1}{u_{N-i}} (\mathbf{f}_i - \mathbf{f}_{i-1}) = v_i (\mathbf{f}_i - \mathbf{f}_{i-1}),$$

we can rewrite (22) as follows:

$$\begin{aligned}
& \mathcal{M}^\top(u)\mathcal{T}(H^A, v)\mathcal{M}(u) \\
&= \frac{1}{2L} \sum_{i=0}^N (\mathbf{e}_{N-i+1} - \mathbf{e}_{N-i})(\mathbf{f}_{i-1} - \mathbf{f}_N)^\top + (\mathbf{f}_{i-1} - \mathbf{f}_N)(\mathbf{e}_{N-i+1} - \mathbf{e}_{N-i})^\top \\
&\quad - \frac{1}{2u_N L} \mathbf{f}_0 \mathbf{f}_0^\top - \frac{1}{2L} \mathbf{f}_N \mathbf{f}_N^\top + \frac{1}{2L} (\mathcal{H}^A)^\top \left[ \sum_{i=0}^N \mathbf{f}_i \mathbf{e}_{N-i}^\top \right] + \frac{1}{2L} \left[ \sum_{i=0}^N \mathbf{e}_{N-i} \mathbf{f}_i^\top \right] \mathcal{H}^A \\
&= - \underbrace{\frac{1}{2u_N L} \mathbf{f}_0 \mathbf{f}_0^\top}_{\mathbf{D}_2} + \underbrace{\frac{1}{2L} \left[ \sum_{i,j} h_{N-i, N-j} \mathbf{f}_j \mathbf{e}_{N-i}^\top \right] + \frac{1}{2} \left[ \sum_{i,j} h_{N-j, N-i} \mathbf{e}_{N-i} \mathbf{f}_j^\top \right]}_{\mathbf{B}_2} \\
&\quad + \underbrace{\left[ \frac{1}{2L} \sum_{i=0}^N (\mathbf{e}_{N-i+1} - \mathbf{e}_{N-i}) \mathbf{f}_{i-1}^\top + \mathbf{f}_{i-1} (\mathbf{e}_{N-i+1} - \mathbf{e}_{N-i})^\top - \frac{1}{2L} (\mathbf{e}_0 \mathbf{f}_N^\top + \mathbf{f}_N \mathbf{e}_0^\top) \right]}_{\mathbf{C}_2} - \underbrace{\frac{1}{2L} \mathbf{f}_N \mathbf{f}_N^\top}_{\mathbf{A}_2}. \tag{28}
\end{aligned}$$

For the final step, we compare (28) and (27) term-by-term, by showing  $\mathbf{X}_1 = \mathbf{X}_2$  for  $\mathbf{X} = \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ .

- $\mathbf{A}_1 = \mathbf{A}_2$  comes directly.
- $\mathbf{B}_1 = \mathbf{B}_2$  comes from changing the summation index  $i \rightarrow N - i$  and  $j \rightarrow N - j$ .
- $\mathbf{C}_1 = \mathbf{C}_2$  comes from the expansion of summation.
- $\mathbf{D}_1 = \mathbf{D}_2$  comes from  $\mathbf{f}_0 = u_N \mathbf{e}_N$ .

Therefore,

$$\mathcal{S}(H, u) = \mathcal{M}^\top(u)\mathcal{T}(H^A, v)\mathcal{M}(u),$$

which concludes the proof.

**Remark 1.** We can interpret  $\mathcal{M}$  as a basis transformation, where

$$u_N \mathbf{e}_N = \mathbf{f}_0, \quad u_i (\mathbf{e}_i - \mathbf{e}_{i+1}) = \mathbf{f}_{N-i} - \mathbf{f}_{N-i-1} \quad i = 0, 1, \dots, N-1. \tag{29}$$

**Remark 2.** To clarify, the quantifier  $[\forall \nabla f(x_0), \dots, \nabla f(x_N)]$  in (C1) means  $\nabla f(x_0), \dots, \nabla f(x_N)$  can be any arbitrary vectors in  $\mathbb{R}^d$ . This is different from  $[\nabla f(x_0), \dots, \nabla f(x_N)]$  be gradient of some  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ . The same is true for (C2).

## B Omitted calculation of Section 2

### B.1 Calculation of $H$ matrices

**OGM and OGM-G.** [26, Proposition 3] provides a recursive formula for (OGM) as follows:

$$h_{k+1, i} = \begin{cases} \frac{\theta_k - 1}{\theta_{k+1}} h_{k, i} & i = 0, \dots, k-2 \\ \frac{\theta_k - 1}{\theta_{k+1}} (h_{k, k-1} - 1) & i = k-1 \\ 1 + \frac{2\theta_k - 1}{\theta_{k+1}} & i = k. \end{cases} \tag{30}$$

If  $k > i$ ,

$$\begin{aligned}
h_{k+1, k-1} &= \frac{\theta_k - 1}{\theta_{k+1}} \frac{2\theta_{k-1} - 1}{\theta_k}, \\
h_{k+1, i} &= \frac{\theta_k - 1}{\theta_{k+1}} h_{k, i} = \dots = \left( \prod_{l=i+2}^k \frac{\theta_l - 1}{\theta_{l+1}} \right) h_{i+1, i-1} = \left( \prod_{l=i+1}^k \frac{\theta_l - 1}{\theta_{l+1}} \right) \frac{2\theta_i - 1}{\theta_{i+1}}.
\end{aligned}$$

Thus  $H_{\text{OGM}}$  can be calculated as

$$H_{\text{OGM}}(k+1, i+1) = \begin{cases} 0 & i > k \\ 1 + \frac{2\theta_{k-1}}{\theta_{k+1}} & i = k \\ \left( \prod_{l=i+1}^k \frac{\theta_l - 1}{\theta_{l+1}} \right) \frac{2\theta_i - 1}{2\theta_{i+1}} & i < k. \end{cases}$$

Recursive formula [29] of (OGM-G) is as following:

$$h_{k+1, i} = \begin{cases} \frac{\theta_{N-i-1} - 1}{\theta_{N-i}} h_{k+1, i+1} & i = 0, \dots, k-2 \\ \frac{\theta_{N-k} - 1}{\theta_{N-k+1}} (h_{k+1, i} - 1) & i = k-1 \\ 1 + \frac{2\theta_{N-k-1} - 1}{\theta_{N-k}} & i = k. \end{cases} \quad (31)$$

If  $k > i$ ,

$$h_{k+1, k-1} = \frac{\theta_{N-k} - 1}{\theta_{N-k+1}} \frac{2\theta_{N-k-1} - 1}{\theta_{N-k}},$$

$$h_{k+1, i} = \frac{\theta_{N-i-1} - 1}{\theta_{N-i}} h_{k+1, i+1} = \dots = \left( \prod_{l=N-k+1}^{N-i-1} \frac{\theta_l - 1}{\theta_{l+1}} \right) h_{k+1, k-1} = \left( \prod_{l=N-k}^{N-i-1} \frac{\theta_l - 1}{\theta_{l+1}} \right) \frac{2\theta_{N-k-1} - 1}{\theta_{N-k}}.$$

Thus  $H_{\text{OGM-G}}$  can be calculated as

$$H_{\text{OGM-G}}(k+1, i+1) = \begin{cases} 0 & i > k \\ 1 + \frac{2\theta_{N-k-1} - 1}{\theta_{N-k}} & i = k \\ \left( \prod_{l=N-k}^{N-i-1} \frac{\theta_l - 1}{\theta_{l+1}} \right) \frac{2\theta_{N-k-1} - 1}{\theta_{N-k}} & i < k \end{cases},$$

which gives  $H_{\text{OGM-G}} = H_{\text{OGM}}^A$ .

**Gradient Descent.** For (GD),  $H(i+1, k+1) = h_{i+1, k} = h\delta_{i+1, k+1}$ , where  $\delta_{i,j}$  is the Kronecker delta. Therefore,  $H_{\text{GD}} = H_{\text{GD}}^A$ .

**OBL-F<sub>b</sub> and OBL-G<sub>b</sub>.** Recall  $\gamma = \sqrt{\frac{N(N+1)}{2}}$ . We obtain the recursive formula of the  $H$  matrix of (OBL-F<sub>b</sub>).

$$h_{k+1, i} = \begin{cases} \frac{k}{k+3} h_{k, i} & k = 0, \dots, N-2, i = 0, \dots, k-2 \\ 1 + \frac{2k}{k+3} & k = 0, \dots, N-2, i = k \\ \frac{k}{k+3} (h_{k, k-1} - 1) & k = 1, \dots, N-2, i = k-1 \\ 1 + \frac{N-1}{\gamma+1} & k = N-1, i = N-1 \\ \frac{N-1}{2(\gamma+1)} (h_{N-1, N-2} - 1) & k = N-1, i = N-2 \\ \frac{N-1}{2(\gamma+1)} h_{N-1, i} & k = N-1, i = 0, \dots, N-3 \end{cases}. \quad (32)$$

By using the above formula, we obtain

$$H_{\text{OBL-F}_b}(k+1, i+1) = \begin{cases} 1 + \frac{2k}{k+3} & k = 0, \dots, N-2, i = k \\ \frac{2i(i+1)(i+2)}{(k+1)(k+2)(k+3)} & k = 0, \dots, N-2, i = 0, \dots, k-1 \\ 1 + \frac{N-1}{\gamma+1} & k = N-1, i = N-1 \\ \frac{i(i+1)(i+2)}{(\gamma+1)N(N+1)} & k = N-1, i = 0, \dots, N-2 \end{cases}.$$

Similarly, we achieve the following recursive formula of the  $H$  matrix of (OBL-G<sub>b</sub>).

$$h_{k+1,i} = \begin{cases} \frac{N-k-1}{N-k+2} h_{k,i} & k = 1, \dots, N-1, i = 0, \dots, k-2 \\ 1 + \frac{2(N-k-1)}{N-k+2} & k = 1, \dots, N-1, i = k \\ \frac{N-k-1}{N-k+2} (h_{k,k-1} - 1) & k = 1, \dots, N-1, i = k-1 \\ 1 + \frac{N-1}{\gamma+1} & k = 0, i = 0 \end{cases}. \quad (33)$$

By using the above recursive formula, we obtain

$$H_{\text{OBL-G}_b}(k+1, i+1) = \begin{cases} 1 + \frac{N-1}{\gamma+1} & k = 0, i = 0 \\ \frac{(N-k-1)(N-k)(N-k+1)}{(\gamma+1)N(N+1)} & k = 1, \dots, N-1, i = 0 \\ 1 + \frac{2(N-k-1)}{N-k+2} & k = 1, \dots, N-1, i = k \\ \frac{2(N-k-1)(N-k)(N-k+1)}{(N-i)(N-i+1)(N-i+2)} & k = 1, \dots, N-1, i = 1, \dots, k-1 \end{cases}.$$

Thus  $H_{\text{OBL-F}_b} = H_{\text{OBL-G}_b}^A$ .

## B.2 Calculation of energy functions

**Calculation of  $\mathbf{U}$  and  $\mathbf{V}$  with  $H$  matrix** In this paragraph, we calculate  $\mathbf{U}$  and  $\mathbf{V}$ . Recall (13) and (14).

First, we put  $x_{k+1} - x_k = -\frac{1}{L} \sum_{i=0}^k h_{k+1,i} \nabla f(x_i)$  to (11). We have

$$\begin{aligned} \mathbf{U} &= -\frac{1}{2L} \left\| \sum_{i=0}^N (u_i - u_{i-1}) \nabla f(x_i) \right\|^2 + \sum_{i=0}^N (u_i - u_{i-1}) \left( \langle \nabla f(x_i), x_0 - x_i \rangle + \frac{1}{2L} \|\nabla f(x_i)\|^2 \right) \\ &\quad + \sum_{i=0}^{N-1} u_i \left( \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle + \frac{1}{2L} \|\nabla f(x_i) - \nabla f(x_{i+1})\|^2 \right) \\ &= -\frac{1}{2L} \left\| \sum_{i=0}^N (u_i - u_{i-1}) \nabla f(x_i) \right\|^2 + \sum_{i=0}^N \frac{u_i - u_{i-1}}{L} \left( \left\langle \nabla f(x_i), \sum_{l=0}^{i-1} \sum_{j=0}^l h_{l+1,j} \nabla f(x_j) \right\rangle + \frac{1}{2} \|\nabla f(x_i)\|^2 \right) \\ &\quad + \sum_{i=0}^{N-1} \frac{u_i}{L} \left( \left\langle \nabla f(x_{i+1}), \sum_{j=0}^i h_{i+1,j} \nabla f(x_j) \right\rangle + \frac{1}{2} \|\nabla f(x_i) - \nabla f(x_{i+1})\|^2 \right). \end{aligned}$$

By arranging, we obtain

$$\mathbf{U} = \sum_{0 \leq j \leq i \leq N} \frac{s_{i,j}}{L} \langle \nabla f(x_i), \nabla f(x_j) \rangle$$

where

$$s_{i,j} = \begin{cases} -\frac{1}{2}(u_N - u_{N-1})^2 + \frac{1}{2}u_N & j = i, i = N \\ -\frac{1}{2}(u_i - u_{i-1})^2 + u_i & j = i, i = 0, \dots, N-1 \\ u_i h_{i,i-1} - u_{i-1} - (u_i - u_{i-1})(u_{i-1} - u_{i-2}) & j = i-1 \\ (u_i - u_{i-1}) \sum_{l=j}^{i-1} h_{l+1,j} + u_{i-1} h_{i,j} - (u_i - u_{i-1})(u_j - u_{j-1}) & j = 0, \dots, i-2. \end{cases}. \quad (34)$$

Recall that we defined  $u_{-1} = 0$ . Next, put  $y_{k+1} - y_k = -\frac{1}{L} \sum_{i=0}^k h_{k+1,i} \nabla f(y_i)$  to (12). We have

$$\begin{aligned}
\mathbf{V} &= \mathcal{V}_N - \frac{1}{2L} \|\nabla f(y_N)\|^2 \\
&= \frac{v_0}{2L} \|\nabla f(y_N)\|^2 + \sum_{i=0}^{N-1} v_{i+1} \left( \langle \nabla f(y_{i+1}), y_i - y_{i+1} \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_{i+1})\|^2 \right) \\
&\quad + \sum_{i=0}^{N-1} (v_{i+1} - v_i) \left( \langle \nabla f(y_i), y_N - y_i \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2 \right) - \frac{1}{2L} \|\nabla f(y_N)\|^2 \\
&= \frac{v_0 - 1}{2L} \|\nabla f(y_N)\|^2 + \sum_{i=0}^{N-1} v_{i+1} \left( \left\langle \nabla f(y_{i+1}), \frac{1}{L} \sum_{j=0}^i h_{i+1,j} \nabla f(y_j) \right\rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_{i+1})\|^2 \right) \\
&\quad + \sum_{i=0}^{N-1} (v_{i+1} - v_i) \left( \left\langle \nabla f(y_i), -\frac{1}{L} \sum_{l=i}^{N-1} \sum_{j=0}^l h_{l+1,j} \nabla f(y_j) \right\rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2 \right).
\end{aligned}$$

By arranging, we obtain

$$\mathbf{V} = \sum_{0 \leq j \leq i \leq N} \frac{t_{i,j}}{L} \langle \nabla f(y_i), \nabla f(y_j) \rangle$$

where

$$t_{i,j} = \begin{cases} \frac{v_1}{2} + \frac{v_1 - v_0}{2} - (v_1 - v_0) \sum_{l=0}^{N-1} h_{l+1,0} & i = 0, j = i \\ \frac{v_{i+1} + v_i}{2} + \frac{v_{i+1} - v_i}{2} - (v_{i+1} - v_i) \sum_{l=i}^{N-1} h_{l+1,i} & i = 1, \dots, N-1, j = i \\ \frac{v_0 - 1}{2} + \frac{v_N}{2} + \sum_{i=0}^{N-1} \frac{v_{i+1} - v_i}{2} & i = N, j = i \\ v_i h_{i,i-1} - v_i - (v_{i+1} - v_i) \sum_{l=i}^{N-1} h_{l+1,i-1} - (v_i - v_{i-1}) \sum_{l=i}^{N-1} h_{l+1,i} & i = 1, \dots, N-1, j = i-1 \\ v_N h_{N,N-1} - v_N - (v_N - v_{N-1}) & i = N, j = i-1 \\ v_i h_{i,j} - (v_{i+1} - v_i) \sum_{l=i}^{N-1} h_{l+1,j} - (v_{j+1} - v_j) \sum_{l=i}^{N-1} h_{l+1,i} & i = 2, \dots, N-1, j = 0, \dots, i-2 \\ v_N h_{N,j} - (v_{j+1} - v_j) & i = N, j = 0, \dots, N-2 \end{cases} \quad (35)$$

Now we calculate  $\{s_{ij}\}$  and  $\{t_{ij}\}$  for [(OGM), (OGM-G)] [(OBL-F<sub>b</sub>), (OBL-G<sub>b</sub>)] and [(GD), (GD)].

### B.2.1 Calculation of energy function of (OGM) and (OGM-G)

We will show  $s_{ij} = 0$  and  $t_{ij} = 0$  for all  $i, j$ . By the definition of  $\{u_i\}_{i=-1}^N$  and  $\{\theta_i\}_{i=-1}^N$ ,

$$\begin{aligned}
u_i - u_{i-1} &= 2\theta_i^2 - 2\theta_{i-1}^2 = 2\theta_i \quad 0 \leq i \leq N-1, \\
u_N - u_{N-1} &= \theta_N^2 - 2\theta_{N-1}^2 = \theta_N.
\end{aligned}$$

Therefore, we have

$$s_{i,i} = \begin{cases} -\frac{1}{2}(u_N - u_{N-1})^2 + \frac{1}{2}u_N = -\theta_N + \theta_N = 0 & j = i, i = N \\ -\frac{1}{2}(u_i - u_{i-1})^2 + u_i = -\theta_i + \theta_i = 0 & j = i, i = 0, \dots, N-1. \end{cases}$$

Now we claim that  $s_{ij} = 0$  when  $j \neq i$ . In the case that  $i = j + 1$ , we have

$$\begin{aligned} s_{i,i-1} &= u_i h_{i,i-1} - u_{i-1} - (u_i - u_{i-1})(u_{i-1} - u_{i-2}) \\ &= u_i \left( \frac{2\theta_{i-1} - 1}{\theta_i} + 1 \right) - u_{i-1} - (u_i - u_{i-1})(u_{i-1} - u_{i-2}) \\ &= \begin{cases} 2\theta_i^2 \left( \frac{2\theta_{i-1} - 1}{\theta_i} + 1 \right) - 2\theta_{i-1}^2 - 4\theta_i \theta_{i-1} & 0 \leq i \leq N-1 \\ \theta_N^2 \left( \frac{2\theta_{N-1} - 1}{\theta_N} + 1 \right) - 2\theta_{N-1}^2 - 2\theta_{N-1} \theta_N & i = N \end{cases} \\ &= 0. \end{aligned}$$

We show  $s_{ij} = 0$  for  $j \neq i$  with induction on  $i$ , i.e., proving

$$\left[ s_{i,j} = (u_i - u_{i-1}) \sum_{l=j}^{i-1} h_{l+1,j} + u_{i-1} h_{i,j} - (u_i - u_{i-1})(u_j - u_{j-1}) = 0, \quad j = 0, \dots, i-2 \right]. \quad (36)$$

First we prove (36) for  $i = j + 2$ .

$$\begin{aligned} &(u_{j+2} - u_{j+1})(h_{j+1,j} + h_{j+2,j}) + u_{j+1} h_{j+2,j} - (u_{j+2} - u_{j+1})(u_j - u_{j-1}) \\ &= (u_{j+2} - u_{j+1})h_{j+1,j} + u_{j+2} h_{j+2,j} - (u_{j+2} - u_{j+1})(u_j - u_{j-1}) \\ &= \begin{cases} 2\theta_{j+2} h_{j+1,j} + 2\theta_{j+2}^2 h_{j+2,j} - 4\theta_{j+2} \theta_j & 0 \leq j \leq N-3 \\ \theta_N h_{N-1,N-2} + \theta_N^2 h_{N,N-2} - 2\theta_N \theta_{N-2} & j = N-2 \end{cases} \\ &= 0. \end{aligned}$$

Next, assume (36) for  $i = i_0$ . When  $i = i_0 + 1$ ,

$$\begin{aligned} &(u_{i_0+1} - u_{i_0}) \sum_{l=j}^{i_0} h_{l+1,j} + u_{i_0} h_{i_0+1,j} - (u_{i_0+1} - u_{i_0})(u_j - u_{j-1}) \\ &= (u_{i_0+1} - u_{i_0}) \left( \sum_{l=j}^{i_0-1} h_{l+1,j} + h_{i_0+1,j} \right) + u_{i_0} h_{i_0+1,j} - (u_{i_0+1} - u_{i_0})(u_j - u_{j-1}) \\ &= (u_{i_0+1} - u_{i_0}) \left( \frac{(u_{i_0} - u_{i_0-1})(u_j - u_{j-1}) - u_{i_0-1} h_{i_0,j}}{u_{i_0} - u_{i_0-1}} + h_{i_0+1,j} \right) + u_{i_0} h_{i_0+1,j} \\ &\quad - (u_{i_0+1} - u_{i_0})(u_j - u_{j-1}) \\ &= u_{i_0+1} h_{i_0+1,j} - \frac{u_{i_0-1}(u_{i_0+1} - u_{i_0})}{u_{i_0} - u_{i_0-1}} h_{i_0,j} \\ &= \begin{cases} 2\theta_{i_0+1}^2 h_{i_0+1,j} - \frac{4\theta_{i_0-1}^2 \theta_{i_0+1}}{2\theta_{i_0}} h_{i_0,j} & 0 \leq i_0 \leq N-2 \\ \theta_N^2 h_{i_0+1,j} - \frac{2\theta_{N-2}^2 \theta_N}{2\theta_{N-1}} h_{i_0,j} & i_0 = N-1 \end{cases} \\ &= 0 \end{aligned}$$

where the second equality comes from the induction hypothesis, and the third equality comes from (30). In sum, we proved  $s_{ij} = 0$  for every  $i$  and  $j$ , which implies  $\mathbf{U} = 0$ .

Next, we will claim that  $t_{ij} = 0$  for all  $i, j$ . Firstly, explicit formula of  $H_{\text{OGM-G}}(k+1, i+1)$  first. When  $k > i$ ,

$$\begin{aligned} H_{\text{OGM-G}}(k+1, i+1) &= \frac{\theta_{N-k} - 1}{\theta_{N-k+1}} \frac{\theta_{N-k+1} - 1}{\theta_{N-k+2}} \dots \frac{\theta_{N-i-1} - 1}{\theta_{N-i}} \frac{2\theta_{N-k-1} - 1}{\theta_{N-k}} \\ &= \frac{\theta_{N-k}^2 - \theta_{N-k}}{\theta_{N-k} \theta_{N-k+1}} \frac{\theta_{N-k+1}^2 - \theta_{N-k+1}}{\theta_{N-k+1} \theta_{N-k+2}} \dots \frac{\theta_{N-i-1}^2 - \theta_{N-i-1}}{\theta_{N-i-1} \theta_{N-i}} \frac{2\theta_{N-k-1} - 1}{\theta_{N-k}} \\ &= \frac{\theta_{N-k-1}^2}{\theta_{N-k} \theta_{N-k+1}} \frac{\theta_{N-k}^2}{\theta_{N-k+1} \theta_{N-k+2}} \dots \frac{\theta_{N-i-2}^2}{\theta_{N-i-1} \theta_{N-i}} \frac{2\theta_{N-k-1} - 1}{\theta_{N-k}} \\ &= \frac{\theta_{N-k-1}^2 (2\theta_{N-k-1} - 1)}{\theta_{N-i-1}^2 \theta_{N-i}}. \end{aligned}$$

To calculate  $\{t_{i,j}\}$ , it is enough to deal with the sum  $\sum_{l=i}^{N-1} h_{l+1,j}$ , which can be expressed as

$$\sum_{l=i}^{N-1} h_{l+1,j} = \begin{cases} \frac{\theta_{N+1}}{2} & i = 0, j = i \\ \theta_{N-i} & i = 1, \dots, N-1, j = i \\ \frac{\theta_{N-i-1}^4}{\theta_{N-j}\theta_{N-j-1}^2} & i = 1, \dots, N-1, j = 0, \dots, i-1 \end{cases}. \quad (37)$$

By inserting (37) in (35),  $[t_{ij} = 0, \forall i, j]$  is obtained, which implies  $\mathbf{V} = 0$ . (37) and (35) are also stated in [29, Lemma 6.1].

### B.2.2 Calculation of energy function of (OBL-F<sub>b</sub>) and (OBL-G<sub>b</sub>)

First we calculate  $\{s_{ij}\}$  for (OBL-F<sub>b</sub>). Recall  $u_i = \frac{(i+1)(i+2)}{2}$  for  $0 \leq i \leq N-1$  and  $u_N = \gamma^2 + \gamma$  where  $\gamma = \sqrt{N(N+1)/2}$ . When  $j = i$ ,

$$\begin{aligned} s_{i,i} &= \begin{cases} -\frac{1}{2}(u_N - u_{N-1})^2 + \frac{1}{2}u_N & i = N \\ -\frac{1}{2}(u_i - u_{i-1})^2 + u_i & 0 \leq i \leq N-1 \end{cases} \\ &= \begin{cases} \frac{\gamma}{2} = \frac{u_N - u_{N-1}}{2} & i = N \\ -\frac{1}{2}(i+1)^2 + \frac{(i+1)(i+2)}{2} = \frac{u_i - u_{i-1}}{2} & 0 \leq i \leq N-1 \end{cases}. \end{aligned}$$

Now we claim that  $s_{ij} = 0$  when  $j \neq i$ . In the case  $j = i-1$ , we have

$$\begin{aligned} s_{i,i-1} &= u_i h_{i,i-1} - u_{i-1} - (u_i - u_{i-1})(u_{i-1} - u_{i-2}) \\ &= \begin{cases} \frac{(i+1)(i+2)}{2} h_{i,i-1} - \frac{i(i+1)}{2} - (i+1)i & 0 \leq i \leq N-1 \\ (\gamma^2 + \gamma) h_{N,N-1} - \frac{N(N+1)}{2} - \gamma N & i = N \end{cases} \\ &= 0. \end{aligned}$$

We show  $s_{ij} = 0$  for  $j \neq i$  with induction on  $i$ , i.e., proving

$$\left[ s_{i,j} = (u_i - u_{i-1}) \sum_{l=j}^{i-1} h_{l+1,j} + u_{i-1} h_{i,j} - (u_i - u_{i-1})(u_j - u_{j-1}) = 0 \quad j = 0, \dots, i-2 \right]. \quad (38)$$

(38) holds when  $i = j+2$  since

$$\begin{aligned} & (u_{j+2} - u_{j+1})(h_{j+1,j} + h_{j+2,j}) + u_{j+1} h_{j+2,j} - (u_{j+2} - u_{j+1})(u_j - u_{j-1}) \\ &= (u_{j+2} - u_{j+1})h_{j+1,j} + u_{j+2} h_{j+2,j} - (u_{j+2} - u_{j+1})(u_j - u_{j-1}) \\ &= \begin{cases} (j+3)h_{j+1,j} + \frac{(j+3)(j+4)}{2} h_{j+2,j} - (j+3)(j+1) & 0 \leq j \leq N-3 \\ \gamma h_{N-1,N-2} + (\gamma^2 + \gamma) h_{N,N-2} - \gamma & j = N-2 \end{cases} \\ &= 0. \end{aligned}$$

Assume (38) for  $i = i_0$ . For  $i = i_0 + 1$ ,

$$\begin{aligned}
& (u_{i_0+1} - u_{i_0}) \sum_{l=j}^{i_0} h_{l+1,j} + u_{i_0} h_{i_0+1,j} - (u_{i_0+1} - u_{i_0})(u_j - u_{j-1}) \\
&= (u_{i_0+1} - u_{i_0}) \left( \sum_{l=j}^{i_0-1} h_{l+1,j} + h_{i_0+1,j} \right) + u_{i_0} h_{i_0+1,j} - (u_{i_0+1} - u_{i_0})(u_j - u_{j-1}) \\
&= (u_{i_0+1} - u_{i_0}) \left( \frac{(u_{i_0} - u_{i_0-1})(u_j - u_{j-1}) - u_{i_0-1} h_{i_0,j}}{u_{i_0} - u_{i_0-1}} + h_{i_0+1,j} \right) + u_{i_0} h_{i_0+1,j} \\
&\quad - (u_{i_0+1} - u_{i_0})(u_j - u_{j-1}) \\
&= u_{i_0+1} h_{i_0+1,j} - \frac{u_{i_0-1}(u_{i_0+1} - u_{i_0})}{u_{i_0} - u_{i_0-1}} h_{i_0,j} \\
&= \begin{cases} \frac{(i_0+2)(i_0+3)}{2} h_{i_0+1,j} - \frac{i_0(i_0+1)(i_0+2)}{2(i_0+1)} h_{i_0,j} & 0 \leq i_0 \leq N-2 \\ (\gamma^2 + \gamma) h_{N,j} - \frac{(N-1)N\gamma}{2N} h_{N-1,j} & i_0 = N-1 \end{cases} \\
&= 0.
\end{aligned}$$

Next, we calculate  $\{t_{ij}\}$  for (OBL-G<sub>b</sub>). We need to deal with the sum  $\sum_{l=k}^{N-1} h_{l+1,i}$ , which can be expressed as

$$\sum_{l=i}^{N-1} h_{l+1,j} = \begin{cases} 1 + \frac{(N+2)(N-1)}{4(\gamma+1)} & i = 0, j = 0 \\ \frac{(N-i+2)(N-i+1)(N-i)(N-i-1)}{4(\gamma+1)N(N+1)} & i = 1, \dots, N-1, j = 0 \\ \frac{(N-i+2)(N-i+1)(N-i)(N-i-1)}{2(N-j)(N-j+1)(N-j+2)} & i = j+1, \dots, N-1, j = 1, \dots, N-1 \\ 1 + \frac{N-i-1}{2} & i = j, j = 1, \dots, N-1 \end{cases}.$$

By combining  $v_0 = \frac{1}{\gamma^2 + \gamma}$ ,  $v_i = \frac{1}{(N-i+1)(N-i+2)}$  for  $1 \leq i \leq N$  and (35), we obtain

$$\begin{aligned}
t_{ij} &= \begin{cases} 1 & i = N, j = N \\ \frac{1}{2N(N+1)} - \frac{v_0}{2} & i = 0, j = 1 \\ v_0 - \frac{1}{N(N+1)} & i = N, j = 0 \\ \frac{1}{(N-i)(N-i+1)(N-i+2)} & i = 1, \dots, N-1, j = i \\ -\frac{2}{(N-i)(N-i+1)(N-i+2)} & i = N, j = 1, \dots, N-1 \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \frac{v_N}{2} & i = N, j = N \\ \frac{v_{i+1} - v_i}{2} & i = 0, \dots, N-1, j = i \\ -v_{i+1} + v_i & i = N, j = 0, \dots, N-1 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Therefore,

$$\mathbf{V} = \sum_{0 \leq j \leq i \leq N} \frac{t_{ij}}{L} \langle \nabla f(y_i), \nabla f(y_j) \rangle = \frac{v_0}{2L} \|\nabla f(y_N)\|^2 + \sum_{i=0}^{N-1} \frac{v_{i+1} - v_i}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2.$$

### B.2.3 Calculation of energy function of GD

We calculate  $\{t_{ij}\}$  first. Recall that  $\{v_i\}_{i=0}^N = \left(\frac{1}{2N+1}, \dots, \frac{N+i}{(2N+1)(N-i+1)}, \dots\right)$  and  $h_{i+1,k} = \delta_{i,k}$  to (35), and making  $\{t_{ij}\}$  symmetric gives us

$$t_{ij} = \begin{cases} \frac{1}{2}v_0 & i = j, i = 0 \\ v_i & i = j, 1 \leq i \leq N-1 \\ v_N - \frac{1}{2} & i = j, i = N \\ \frac{1}{2}(v_{\min(i,j)} - v_{\min(i,j)+1}) & i \neq j. \end{cases}$$

We can verify that the matrix  $\{t_{ij}\}_{0 \leq i, j \leq N}$  is diagonally dominant:  $t_{ii} = |\sum_{j \neq i} t_{ij}|$ . Therefore,

$\sum_{0 \leq i, j \leq N-1} \frac{t_{ij}}{L} \langle \nabla f(y_i), \nabla f(y_j) \rangle \geq 0$  for any  $\{\nabla f(y_i)\}_{i=0}^N$ . This proof is essentially the same as the proof in [29], but we repeat it here with our notation for the sake of completeness.

Next, we prove that (GD) with  $h = 1$  and  $\{u_i\}_{i=0}^N = \left(\dots, \frac{(2N+1)(i+1)}{2N-i}, \dots, 2N+1\right)$  satisfies (C1), by showing more general statement:

$$\left[ \text{(GD) and } \{u_i\}_{i=0}^N = \left(\dots, \frac{(2Nh+1)(i+1)}{2N-i}, \dots, 2Nh+1\right) \text{ satisfies (C1)} \right] \quad (39)$$

Note that (39) gives

$$(2Nh+1)(f(x_N) - f^*) \leq \mathcal{U}_N \leq \mathcal{U}_{-1} = \frac{L}{2} \|x_0 - x^*\|^2. \quad (40)$$

Later in the proof of Corollary 2, we will utilize the equation (39). The result (39) is proved in [19, Theorem 3.1], and we give the proof outline here.

In order to demonstrate (39), we will directly expand the expression  $\mathcal{U}_N - u_N (f(x_N) - f^*)$ , instead of employing  $\mathbf{U}$  as a intermediary step. Define  $\{u'_i\}_{i=0}^N := \left(\dots, \frac{(2N+1)(i+1)}{2N-i}, \dots, 2N+1\right)$ . Then

$$\mathcal{U}_N - u_N (f(x_N) - f^*) = \frac{1}{L} \text{Tr}(g^\top g S)$$

where  $g = [\nabla f(x_0) | \dots | \nabla f(x_N) | L(x_0 - x^*)] \in \mathbb{R}^{d \times (N+2)}$  and  $S \in \mathbb{S}^{N+2}$  is given by

$$S = \begin{bmatrix} S' & \lambda \\ \lambda & \frac{1}{2} \end{bmatrix},$$

$$\lambda = [u_0 | u_1 - u_0 | \dots | u_N - u_{N-1}]^\top, \quad S' = \frac{2Nh+1}{2N+1} (hS_0 + (1-h)S_1),$$

$$\begin{aligned} S_0 &= \sum_{i=0}^{N-1} \frac{u'_i}{2} (e_{i+1}e_i^\top + e_i e_{i+1}^\top + (e_i - e_{i+1})(e_i - e_{i+1})^\top) \\ &\quad + \sum_{i=0}^N \frac{u'_i - u'_{i-1}}{2} (e_i(e_0 + \dots + e_{i-1})^\top + (e_0 + \dots + e_{i-1})e_i^\top + e_i e_i^\top) \\ S_1 &= \sum_{i=0}^{N-1} \frac{u'_i}{2} (e_i - e_{i+1})(e_i - e_{i+1})^\top + \sum_{i=0}^N \frac{u'_i - u'_{i-1}}{2} e_i e_i^\top. \end{aligned}$$

Now we will show  $S \succeq 0$  to obtain  $[\mathcal{U}_N - u_N (f(x_N) - f^*) \geq 0, \forall g]$ , which is (C1). By using Sylvester's Criterion,  $S_0 \succ 0$  follows.  $S_1 \succ 0$  follows from the fact that  $S_1$  expressed by the sum of positive semi-definite matrices  $zz^\top$ . Since the convex sum of two positive semi-definite matrices is also positive semi-definite,  $S' = hS_0 + (1-h)S_1 \succ 0$ .

Next, we argue that  $\det S = 0$ . Indeed, take  $\tau = (1, \dots, -(2Nh+1))^\top$  to show  $S\tau = 0$ , which gives  $\det S = 0$ . Note that the determinant of  $S$  can also be expressed by

$$\det(S) = \left(\frac{1}{2} - \lambda^\top (S')^{-1} \lambda\right) \det(S'). \quad (41)$$

We have shown that  $S' \succ 0$ , (41) implies  $\frac{1}{2} - \lambda^\top (S')^{-1} \lambda = 0$ , which is the Schur complement of the matrix  $S$ . By a well-known lemma on the Schur complement, we conclude  $S \succeq 0$ .

### B.3 Omitted proof in Section 2.5

#### B.3.1 Omitted calculation of Corollary 1

Here, we will give the general formulation of H-dual FSFOM of

$$x_{k+1} = x_k + \beta_k (x_k^+ - x_{k-1}^+) + \gamma_k (x_k^+ - x_k), \quad k = 0, \dots, N-1. \quad (42)$$

**Proposition 1.** The H-dual of (42) is

$$y_{k+1} = y_k + \beta'_k (y_k^+ - y_{k-1}^+) + \gamma'_k (y_k^+ - y_k), \quad k = 0, \dots, N-1 \quad (43)$$

where

$$\beta'_k = \frac{\beta_{N-k}(\beta_{N-1-k} + \gamma_{N-1-k})}{\beta_{N-k} + \gamma_{N-k}}$$

$$\gamma'_k = \frac{\gamma_{N-k}(\beta_{N-1-k} + \gamma_{N-1-k})}{\beta_{N-k} + \gamma_{N-k}}$$

for  $k = 0, \dots, N-1$  and  $(\beta_N, \gamma_N)$  is any value that  $\beta_N + \gamma_N \neq 0$ .<sup>3</sup>

*Proof.* The  $H$  matrix  $\{h_{k,i}\}_{0 \leq i < k \leq N}$  satisfies

$$h_{k+1,i} = \begin{cases} 1 + \beta_k + \gamma_k & i = k, k = 0, \dots, N-1 \\ \beta_k (h_{k,i} - 1) & i = k-1, k = 1, \dots, N-1 \\ \beta_k h_{k,i} & i = k-2, k = 2, \dots, N-1 \end{cases}.$$

Therefore,

$$h_{k+1,i} = \left( \prod_{j=i+1}^k \beta_j \right) (\beta_i + \gamma_i + \delta_{k,i})$$

where  $\delta_{k,i}$  is a Kronecker Delta function. Similarly,  $H$  matrix of (43)  $\{g_{k,i}\}_{0 \leq i < k \leq N}$  satisfies

$$\begin{aligned} g_{k+1,i} &= \left( \prod_{j=i+1}^k \beta'_j \right) (\beta'_i + \gamma'_i + \delta_{k,i}) \\ &= \left( \prod_{j=i+1}^k \frac{\beta_{N-j}(\beta_{N-1-j} + \gamma_{N-1-j})}{\beta_{N-j} + \gamma_{N-j}} \right) (\beta_{N-1-i} + \gamma_{N-1-i} + \delta_{k,i}) \\ &= \left( \prod_{j=i+1}^k \beta_{N-j} \right) (\beta_{N-k-1} + \gamma_{N-i-1} + \delta_{N-k-1, N-i-1}). \end{aligned}$$

Thus  $g_{k+1,i} = h_{N-i, N-1-k}$ . □

Now we derive the H-dual of (6) by applying Proposition 1. Note that

$$\beta_k = \frac{(T_k - t_k)t_{k+1}}{t_k T_{k+1}}, \quad \gamma_k = \frac{(t_k^2 - T_k)t_{k+1}}{t_k T_{k+1}}, \quad k = 0, \dots, N-1.$$

Next, define  $\beta_N$  and  $\gamma_N$  as a same form of  $\{\beta_i, \gamma_i\}_{i=0}^{N-1}$  with any  $t_{N+1} > 0$ . Note that

$$\beta_k + \gamma_k = \frac{t_{k+1}(t_k^2 - t_k)}{t_k T_{k+1}} = \frac{t_{k+1}(t_k - 1)}{T_{k+1}}.$$

By applying the formula at Proposition 1, we obtain

$$\beta'_k = \frac{(T_{N-k} - t_{N-k}) \frac{t_{N-k}(t_{N-k-1}-1)}{T_{N-k}}}{t_{N-k}^2 - t_{N-k}} = \frac{(T_{N-k} - t_{N-k})(t_{N-k-1} - 1)}{T_{N-k}(t_{N-k} - 1)} = \frac{T_{N-k-1}(t_{N-k-1} - 1)}{T_{N-k}(t_{N-k} - 1)}$$

and

$$\gamma'_k = \frac{(t_{N-k}^2 - T_{N-k}) \frac{t_{N-k}(t_{N-k-1}-1)}{T_{N-k}}}{t_{N-k}^2 - t_{N-k}} = \frac{(t_{N-k}^2 - T_{N-k})(t_{N-k-1} - 1)}{T_{N-k}(t_{N-k} - 1)}.$$

<sup>3</sup>Here, note that FSFOM (43) is independent with the any choice of  $\beta_N, \gamma_N$  since  $y_1 = y_0 + (\beta'_0 + \gamma'_0)(y_0^+ - y_0) = y_0 + (\beta_{N-1} + \gamma_{N-1})(y_0^+ - y_0)$ .

### B.3.2 Proof of Corollary 2

First, recall (39) when  $0 < h \leq 1$ :

$$\left[ \text{(GD) and } \{u_i\}_{i=0}^N = \left( \dots, \frac{(2Nh+1)(i+1)}{2N-i}, \dots, 2Nh+1 \right) \text{ satisfies (C1)} \right]$$

Additionally, observe that the  $H$  matrix of (GD) is  $\text{diag}(h, \dots, h)$ , which gives the fact that the H-dual of (GD) is itself.

Next, use Theorem 1 to obtain

$$\left[ \text{(GD) with } 0 < h \leq 1 \text{ and } \{v_i\}_{i=0}^N = \left( \frac{1}{2Nh+1}, \dots, \frac{N+i}{(2Nh+1)(N-i+1)}, \dots \right) \text{ satisfies (C2)} \right]. \quad (44)$$

By using the same argument with (5), (44) gives

$$\frac{1}{2L} \|\nabla f(y_N)\|^2 \leq \mathcal{V}_N \leq \mathcal{V}_0 = \frac{1}{2Nh+1} (f(y_0) - f_\star) + \frac{1}{2Nh+1} \llbracket y_N, y_\star \rrbracket \leq \frac{1}{2Nh+1} (f(y_0) - f_\star). \quad (45)$$

In addition, we can achieve the convergence rate of the gradient norm under the initial condition of  $\|x_0 - x_\star\|^2$ :

$$\frac{1}{2L} \|\nabla f(x_{2N})\|^2 \leq \frac{1}{2Nh+1} (f(x_N) - f_\star) \leq \frac{1}{(2Nh+1)^2} \frac{L}{2} \|x_0 - x_\star\|^2$$

and

$$\frac{1}{2L} \|\nabla f(x_{2N+1})\|^2 \leq \frac{1}{2(N+1)h+1} (f(x_N) - f_\star) \leq \frac{1}{(2(N+1)h+1)(2Nh+1)} \frac{L}{2} \|x_0 - x_\star\|^2.$$

The first inequality comes from (45) and the second inequality comes from (40).

### B.3.3 Proof of $\mathcal{A}_\star$ -optimality of (OGM-G) and (OBL-G<sub>b</sub>)

**Definition of  $\mathcal{A}^\star$ -optimal FSFOM.** For the given inequality sets  $\mathcal{L}$ ,  $\mathcal{A}^\star$ -optimal FSFOM with respect to  $[\mathcal{L}, \text{P1}]$  is defined as a FSFOM which  $H$  matrix is the solution of following minimax problem:

$$\begin{aligned} & \underset{H \in \mathbb{R}_L^{N \times N}}{\text{minimize}} \quad \underset{f}{\text{maximize}} \quad f(x_N) - f_\star \\ & \text{subject to.} \quad [x_0, \dots, x_N \text{ are generated by FSFOM with the matrix } H] \\ & \quad [\forall l \in \mathcal{L}, f \text{ satisfies } l] \\ & \quad \|x_0 - x_\star\|^2 \leq R^2 \end{aligned} \quad (46)$$

Similarly,  $\mathcal{A}^\star$ -optimal FSFOM with respect to  $[\mathcal{L}, (\text{P2})]$  is specified with its  $H$  matrix, which is the solution of the following minimax problem:

$$\begin{aligned} & \underset{H \in \mathbb{R}_L^{N \times N}}{\text{minimize}} \quad \underset{f}{\text{maximize}} \quad \frac{1}{2L} \|\nabla f(y_N)\|^2 \\ & \text{subject to.} \quad [y_0, \dots, y_N \text{ are generated by FSFOM with the matrix } H] \\ & \quad [\forall l \in \mathcal{L}, f \text{ satisfies } l] \\ & \quad f(y_0) - f_\star \leq \frac{1}{2} LR^2 \end{aligned} \quad (47)$$

Here  $\mathbb{R}_L^{N \times N}$  is the set of lower triangular matrices. Next we denote the inner maximize problem of (46) and (47) as  $\mathcal{P}_1(\mathcal{L}, H, R)$  and  $\mathcal{P}_2(\mathcal{L}, H, R)$ , respectively. For a more rigorous definition of  $\mathcal{A}^\star$ -optimal FSFOM, refer [44].

**Remark.** We discuss the minimax problems (46) and (47) and their interpretation. Specifically, we consider the meaning of the maximization problem in these minimax problems, which can be thought of as calculating the worst-case performance for a fixed FSFOM with  $H$ . In other words, the minimization problem in (46) and (47) can be interpreted as determining the optimal value of  $H$  that minimizes the worst-case performance.

**Prior works.** The prior works for  $\mathcal{A}^*$ -optimality are summarized as follows. Consider the following sets of inequalities.

$$\begin{aligned}
\mathcal{L}_F &= \{ \llbracket x_i, x_{i+1} \rrbracket \}_{i=0}^{N-1} \cup \{ \llbracket x_\star, x_i \rrbracket \}_{i=0}^N \\
&= \left\{ f(x_i) \geq f(x_{i+1}) + \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle + \frac{1}{2L} \|\nabla f(x_i) - \nabla f(x_{i+1})\|^2 \right\}_{i=0}^{N-1} \\
&\quad \bigcup \left\{ f_\star \geq f(x_i) + \langle \nabla f(x_i), x_\star - x_i \rangle + \frac{1}{2L} \|\nabla f(x_i)\|^2 \right\}_{i=0}^N, \\
\mathcal{L}_G &= \{ \llbracket y_i, y_{i+1} \rrbracket \}_{i=0}^{N-1} \cup \{ \llbracket y_N, y_i \rrbracket \}_{i=0}^{N-1} \cup \{ \llbracket y_N, \star \rrbracket \} \\
&= \left\{ f(y_i) \geq f(y_{i+1}) + \langle \nabla f(y_{i+1}), y_i - y_{i+1} \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_{i+1})\|^2 \right\}_{i=0}^{N-1} \\
&\quad \bigcup \left\{ f(y_N) \geq f(y_i) + \langle \nabla f(y_i), y_N - y_i \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2 \right\}_{i=0}^{N-1} \\
&\quad \bigcup \left\{ f(y_N) \geq f_\star + \frac{1}{2L} \|\nabla f(y_N)\|^2 \right\}, \\
\mathcal{L}_{G'} &= \{ \llbracket y_i, y_{i+1} \rrbracket \}_{i=0}^{N-1} \cup \left\{ \llbracket y_N, y_i \rrbracket - \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2 \right\}_{i=0}^{N-1} \cup \left\{ \llbracket y_N, \star \rrbracket - \frac{1}{2L} \|\nabla f(y_N)\|^2 \right\} \\
&= \left\{ f(y_i) \geq f(y_{i+1}) + \langle \nabla f(y_{i+1}), y_i - y_{i+1} \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_{i+1})\|^2 \right\}_{i=0}^{N-1} \\
&\quad \bigcup \left\{ f(y_N) \geq f(y_i) + \langle \nabla f(y_i), y_N - y_i \rangle \right\}_{i=0}^{N-1} \bigcup \left\{ f(y_N) \geq f_\star \right\}, \\
\mathcal{L}_{F'} &= \{ \llbracket x_i, x_{i+1} \rrbracket \}_{i=0}^{N-1} \cup \left\{ \llbracket x_\star, x_i \rrbracket - \frac{1}{2L} \|\nabla f(x_i)\|^2 \right\}_{i=0}^N \\
&= \left\{ f(x_i) \geq f(x_{i+1}) + \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle + \frac{1}{2L} \|\nabla f(x_i) - \nabla f(x_{i+1})\|^2 \right\}_{i=0}^{N-1} \\
&\quad \bigcup \left\{ f_\star \geq f(x_i) + \langle \nabla f(x_i), x_\star - x_i \rangle \right\}_{i=0}^N,
\end{aligned}$$

and

$$\mathcal{L}_{\text{exact}} = \{ \llbracket x_i, x_j \rrbracket \}_{(i,j) \in \{\star, 0, 1, \dots, N\}^2}.$$

(OGM) is  $\mathcal{A}^*$ -optimal with respect to  $[\mathcal{L}_F, (\text{P1})]$  [26]. Furthermore, (OGM) is also  $\mathcal{A}^*$ -optimal under  $[\mathcal{L}_{\text{exact}}, (\text{P1})]$  which implies that (OGM) is the exact optimal FSFOM with respect to (P1). In addition, the  $\mathcal{A}^*$ -optimality of (OGM-G) with respect to  $[\mathcal{L}_G, (\text{P2})]$  was presented as a conjecture in [29], (OBL-F<sub>b</sub>) is  $\mathcal{A}^*$ -optimal with respect to  $[\mathcal{L}_{F'}, (\text{P1})]$  [44, Theorem 4], and the  $\mathcal{A}^*$ -optimality of (OBL-G<sub>b</sub>) with respect to  $[\mathcal{L}_{G'}, (\text{P2})]$  was presented as a conjecture [44, Conjecture 8].<sup>4</sup>

In the remaining parts, we give the proof of the following two theorems.

**Theorem 4.** (OGM-G) is  $\mathcal{A}^*$ -optimal with respect to  $[\mathcal{L}_G, (\text{P2})]$ .

**Theorem 5.** (OBL-G<sub>b</sub>) is  $\mathcal{A}^*$ -optimal with respect to  $[\mathcal{L}_{G'}, (\text{P2})]$ .

<sup>4</sup>Originally, the inequality set suggested that (OBL-G<sub>b</sub>) would  $\mathcal{A}^*$ -optimal is which  $(f(y_N) \geq f_\star)$  is replaced with  $(f(y_N) \geq f_\star + \frac{1}{2L} \|\nabla f(y_N)\|^2)$  in  $\mathcal{L}_{G'}$ .

**Proof of  $\mathcal{A}^*$ -optimality of (OGM-G)** We provide an alternative formulation of  $\mathcal{P}_2(\mathcal{L}_G, H, R)$  by using the methodology called PEP [19, 52]

$$\begin{aligned}
& \underset{f}{\text{maximize}} && \frac{1}{2L} \|\nabla f(y_N)\|^2 \\
& \text{subject to} && f(y_{i+1}) - f(y_i) + \langle \nabla f(y_{i+1}), y_i - y_{i+1} \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_{i+1})\|^2 \leq 0, \quad i = 0, \dots, N-1 \\
& && f(y_i) - f(y_N) + \langle \nabla f(y_i), y_N - y_i \rangle + \frac{1}{2L} \|\nabla f(y_i) - \nabla f(y_N)\|^2 \leq 0, \quad i = 0, \dots, N-1 \\
& && f_\star + \frac{1}{2L} \|\nabla f(y_N)\|^2 \leq f(y_N) \\
& && f(y_0) - f_\star \leq \frac{1}{2}LR^2 \\
& && y_{i+1} = y_i - \frac{1}{L} \sum_{j=0}^i h_{i+1,j} \nabla f(y_j), \quad i = 0, \dots, N-1
\end{aligned} \tag{48}$$

Next, define  $\mathbf{F}$ ,  $\mathbf{G}$ ,  $\{\mathbf{e}_i\}_{i=0}^N$  and  $\{\mathbf{x}_i\}_{i=0}^N$  as

$$\mathbf{G} := \begin{pmatrix} \langle \nabla f(y_0), \nabla f(y_0) \rangle & \langle \nabla f(y_0), \nabla f(y_1) \rangle & \cdots & \langle \nabla f(y_0), \nabla f(y_N) \rangle \\ \vdots & \vdots & \vdots & \vdots \\ \langle \nabla f(y_N), \nabla f(y_N) \rangle & \langle \nabla f(y_N), \nabla f(y_1) \rangle & \cdots & \langle \nabla f(y_N), \nabla f(y_N) \rangle \end{pmatrix} \in \mathbb{S}^{N+1},$$

$$\mathbf{F} := \begin{pmatrix} f(y_0) - f_\star \\ f(y_1) - f_\star \\ \vdots \\ f(y_N) - f_\star \end{pmatrix} \in \mathbb{R}^{(N+1) \times 1},$$

$\mathbf{e}_i \in \mathbb{R}^{N+1}$  is a unit vector which  $(i+1)$ -th component is 1, and

$$\mathbf{y}_0 := 0, \quad \mathbf{y}_{i+1} := \mathbf{y}_i - \frac{1}{L} \sum_{j=0}^i h_{i+1,j} \mathbf{e}_j \quad i = 0, \dots, N-1.$$

By using the definition of  $\mathbf{F}$ ,  $\mathbf{G}$ ,  $\{\mathbf{e}_i\}_{i=0}^N$  and  $\{\mathbf{y}_i\}_{i=0}^N$ ,  $\mathcal{P}_2(\mathcal{L}_G, H, R)$  can be converted into

$$\begin{aligned}
& \underset{\mathbf{F}, \mathbf{G} \succeq 0}{\text{minimize}} && -\frac{1}{2L} \text{Tr}(\mathbf{G}\mathbf{e}_N\mathbf{e}_N^\top) \\
& \text{subject to} && \mathbf{F}(\mathbf{e}_{i+1} - \mathbf{e}_i)^\top + \text{Tr}(\mathbf{G}\mathbf{A}_i) \leq 0, \quad i = 0, \dots, N-1 \\
& && \mathbf{F}(\mathbf{e}_i - \mathbf{e}_N)^\top + \text{Tr}(\mathbf{G}\mathbf{B}_i) \leq 0, \quad i = 0, \dots, N-1 \\
& && -\mathbf{F}\mathbf{e}_N^\top + \frac{1}{2} \text{Tr}(\mathbf{G}\mathbf{e}_N\mathbf{e}_N^\top) \leq 0 \\
& && \mathbf{F}\mathbf{e}_0^\top - \frac{1}{2}LR^2 \leq 0
\end{aligned} \tag{49}$$

where

$$\begin{aligned}
\mathbf{A}_i &:= \frac{1}{2} \mathbf{e}_{i+1}(\mathbf{y}_i - \mathbf{y}_{i+1})^\top + \frac{1}{2}(\mathbf{y}_i - \mathbf{y}_{i+1})\mathbf{e}_{i+1}^\top + \frac{1}{2L}(\mathbf{e}_i - \mathbf{e}_{i+1})(\mathbf{e}_i - \mathbf{e}_{i+1})^\top \\
\mathbf{B}_i &:= \frac{1}{2} \mathbf{e}_i(\mathbf{y}_N - \mathbf{y}_i)^\top + \frac{1}{2}(\mathbf{y}_N - \mathbf{y}_i)\mathbf{e}_i^\top + \frac{1}{2L}(\mathbf{e}_i - \mathbf{e}_N)(\mathbf{e}_i - \mathbf{e}_N)^\top.
\end{aligned}$$

Moreover, under the condition  $d \geq N+2$ , we can take the Cholesky factorization of  $\mathbf{G}$  to recover the triplet  $\{(y_i, f(y_i), \nabla f(y_i))\}_{i=0}^N$ .<sup>5</sup> Thus (48) and (49) are equivalent. The next step is calculating the Lagrangian of (49) and deriving the Lagrangian dual problem of it. In [54], they argued about the strong duality of (49).

<sup>5</sup>Cholesky factorization is unique if  $\mathbf{G} > 0$  but it may be not unique if  $\mathbf{G} \succeq 0$ . In this case, we choose one representation of Cholesky factorization.

**Fact 1.** Assume  $h_{i+1,i} \neq 0$  for  $0 \leq i \leq N-1$ . Then the strong duality holds between (48) and (49).

Denote the dual variables of each constraints as  $\{\delta_i\}_{i=0}^{N-1}$ ,  $\{\lambda_i\}_{i=0}^{N-1}$ ,  $\delta_N$  and  $\tau$ . Then Lagrangian becomes

$$\mathcal{L}(\delta, \lambda, \tau, \mathbf{F}, \mathbf{G}) = \mathbf{F} \cdot \mathbf{X}^\top + \text{Tr}(\mathbf{G} \cdot \mathbf{T}) - \frac{\tau L}{2} R^2$$

where

$$\begin{aligned} \mathbf{X} &= \sum_{i=0}^{N-1} \delta_i (\mathbf{e}_{i+1} - \mathbf{e}_i) + \sum_{i=0}^{N-1} \lambda_i (\mathbf{e}_i - \mathbf{e}_N) - \delta_N \mathbf{e}_N + \tau \mathbf{e}_0, \\ \mathbf{T} &= \sum_{i=0}^{N-1} \delta_i \mathbf{A}_i + \sum_{i=0}^{N-1} \lambda_i \mathbf{B}_i + \frac{\delta_N}{2} \mathbf{e}_N \mathbf{e}_N^\top - \frac{1}{2} \mathbf{e}_N \mathbf{e}_N^\top. \end{aligned}$$

If  $[\mathbf{X} = 0 \text{ and } \mathbf{T} \succeq 0]$  is false, we can choose  $\mathbf{F}$  and  $\mathbf{G}$  that makes the value of Lagrangian to be  $-\infty$ . Thus the convex dual problem of (49) is

$$\underset{\delta, \lambda, \tau}{\text{maximize}} \underset{\mathbf{F}, \mathbf{G}}{\text{minimize}} \mathcal{L}(\delta, \lambda, \tau, \mathbf{F}, \mathbf{G}) = \begin{cases} \underset{\delta, \lambda, \tau}{\text{maximize}} & -\frac{\tau L}{2} R^2 \\ \text{subject to} & \mathbf{X} = 0, \mathbf{T} \succeq 0 \\ & \delta_i \geq 0, \lambda_i \geq 0, \tau \geq 0 \end{cases}. \quad (50)$$

For the constraint  $\mathbf{X} = 0$ ,  $\lambda_i = \delta_i - \delta_{i-1}$  for  $1 \leq i \leq N-1$ ,  $\tau = \delta_N$  and  $-\delta_0 + \lambda_0 + \delta_N = 0$ . By substituting  $v_{i+1} = \delta_i$  for  $0 \leq i \leq N-1$  and  $v_0 = \delta_N$ , (50) becomes

$$\underset{v_0, \dots, v_N}{\text{maximize}} \underset{\mathbf{F}, \mathbf{G}}{\text{minimize}} \mathcal{L}(\delta, \lambda, \tau, \mathbf{F}, \mathbf{G}) = \begin{cases} \underset{\delta, \lambda, \tau}{\text{maximize}} & -\frac{v_0 L}{2} R^2 \\ \text{subject to} & \mathbf{T} \succeq 0 \\ & 0 \leq v_0 \leq \dots \leq v_N \end{cases}. \quad (51)$$

Therefore if strong duality holds,  $\mathcal{P}_2(\mathcal{L}_G, H, R)$  becomes

$$\begin{aligned} &\underset{v_0, \dots, v_N}{\text{minimize}} \frac{v_0 L}{2} R^2 \\ &\text{subject to. } \mathbf{T} \succeq 0, 0 \leq v_0 \leq \dots \leq v_N. \end{aligned} \quad (52)$$

We can apply a similar argument for  $\mathcal{P}_1(\mathcal{L}_F, H, R)$ . To begin with, define  $\{\mathbf{f}_i\}_{i=-1}^N$  as a unit vector of length  $N+2$  which  $(i+2)$ -component is 1. Additionally, define  $\{\mathbf{x}_i\}_{i=0}^N$ ,  $\{\mathbf{C}_i\}_{i=0}^{N-1}$  and  $\{\mathbf{D}_i\}_{i=0}^N$  as follows:

$$\begin{aligned} \mathbf{x}_0 &:= \mathbf{f}_{-1}, \quad \mathbf{x}_{i+1} := \mathbf{x}_i - \frac{1}{L} \sum_{j=0}^i h_{i+1,j} \mathbf{f}_j \quad i = 0, \dots, N-1, \\ \mathbf{C}_i &:= \frac{1}{2} \mathbf{f}_{i+1} (\mathbf{x}_i - \mathbf{x}_{i+1})^\top + \frac{1}{2} (\mathbf{x}_i - \mathbf{x}_{i+1}) \mathbf{f}_{i+1}^\top + \frac{1}{2L} (\mathbf{f}_i - \mathbf{f}_{i+1}) (\mathbf{f}_i - \mathbf{f}_{i+1})^\top, \\ \mathbf{D}_i &:= -\frac{1}{2} \mathbf{f}_i \mathbf{x}_i^\top - \frac{1}{2} \mathbf{x}_i \mathbf{f}_i^\top + \frac{1}{2L} \mathbf{f}_i \mathbf{f}_i^\top. \end{aligned} \quad (53)$$

Then, if the strong duality holds, the problem

$$\begin{aligned} &\underset{f}{\text{maximize}} f(x_N) - f_\star \\ &\text{subject to. } [x_0, \dots, x_N \text{ are generated by FSFO with the matrix } H] \\ &\quad [\forall l \in \mathcal{L}_F, f \text{ satisfies } l] \\ &\quad \|x_0 - x_\star\|^2 \leq R^2 \end{aligned} \quad (54)$$

is equivalent to

$$\begin{aligned} &\underset{u_0, \dots, u_N}{\text{maximize}} \frac{L}{2u_N} R^2 \\ &\text{subject to. } \mathbf{S} \succeq 0, 0 \leq u_0 \leq \dots \leq u_N, \end{aligned} \quad (55)$$

where

$$\mathbf{S} = \frac{L}{2} \mathbf{f}_{-1} \mathbf{f}_{-1}^\top + \sum_{i=0}^{N-1} u_i \mathbf{C}_i + \sum_{i=0}^N (u_i - u_{i-1}) \mathbf{D}_i.$$

By using Schur's Complement,

$$\mathbf{S} \succeq 0 \quad \Leftrightarrow \quad \mathbf{S}' \succeq 0$$

where

$$\mathbf{S}' = \sum_{i=0}^{N-1} u_i \mathbf{C}_i + \sum_{i=0}^N (u_i - u_{i-1}) \mathbf{D}_i - \frac{1}{2L} \left( \sum_{i=0}^N (u_i - u_{i-1}) \mathbf{f}_i \right) \left( \sum_{i=0}^N (u_i - u_{i-1}) \mathbf{f}_i \right)^\top.$$

Hence (55) is equivalent to

$$\begin{aligned} & \underset{u_0, \dots, u_N}{\text{maximize}} \quad \frac{L}{2u_N} R^2 \\ & \text{subject to.} \quad \mathbf{S}' \succeq 0, \quad 0 \leq u_0 \leq \dots \leq u_N. \end{aligned} \quad (56)$$

Now we will prove the following proposition.

**Proposition 2.** Consider a matrix  $H = \{h_{i,j}\}_{0 \leq j < i \leq N}$  and  $H^A$ . If  $h_{i+1,i} \neq 0$  for  $0 \leq i \leq N-1$  and treating the solution of infeasible maximize problem as 0, the optimal values of  $\mathcal{P}_1(\mathcal{L}_F, H, R)$  and  $\mathcal{P}_2(\mathcal{L}_G, H^A, R)$  are same.

*Proof.* To simplify the analysis, we can consider  $\{\mathbf{f}_i\}_{i=0}^N$  as a length  $N+1$  unit vector, as all terms with  $\mathbf{f}_{-1}$  can be eliminated by using  $\mathbf{S}'$  instead of  $\mathbf{S}$ . With this simplification, both  $\mathbf{S}'$  and  $\mathbf{T}$  belong to  $\mathbb{S}^{N+1}$ .

Next, let  $0 < u_0 \leq u_1 \leq \dots \leq u_N$  and  $v_i = \frac{1}{u_{N-i}}$  for  $0 \leq i \leq N$ , noting that  $0 < v_0 \leq \dots \leq v_N$ . It is important to observe that  $\mathbf{S}'$  and  $\mathbf{T}$  can be expressed in terms of  $\mathcal{S}(H, u)$  and  $\mathcal{T}(H^A, v)$ , respectively. Furthermore, in the proof of Theorem 1 (A), we proved that  $\mathcal{S}(H, u) = \mathcal{M}(u)^\top \mathcal{T}(H^A, v) \mathcal{M}(u)$  for some invertible  $\mathcal{M}(u)$ . Thus,

$$\mathcal{S}(H, u) \succeq 0 \quad \Leftrightarrow \quad \mathcal{T}(H^A, v) \succeq 0.$$

Therefore, we obtain

$$(a_0, \dots, a_N) \in \{(u_0, \dots, u_N) \mid \mathbf{S}' \succeq 0, 0 < u_0 \leq \dots \leq u_N\} \quad (57)$$

if and only if

$$\left( \frac{1}{a_N}, \dots, \frac{1}{a_0} \right) \in \{(v_0, \dots, v_N) \mid \mathbf{T} \succeq 0, 0 < v_0 \leq \dots \leq v_N\}. \quad (58)$$

For the next step, we claim that the optimal value of (56) and

$$\begin{aligned} & \underset{u_0, \dots, u_N}{\text{maximize}} \quad \frac{L}{2u_N} R^2 \\ & \text{subject to.} \quad \mathbf{S}' \succeq 0, \quad 0 < u_0 \leq \dots \leq u_N \end{aligned} \quad (59)$$

are same, i.e., consider the case when all  $u_i$  are positive is enough. To prove that, assume there exists  $0 = u_0 = \dots = u_k, 0 < u_{k+1} \leq \dots \leq u_N$  and  $H$  that satisfy  $\mathbf{S}' \succeq 0$ . Next, observe that the  $\mathbf{f}_k \mathbf{f}_k^\top$  component of  $\mathbf{S}'$  is 0 but  $\mathbf{f}_{k+1} \mathbf{f}_k^\top$  component of  $\mathbf{S}'$  is  $u_{k+1} h_{k+1,k} \neq 0$ , which makes  $\mathbf{S}' \succeq 0$  impossible.

When the optimal value of (52) is 0, it implies  $\{(v_0, \dots, v_N) \mid \mathbf{T} \succeq 0, 0 < v_0 \leq \dots \leq v_N\}$  is an empty set. Therefore,  $\{(u_0, \dots, u_N) \mid \mathbf{S}' \succeq 0, 0 < u_0 \leq \dots \leq u_N\}$  is also empty set. Since (59) and (56) have the same optimal value, the optimal value of (56) is 0.

If the optimal value of (52) is positive, the optimal values of (56) and (52) are the same since (57) and (58) are equivalent.  $\square$

*Proof of Theorem 4.*  $H_{\text{OGM}}$  is the solution of (46) since (OGM) is  $\mathcal{A}^*$ -optimal with respect to  $\mathcal{L}_F$  and (P1). Additionally, if

$$(46) = \underset{H, h_{i+1}, i \neq 0}{\text{maximize}} \mathcal{P}_1(\mathcal{L}_F, H, R) \quad (60)$$

holds,  $H_{\text{OGM}}$  is the solution to the following problem due to the strong duality.

$$\begin{aligned} & \underset{H, h_{i+1}, i \neq 0}{\text{maximize}} \underset{u_0, \dots, u_N}{\text{maximize}} \frac{L}{2u_N} R^2 \\ & \text{subject to. } \mathbf{S} \succeq 0, \quad 0 \leq u_0 \leq \dots \leq u_N. \end{aligned} \quad (61)$$

Applying Proposition 2 and using the fact  $H_{\text{OGM-G}} = H_{\text{OGM}}^A$  provides that  $H_{\text{OGM-G}}$  is the solution of

$$\begin{aligned} & \underset{H, h_{i+1}, i \neq 0}{\text{maximize}} \underset{v_0, \dots, v_N}{\text{minimize}} \frac{v_0 L}{2} R^2 \\ & \text{subject to. } \mathbf{T} \succeq 0, \quad 0 \leq v_0 \leq \dots \leq v_N. \end{aligned} \quad (62)$$

Finally, if

$$(47) = \underset{H, h_{i+1}, i \neq 0}{\text{maximize}} \mathcal{P}_1(\mathcal{L}_F, H, R) \quad (63)$$

holds, the optimal solution of (47) is the  $H_{\text{OGM-G}}$ , which proves the  $\mathcal{A}^*$ -optimality of (OGM-G) with respect to  $\mathcal{L}_G$  and (P2). The proof of (60) and (63) uses the continuity argument with  $H$  and please refer [44, Claim 4].  $\square$

**Remark of Proof of  $\mathcal{A}^*$ -optimality of (OGM-G).** We proved that (OGM-G) is  $\mathcal{A}^*$ -optimal if we use the subset of cocoercivity inequalities. Therefore, it is still open whether (OGM-G) is optimal or not among the  $L$ -smooth convex function's gradient minimization method.

**Proof of  $\mathcal{A}^*$ -optimality of (OBL-G<sub>b</sub>).** We provide the proof that the  $H$  matrix of (OBL-G<sub>b</sub>) is the solution of

$$\underset{H \in \mathbb{R}_L^{N \times N}}{\text{minimize}} \mathcal{P}_2(\mathcal{L}_G, H, R) \quad (64)$$

To begin with, bring to mind the  $\mathcal{A}^*$ -optimality of (OBL-F<sub>b</sub>): (OBL-F<sub>b</sub>) is  $\mathcal{A}^*$ -optimal with respect to the  $[\mathcal{L}_{F'}, P1]$ , i.e., the  $H$  matrix of (OBL-F<sub>b</sub>) is the solution of

$$\underset{H \in \mathbb{R}_L^{N \times N}}{\text{minimize}} \mathcal{P}_1(\mathcal{L}_F, H, R). \quad (65)$$

To prove the  $\mathcal{A}^*$ -optimality of (OBL-G<sub>b</sub>), we use the  $\mathcal{A}^*$ -optimality of (OBL-F<sub>b</sub>).

Under the assumption of strong duality, we could change  $\mathcal{P}_1(\mathcal{L}_F, H, R)$  into the following SDP:

$$\begin{aligned} & \underset{u_0, \dots, u_N}{\text{maximize}} \frac{L}{2u_N} R^2 \\ & \text{subject to. } \mathbf{S}_1 \succeq 0, \quad 0 \leq u_0 \leq \dots \leq u_N \end{aligned} \quad (66)$$

where

$$\mathbf{S}_1 = \frac{L}{2} \mathbf{f}_{-1} \mathbf{f}_{-1}^\top + \sum_{i=0}^{N-1} u_i \mathbf{C}_i + \sum_{i=0}^N (u_i - u_{i-1}) \left( \mathbf{D}_i - \frac{1}{2L} \mathbf{f}_i \mathbf{f}_i^\top \right).$$

Here we used the same notation with (53). Each  $\frac{1}{2L} \mathbf{f}_i \mathbf{f}_i^\top$  term is subtracted from the original  $\mathbf{S}$  since we consider the inequality  $\llbracket x_i, x_\star \rrbracket - \frac{1}{2L} \|\nabla f(x_i)\|^2$  instead of  $\llbracket x_i, x_\star \rrbracket$  for  $\mathcal{L}_{F'}$ . Moreover, (66) is equivalent to

$$\begin{aligned} & \underset{u_0, \dots, u_N}{\text{maximize}} \frac{L}{2u_N} R^2 \\ & \text{subject to. } \mathbf{S}'_1 \succeq 0, \quad 0 \leq u_0 \leq \dots \leq u_N \end{aligned} \quad (67)$$

where

$$\mathbf{S}'_1 = \sum_{i=0}^{N-1} u_i \mathbf{C}_i + \sum_{i=0}^N (u_i - u_{i-1}) \left( \mathbf{D}_i - \frac{1}{2L} \mathbf{f}_i \mathbf{f}_i^\top \right) - \frac{1}{2L} \left( \sum_{i=0}^N (u_i - u_{i-1}) \mathbf{f}_i \right) \left( \sum_{i=0}^N (u_i - u_{i-1}) \mathbf{f}_i \right)^\top.$$

Similarly, under the assumption of strong duality,  $\mathcal{P}_2(\mathcal{L}_G, H, R)$  is equivalent to

$$\begin{aligned} & \underset{v_0, \dots, v_N}{\text{minimize}} \quad \frac{v_0 L}{2} R^2 \\ & \text{subject to.} \quad \mathbf{T}_1 \succeq 0, \quad 0 \leq v_0 \leq \dots \leq v_N \end{aligned} \quad (68)$$

where

$$\mathbf{T}_1 = \sum_{i=0}^{N-1} v_{i+1} \mathbf{A}_i + \sum_{i=0}^{N-1} (v_{i+1} - v_i) \left( \mathbf{B}_i - \frac{1}{2L} (\mathbf{e}_i - \mathbf{e}_N)(\mathbf{e}_i - \mathbf{e}_N)^\top \right) - \frac{1}{2} \mathbf{e}_N \mathbf{e}_N^\top.$$

Now we will prove the following proposition.

**Proposition 3.** Consider a matrix  $H = \{h_{k,i}\}_{0 \leq i < k \leq N}$  and  $H^A$ . If  $h_{i+1,i} \neq 0$  for  $0 \leq i \leq N-1$  and treating the solution of infeasible maximize problem as 0, the optimal values of  $\mathcal{P}_1(\mathcal{L}_F, H, R)$  and  $\mathcal{P}_2(\mathcal{L}_G, H^A, R)$  are same.

*Proof.* The proof structure is the same as the proof of Proposition 2. First consider  $\{\mathbf{f}_i\}_{i=0}^N$  as length  $N+1$ , which gives  $\mathbf{S}'_1, \mathbf{T}_1 \in \mathbb{S}^{N+1}$ . Furthermore, in the proof of Appendix A, we proved that

$$\mathbf{S}'_1 = \mathcal{M}(u)^\top \mathbf{T}_1 \mathcal{M}(u).$$

Therefore,

$$\mathbf{S}'_1 \succeq 0 \quad \Leftrightarrow \quad \mathbf{T} \succeq 0.$$

The other steps are the same as the proof of Proposition 2. □

*Proof of Theorem 5.*  $H_{\text{OGM}}$  is the solution of (46) since (OGM) is  $\mathcal{A}^*$ -optimal with respect to  $\mathcal{L}_F$  and (P1). Additionally, if

$$(46) = \underset{H, h_{i+1,i} \neq 0}{\text{maximize}} \mathcal{P}_1(\mathcal{L}_F, H, R) \quad (69)$$

holds,  $H_{\text{OGM}}$  is the solution to the following problem due to the strong duality.

$$\begin{aligned} & \underset{H, h_{i+1,i} \neq 0}{\text{maximize}} \quad \underset{u_0, \dots, u_N}{\text{maximize}} \quad \frac{L}{2u_N} R^2 \\ & \text{subject to.} \quad \mathbf{S} \succeq 0, \quad 0 \leq u_0 \leq \dots \leq u_N. \end{aligned} \quad (70)$$

Applying Proposition 2 and using the fact  $H_{\text{OGM-G}} = H_{\text{OGM}}^A$  provides that  $H_{\text{OGM-G}}$  is the solution of

$$\begin{aligned} & \underset{H, h_{i+1,i} \neq 0}{\text{maximize}} \quad \underset{v_0, \dots, v_N}{\text{minimize}} \quad \frac{v_0 L}{2} R^2 \\ & \text{subject to.} \quad \mathbf{T} \succeq 0, \quad 0 \leq v_0 \leq \dots \leq v_N. \end{aligned} \quad (71)$$

Finally, if

$$(47) = \underset{H, h_{i+1,i} \neq 0}{\text{maximize}} \mathcal{P}_1(\mathcal{L}_F, H, R) \quad (72)$$

holds, the optimal solution of (47) is the  $H_{\text{OGM-G}}$ , which proves the  $\mathcal{A}^*$ -optimality of (OGM-G) with respect to  $\mathcal{L}_G$  and (P2). The proof of (69) and (72) uses the continuity argument with  $H$  and please refer [44, Claim 4]. □

## C Omitted parts in Section 3

### C.1 Omitted calculations in Section 3.1

Strictly speaking, (7) is not a differential equation but rather a diffeo-integral equation. However, if  $H$  is separable, i.e., when  $H(t, s) = e^{\beta(s) - \gamma(t)}$  for some  $\beta, \gamma: [0, T] \rightarrow \mathbb{R}$ , then (7) can be reformulated as an ODE.

Assume the process (7) is well-defined. We can alternatively write (7) as

$$X(0) = x_0, \quad \dot{X}(t) = -e^{-\gamma(t)} \int_0^t e^{\beta(s)} \nabla f(X(s)) ds. \quad (73)$$

By multiplying  $e^{\gamma(t)}$  each side and differentiating, we obtain

$$\ddot{X}(t) + \dot{\gamma}(t) \dot{X}(t) + e^{\beta(t) - \gamma(t)} \nabla f(X(t)) = 0.$$

The H-dual of (73) is

$$Y(0) = x_0, \quad \dot{Y}(t) = -e^{\beta(T-t)} \int_0^t e^{-\gamma(T-s)} \nabla f(Y(s)) ds.$$

Under the well-definedness, by multiplying  $e^{-\beta(T-t)}$  each side and differentiating, we obtain

$$\ddot{Y}(t) + \dot{\beta}(T-t) \dot{Y}(t) + e^{\beta(T-t) - \gamma(T-t)} \nabla f(Y(t)) = 0.$$

When  $\beta(s) = \gamma(t) = r \log t$ , two ODEs become

$$\ddot{X}(t) + \frac{r}{t} \dot{X}(t) + \nabla f(X(t)) = 0 \quad (74)$$

and

$$\ddot{Y}(t) + \frac{r}{T-t} \dot{Y}(t) + \nabla f(Y(t)) = 0, \quad (75)$$

which are introduced in Section 3.2.

For the calculations of energy functions of (74) and (75), refer [51, Section 3.1, Section 4.2]. They proved that

$$\begin{aligned} (r-1) \|X(0) - x_\star\|^2 &= T^2 (f(X(T)) - x_\star) + \frac{1}{2} \left\| T \dot{X}(T) + 2(X(T) - x_\star) \right\|^2 \\ &\quad + (r-3) \|X(T) - x_\star\|^2 + \int_0^T (r-3)s \left\| \dot{X}(s) \right\|^2 ds - \int_0^T 2s [X(s), x_\star] ds \end{aligned}$$

holds for (74) and

$$\begin{aligned} &\frac{1}{T^2} (f(Y(0)) - f(Y(T))) + \frac{-r+3}{T^4} \|Y(0) - Y(T)\|^2 \\ &= \frac{1}{4(r-1)} \|\nabla f(Y(T))\|^2 + \int_0^T \frac{r-3}{(T-s)^5} \left\| (T-s) \dot{Y}(s) + 2(Y(s) - Y(T)) \right\|^2 ds - \int_0^T \frac{2}{(T-s)^3} [Y(s), Y(T)] ds \end{aligned}$$

holds for (75). After arranging terms, we obtain the results in Section 3.2. <sup>6</sup>

### C.2 Proof of Theorem 2

To begin with, we give a formal version of Theorem 2. Consider two following conditions.

$$u(T) (f(X(T)) - f^\star) \leq \mathcal{U}(T), \quad (\forall X(0), x^\star, \{\nabla f(X(s))\}_{s \in [0, T]} \in \mathcal{A}_1). \quad (C3')$$

$$\frac{1}{2} \|\nabla f(Y(T))\|^2 \leq \mathcal{V}(T), \quad (\forall Y(0), \{\nabla f(Y(s))\}_{s \in [0, T]} \in \mathcal{A}_2). \quad (C4')$$

where  $\mathcal{A}_1$  and  $\mathcal{A}_2$  are family of vectors which makes Fubini's Theorem can be applied and will be defined later in this section.

<sup>6</sup>In [51], they considered the ODE which gradient term is  $2\nabla f(Y(t))$  instead of  $\nabla f(Y(t))$ .

**Theorem 6** (Formal version of Theorem 2). Assume the C-FSFOMs (7) with  $H$  and  $H^A$  are well-defined in the sense that solutions to the diffeo-integral equations exist. Consider differentiable functions  $u, v: (0, T) \rightarrow \mathbb{R}$  that  $v(t) = \frac{1}{u(T-t)}$  for  $t \in [0, T]$  and

- (i)  $\lim_{s \rightarrow 0} u(t) = 0$
- (ii)  $\lim_{s \rightarrow -T} v(s) (f(Y(s)) - f(Y(T)) + \langle \nabla f(Y(T)), Y(T) - Y(s) \rangle) = 0$ .
- (iii)  $f$  is  $L$ -smooth and convex.

Then the following holds.

$$[(C3') \text{ is satisfied with } u(\cdot) \text{ and } H] \Leftrightarrow [(C4') \text{ is satisfied with } v(\cdot) \text{ and } H^A].$$

**Calculations of energy functions via transformation** Frist of all, we calculate  $\mathcal{U}(T) - u(T)(f(X(T)) - f^*)$ .

$$\begin{aligned} & \mathcal{U}(T) - u(T) (f(X(T)) - f^*) \\ &= \frac{1}{2} \|X(0) - x^*\|^2 + \int_0^T u'(s) (f(X(s)) - f^* + \langle \nabla f(X(s)), x^* - X(s) \rangle) ds - u(T) (f(X(T)) - f^*) ds \\ &= \frac{1}{2} \|X(0) - x^*\|^2 + \int_0^T u'(s) \langle \nabla f(X(s)), x^* - X(s) \rangle ds - \int_0^T u(s) \langle \nabla f(X(s)), \dot{X}(s) \rangle ds \\ &= \frac{1}{2} \|X(0) - x^*\|^2 + \int_0^T u'(s) \langle \nabla f(X(s)), x^* - X(0) \rangle ds \\ & \quad + \int_0^T u'(s) \langle \nabla f(X(s)), X(0) - X(s) \rangle ds - \int_0^T u(s) \langle \nabla f(X(s)), \dot{X}(s) \rangle ds \\ &= \frac{1}{2} \left\| X(0) - x^* - \int_0^T u'(s) \nabla f(X(s)) ds \right\|^2 - \frac{1}{2} \left\| \int_0^T u'(s) \nabla f(X(s)) ds \right\|^2 \\ & \quad + \int_0^T u'(s) \langle \nabla f(X(s)), X(0) - X(s) \rangle ds - \int_0^T u(s) \langle \nabla f(X(s)), \dot{X}(s) \rangle ds. \end{aligned}$$

We used parts of integration and  $u(0) := \lim_{s \rightarrow 0} u(s) = 0$ . Since  $X(0) - x^*$  can have any value, (C3) is equivalent to

$$\begin{aligned} & -\frac{1}{2} \left\| \int_0^T u'(s) \nabla f(X(s)) ds \right\|^2 \\ & \quad + \int_0^T u'(s) \langle \nabla f(X(s)), X(0) - X(s) \rangle ds - \int_0^T u(s) \langle \nabla f(X(s)), \dot{X}(s) \rangle ds \geq 0 \quad (76) \end{aligned}$$

holds for any  $\{\nabla f(X(s))\}_{s \in [0, T]}$ .

Now define a transformation  $\{f_s \in \mathbb{R}^d\}_{s \in [0, T]} \rightarrow \{g_s \in \mathbb{R}^d\}_{s \in [0, T]}$  as

$$g_s := u(s) f_s + \int_s^T u'(z) f_z dz, \quad s \in [0, T] \quad (77)$$

and  $\{g_s \in \mathbb{R}^d\}_{s \in [0, T]} \rightarrow \{f_s \in \mathbb{R}^d\}_{s \in [0, T]}$ .

$$f_s := \frac{1}{u(T)} g_0 + \frac{1}{u(s)} (g_s - g_0) - \int_s^T \frac{u'(b)}{u(b)^2} (g_b - g_0) db, \quad s \in [0, T]. \quad (78)$$

One can show that the above two transformations are in the inverse relationship. Next, we calculate  $\mathcal{U}(T)$ . Define  $f_s := \nabla f(X(s))$ . Under the transformation (77), we can find a simple expression of

(76).

$$\begin{aligned}
(76) &= -\frac{1}{2} \|g_0\|^2 - \int_0^T u'(s) \langle f_s, X(s) - X(0) \rangle ds - \int_0^T u(s) \langle f_s, \dot{X}(s) \rangle ds \\
&= -\frac{1}{2} \|g_0\|^2 - \int_0^T u'(s) \left\langle f_s, \int_0^s \dot{X}(a) da \right\rangle ds + \int_0^T u(s) \left\langle f_s, \int_0^s \dot{X}(a) da \right\rangle ds \\
&= -\frac{1}{2} \|g_0\|^2 - \int_0^T u'(s) \left\langle f_s, \int_0^s \dot{X}(a) da \right\rangle ds + \int_0^T \int_0^s H(s, a) \langle u(s) f_s, f_a \rangle dad s \quad (79) \\
&\stackrel{(\circ)}{=} -\frac{1}{2} \|g_0\|^2 - \int_0^T \left\langle \dot{X}(a), \int_a^T u'(s) f_s ds \right\rangle da + \int_0^T \int_0^s H(s, a) \langle u(s) f_s, f_a \rangle dad s \\
&= -\frac{1}{2} \|g_0\|^2 + \int_0^T \int_0^s H(s, a) \langle f_a, g_s \rangle dad s.
\end{aligned}$$

We used Fubini's Theorem at  $(\circ)$ . Next we calculate  $\mathcal{V}(T)$ .

$$\begin{aligned}
\mathcal{V}(T) &= \frac{1}{u(T)} (f(Y(0)) - f(Y(T))) + \int_0^T \frac{d}{ds} \frac{1}{u(T-s)} (f(Y(s)) - f(Y(T)) + \langle \nabla f(Y(s)), Y(T) - Y(s) \rangle) ds \\
&= \frac{1}{u(T)} (f(Y(0)) - f(Y(T))) + \int_0^T \frac{u'(T-s)}{u(T-s)^2} \langle Y(T) - Y(s), \nabla f(Y(s)) - \nabla f(Y(T)) \rangle ds \\
&\quad + \int_0^T \frac{d}{ds} \frac{1}{u(T-s)} (f(Y(s)) - f(Y(T)) + \langle \nabla f(Y(T)), Y(T) - Y(s) \rangle) ds \\
&\stackrel{(\circ)}{=} \frac{1}{u(T)} \langle \nabla f(Y(T)), Y(0) - Y(T) \rangle + \int_0^T \frac{u'(T-s)}{u(T-s)^2} \langle Y(T) - Y(s), \nabla f(Y(s)) - \nabla f(Y(T)) \rangle ds \\
&\quad - \int_0^T \frac{1}{u(T-s)} \langle \dot{Y}(s), \nabla f(Y(s)) - \nabla f(Y(T)) \rangle ds.
\end{aligned}$$

$(\circ)$  comes from parts of integration and the assumption

$$\lim_{s \rightarrow T} v(s) (f(Y(s)) - f(Y(T)) + \langle \nabla f(X(T)), Y(T) - Y(s) \rangle) = 0.$$

To clarify,  $u'(T-s) = u'(z)|_{z=T-s}$ . Now briefly write  $g_s := \nabla f(Y(T-s))$ . Then

$$\begin{aligned}
&\mathcal{V}(T) - \frac{1}{2} \|\nabla f(Y(T))\|^2 \\
&= -\frac{1}{2} \|g_0\|^2 - \frac{1}{u(T)} \int_0^T \langle g_0, \dot{Y}(s) \rangle ds + \int_0^T \frac{u'(T-s)}{u(T-s)^2} \langle Y(T) - Y(s), g_{T-s} - g_0 \rangle ds \\
&\quad - \int_0^T \frac{1}{u(T-s)} \langle \dot{Y}(s), g_{T-s} - g_0 \rangle ds \\
&= -\frac{1}{2} \|g_0\|^2 + \int_0^T \frac{u'(T-s)}{u(T-s)^2} \left\langle \int_s^T \dot{Y}(a) da, g_{T-s} - g_0 \right\rangle ds \\
&\quad - \frac{1}{u(T)} \int_0^T \langle g_0, \dot{Y}(s) \rangle ds - \int_0^T \frac{1}{u(T-s)} \langle \dot{Y}(s), g_{T-s} - g_0 \rangle ds \\
&\stackrel{(\circ)}{=} -\frac{1}{2} \|g_0\|^2 + \int_0^T \left\langle -\frac{1}{u(T)} g_0 - \frac{1}{u(T-s)} (g_{T-s} - g_0) + \int_0^s \frac{u'(T-b)}{u(T-b)^2} (g_{T-b} - g_0) db, \dot{Y}(s) \right\rangle ds. \quad (80)
\end{aligned}$$

We use Fubini's Theorem at (o). Finally, using (78), we obtain

$$\begin{aligned}
& \mathcal{V}(T) - \frac{1}{2} \|\nabla f(Y(T))\|^2 \\
&= -\frac{1}{2} \|g_0\|^2 - \int_0^T \langle f_{T-s}, \dot{Y}(s) \rangle ds \\
&= -\frac{1}{2} \|g_0\|^2 + \int_0^T \left\langle f_{T-s}, \int_0^s H^A(s, a) g_{T-a} da \right\rangle ds \\
&= -\frac{1}{2} \|g_0\|^2 + \int_0^T \left\langle f_{T-s}, \int_0^s H(T-a, T-s) g_{T-a} da \right\rangle ds \\
&= -\frac{1}{2} \|g_0\|^2 + \int_0^T \int_0^s H(s, a) \langle f_a, g_s \rangle da ds.
\end{aligned} \tag{81}$$

**Proof of Theorem 2** Define  $\mathcal{A}_1$  and  $\mathcal{A}_2$  as follows.

$$\begin{aligned}
\mathcal{A}_1 &= \{ \{f_s\}_{s \in [0, T]} \mid \text{Analysis in the previous paragraph holds} \}, \\
\mathcal{A}_2 &= \{ \{g_s\}_{s \in [0, T]} \mid \text{Analysis in the previous paragraph holds} \}.
\end{aligned}$$

Now we prove Theorem 2. We have shown that (C3') is equivalent to (76)  $\geq 0$  for all  $\{\nabla f(X(s))\}_{s \in [0, T]} \in \mathcal{A}_1$ . By definition of  $\mathcal{A}_1$ , it is equivalent to

$$-\frac{1}{2} \|g_0\|^2 + \int_0^T \int_0^s H(s, a) \langle f_a, g_s \rangle da ds \geq 0$$

for any  $\{g_s\}_{s \in [0, T]} \in \mathcal{A}_2$ . Moreover, by (81), it is also equivalent to

$$-\frac{1}{2} \|g_0\|^2 + \int_0^T \int_0^s H(s, a) \langle f_a, g_s \rangle da ds = \mathcal{V}(T) - \frac{1}{2} \|\nabla f(Y(T))\|^2 \geq 0$$

for any  $\{g_s\}_{s \in [0, T]} \in \mathcal{A}_2$ , which is (C4').

### C.3 Omitted parts in Section 3.3

**Regularity of (10) at  $t = -T$ .** To begin with, note that ODE (10) can be expressed as

$$\begin{cases} \dot{W}(t) = -\frac{2p-1}{T-t} W(t) - Cp^2(T-t)^{p-2} \nabla f(Y(t)) \\ \dot{Y}(t) = W(t) \end{cases}$$

for  $t \in (0, T)$ . Since right hand sides are Lipschitz continuous with respect to  $W$  and  $Y$  in any closed interval  $[0, s] \in [0, T)$ , solution  $(Y, W)$  uniquely exists that satisfies above ODE with initial condition  $(Y(0), W(0)) = (y_0, 0)$ . Next, we give the proof of regularity of ODE (10) at terminal time  $T$  in the following order. The proof structure is based on the regularity proof in [51]:

- (i)  $\sup_{t \in [0, T)} \|\dot{Y}(t)\|$  is bounded
- (ii)  $Y(t)$  can be continuously extended to  $T$
- (iii)  $\lim_{t \rightarrow T^-} \|\dot{Y}(t)\| = 0$
- (iv)  $\lim_{t \rightarrow T^-} \frac{\dot{Y}(t)}{(T-t)^{p-1}} = Cp \nabla f(Y(T))$ .

**Step (i):**  $\sup_{t \in [0, T)} \|\dot{Y}(t)\|$  is bounded. ODE (10) is equivalent to

$$\frac{1}{(T-s)^{p-2}} \ddot{Y}(s) + \frac{2p-1}{(T-s)^{p-1}} \dot{Y}(s) + Cp^2 \nabla f(Y(s)) = 0. \tag{82}$$

By multiplying  $\dot{Y}(s)$  and integrating from 0 to  $t$ , we obtain

$$\int_0^t \frac{1}{(T-s)^{p-2}} \langle \ddot{Y}(s), \dot{Y}(s) \rangle ds + \int_0^t \frac{2p-1}{(T-s)^{p-1}} \|\dot{Y}(s)\|^2 ds + Cp^2 \int_0^t \langle \dot{Y}(s), \nabla f(Y(s)) \rangle ds = 0$$

and integration by parts gives us

$$\frac{1}{2(T-t)^{p-2}} \|\dot{Y}(t)\|^2 - \frac{1}{2T^{p-2}} \|\dot{Y}(0)\|^2 + \int_0^t \frac{2p-1}{(T-s)^{p-1}} \|\dot{Y}(s)\|^2 ds + Cp^2 (f(Y(t)) - f(Y(0))) = 0.$$

Define  $\Psi(t): [0, T] \rightarrow \mathbb{R}$  as

$$\Psi(t) = \frac{1}{2(T-t)^{p-2}} \|\dot{Y}(t)\|^2 + Cp^2 (f(Y(t)) - f(Y(0))).$$

Since

$$\dot{\Psi}(t) = -\frac{2p-1}{(T-s)^{p-1}} \|\dot{Y}(s)\|^2 ds,$$

$\Psi(t)$  is a nonincreasing function. Thus

$$\|\dot{Y}(t)\|^2 = 2(T-t)^{p-2} (\Psi(t) + Cp^2 (f(Y(0)) - f(Y(t)))) \leq 2T^{p-2} (\Psi(0) + Cp^2 (f(Y(0)) - f_*)), \quad (83)$$

and the right hand side of (83) is constant, which implies  $M := \sup_{t \in [0, T]} \|\dot{Y}(t)\| < \infty$ .

**Step (ii):  $Y(t)$  can be continuously extended to  $T$ .** We can prove  $Y(t)$  is uniformly continuous due to the following analysis.

$$\|Y(t+\delta) - Y(t)\| = \left\| \int_t^{t+\delta} \dot{Y}(s) ds \right\| \leq \int_t^{t+\delta} \|\dot{Y}(s)\| ds \leq \delta M.$$

Since a uniformly continuous function  $g: D \rightarrow \mathbb{R}^d$  can be extended continuously to  $\bar{D}$ ,  $Y: [0, T) \rightarrow \mathbb{R}^d$  can be extended to  $[0, T]$ .

**Step (iii):  $\lim_{t \rightarrow T^-} \|\dot{Y}(t)\| = 0$ .** We first prove the limit  $\lim_{t \rightarrow T^-} \|\dot{Y}(t)\| = 0$  exists. From C.3, we know  $\lim_{t \rightarrow T^-} Y(t)$  exists and by continuity of  $f$ ,  $\lim_{t \rightarrow T^-} f(Y(t))$  also exists. Moreover,  $\Psi(t)$  is non-increasing and

$$\Psi(t) = \frac{1}{2(T-t)^{p-2}} \|\dot{Y}(t)\|^2 + Cp^2 (f(Y(t)) - f(Y(0))) \geq Cp^2 (f_* - f(Y(0))),$$

thus  $\lim_{t \rightarrow T^-} \Psi(t)$  exists. Therefore,  $\lim_{t \rightarrow T^-} \frac{\|\dot{Y}(t)\|^2}{(T-t)^{p-2}}$  exists, which implies  $\lim_{t \rightarrow T^-} \|\dot{Y}(t)\| = 0$  when  $p > 2$ . For the case  $p = 2$ ,  $\lim_{t \rightarrow T^-} \|\dot{Y}(t)\|$  exists, and  $\lim_{t \rightarrow T^-} \|\dot{Y}(t)\| = 0$  follows since we have

$$\Psi(t) = \frac{1}{2T^p} \|\dot{Y}(0)\|^2 - \int_0^t \frac{3}{(T-s)^{p-1}} \|\dot{Y}(s)\|^2 ds$$

is bounded below. Thus the value of integration is finite.

**Step (iv):  $\lim_{t \rightarrow T^-} \frac{\dot{Y}(t)}{(T-t)^{p-1}} = Cp \nabla f(Y(T))$ .** The (7) form of process  $Y$  is

$$\dot{Y}(t) = - \int_0^t \frac{Cp^2 (T-s)^{2p-1}}{(T-s)^{p+1}} \nabla f(Y(s)) ds.$$

By dividing  $(T-t)^{p-1}$  each side, we obtain

$$\frac{\dot{Y}(t)}{(T-t)^{p-1}} = -(T-t)^p \int_0^t \frac{Cp^2}{(T-s)^{p+1}} \nabla f(Y(s)) ds.$$

By the result of C.3, we can apply L'Hopital's rule, which gives

$$\lim_{t \rightarrow T^-} \frac{\dot{Y}(t)}{(T-t)^{p-1}} = - \lim_{t \rightarrow T^-} \frac{\int_0^t \frac{Cp^2}{(T-s)^{p+1}} \nabla f(Y(s)) ds}{(T-t)^{-p}} = Cp \nabla f(Y(T)).$$

Also, since  $\nabla f$  is  $L$ -Lipshitz,  $\|\nabla f(Y(t)) - \nabla f(Y(T))\| \leq L \|Y(t) - Y(T)\|$ . Therefore,

$$\lim_{t \rightarrow T^-} \frac{\nabla f(Y(t)) - \nabla f(Y(T))}{(T-t)^\beta} = 0 \quad (84)$$

for any  $\beta < p$ .

**Applying Theorem 2 to (10).** For the case  $p = 2$ , refer [51]. Now we consider the case  $p > 2$ , with  $u(t) = Ct^p$  and  $v(t) = \frac{1}{C(T-t)^p}$ . First, we verify conditions (i) and (ii) in Theorem 6. (i) holds since  $u(0) = 0$ , and (ii) holds since

$$\begin{aligned} & \lim_{s \rightarrow T^-} v(s) (f(Y(s)) - f(Y(T)) + \langle \nabla f(Y(T)), Y(T) - Y(s) \rangle) \\ &= \frac{1}{C} \lim_{s \rightarrow T^-} \frac{f(Y(s)) - f(Y(T)) + \langle \nabla f(Y(T)), Y(T) - Y(s) \rangle}{(T-s)^p} \\ &\stackrel{(\circ)}{=} \frac{1}{C} \lim_{s \rightarrow T^-} \frac{\langle \dot{Y}(s), \nabla f(Y(s)) - \nabla f(Y(T)) \rangle}{-p(T-s)^{p-1}} \\ &\stackrel{(\bullet)}{=} -\frac{1}{pC} \lim_{s \rightarrow T^-} \left\langle \frac{\dot{Y}(s)}{(T-s)^{p-1}}, \nabla f(Y(s)) - \nabla f(Y(T)) \right\rangle \\ &= 0. \end{aligned}$$

( $\circ$ ) uses L'Hospital's rule and ( $\bullet$ ) uses the limit results at the previous paragraph.

To prove

$$\frac{1}{2} \|\nabla f(Y(T))\|^2 \leq \frac{1}{CT^p} (f(Y(0)) - f(Y(T))),$$

we carefully check that for any  $L$ -smooth convex function  $f$ ,  $\{\nabla f(Y(s))\}_{s \in [0, T]} \in \mathcal{A}_2$ .

**Verification of (80).** Fubini's Theorem is used for

$$\begin{aligned} & \int_{0 \leq s, a \leq T} \mathbf{1}_{s \leq a} \frac{u'(T-s)}{u(T-s)^2} \langle \dot{Y}(a), (\nabla f(Y(s)) - \nabla f(Y(T))) \rangle \\ &= \int_{0 \leq s, a \leq T} \mathbf{1}_{s \leq a} \dot{Y}(a) (\nabla f(Y(s)) - \nabla f(Y(T))) \frac{p}{C(T-s)^{p+1}}. \end{aligned}$$

The above function is continuous in  $0 \leq s, a \leq T$  due to  $\lim_{t \rightarrow T^-} \frac{\dot{Y}(t)}{(T-t)^{p-1}} = Cp \nabla f(Y(T))$  and (84) with  $\beta = 2 < p$ .

**Verification of (79).** First,

$$\lim_{t \rightarrow T^-} \frac{Y(t) - Y(T)}{(T-t)^p} = \lim_{t \rightarrow T^-} \frac{\dot{Y}(t)}{-p(T-t)^{p-1}} = -C \nabla f(Y(T))$$

holds from (iv). Thus  $\sup_{b \in (0, T]} \left\| \frac{Y(T-b) - Y(T)}{b^p} \right\| < \infty$ . Now we will show

$$\lim_{a \rightarrow +0} a^{1+\epsilon} f_a = 0, \quad \forall \epsilon > 0.$$

Observe

$$f_a = \frac{1}{CT^p} g_0 + \frac{1}{CT^p} (g_a - g_0) - \int_a^T \frac{p}{Cb^{p+1}} (g_b - g_a) db$$

and

$$\begin{aligned} \|a^{1+\epsilon} f_a\| &= \left\| \frac{a^{1+\epsilon}}{CT^p} g_0 + \frac{a^{1+\epsilon}}{CT^p} (g_a - g_0) - \int_a^T \frac{pa^{1+\epsilon}}{Cb^{p+1}} (g_b - g_a) db \right\| \\ &\leq \left\| \frac{a^{1+\epsilon}}{CT^p} g_0 + \frac{a^{1+\epsilon}}{CT^p} (g_a - g_0) \right\| + \int_a^T \frac{pa^{1+\epsilon}}{Cb^{p+1}} \|g_b - g_0\| db + \int_a^T \frac{pa^{1+\epsilon}}{Cb^{p+1}} \|g_a - g_0\| db \\ &\leq \left\| \frac{a^{1+\epsilon}}{CT^p} g_0 + \frac{a^{1+\epsilon}}{CT^p} (g_a - g_0) \right\| + \int_a^T \frac{Lpa^{1+\epsilon}}{Cb^{p+1}} \|Y(T-b) - Y(T)\| db + \int_a^T \frac{Lpa^{1+\epsilon}}{Cb^{p+1}} \|Y(T-a) - Y(T)\| db \\ &\leq \left\| \frac{a^{1+\epsilon}}{CT^p} g_0 + \frac{a^{1+\epsilon}}{CT^p} (g_a - g_0) \right\| + \int_a^T Lpa^\epsilon \left\| \frac{Y(T-b) - Y(T)}{b^p} \right\| db \\ &\quad + \frac{La^{1+\epsilon}}{C} \left( \frac{1}{a^p} - \frac{1}{T^p} \right) \|Y(T-a) - Y(T)\|. \end{aligned}$$

By using the boundness of  $\frac{Y(T-b)-Y(T)}{b^p}$ , we obtain the desired result.

Next,

$$\begin{aligned} \|a^{2\epsilon-p+2} \dot{X}(a)\| &= \left\| \int_0^a a^{\epsilon-p+2} H(a,b) f_b db \right\| \\ &\leq \int_0^a \frac{Cp^2 b^{2p-1}}{a^{2p-1-2\epsilon}} \|f_b\| db \\ &= \int_0^a \frac{Cp^2 b^{2p-2-\epsilon}}{a^{2p-1-2\epsilon}} \|b^{1+\epsilon} f_b\| db \\ &= \left( \sup_{b \in [0,a]} \|b^{1+\epsilon} f_b\| \right) \frac{Cp^2}{2p-1-\epsilon} a^\epsilon, \end{aligned}$$

which gives

$$\lim_{a \rightarrow +0} a^{2\epsilon-p+2} \dot{X}(a) = 0, \quad \forall 0 < \epsilon < 2p-1.$$

For the final step, recall that Fubini's Theorem is used for

$$\int_{0 \leq a, s \leq T} \mathbf{1}_{a \leq s} s^{p-1} \langle f_s, \dot{X}(a) \rangle dad s.$$

To prove that the above function is continuous, take any  $0 < \epsilon < \frac{p-2}{2}$ . Then

$$\lim_{0 \leq a \leq s, s \rightarrow +0} \|s^{p-1} \langle f_s, \dot{X}(a) \rangle\| \leq \lim_{0 \leq a \leq s, s \rightarrow +0} \left\| \frac{s^{-2\epsilon+2p-3}}{a^{-2\epsilon+p-2}} \langle \dot{X}(a), f_s \rangle \right\| = \lim_{s \rightarrow +0} \|s^{-2\epsilon+2p-3} f_s\| \cdot \lim_{a \rightarrow +0} \left\| \frac{\dot{X}(a)}{a^{-2\epsilon+p-2}} \right\| = 0,$$

which gives the desired result.

## D Omitted calculation in Section 4

### D.1 Preliminaries

**General proximal step and composite FSFOM** For  $a > 0$ , we newly define a prox-grad step with step size  $\frac{1}{L} \frac{1}{\alpha}$  as follows:

$$y^{\oplus, \alpha} := \operatorname{argmin}_{z \in \mathbb{R}^d} \left( f(y) + \langle \nabla f(y), z - y \rangle + g(z) + \frac{\alpha L}{2} \|z - y\|^2 \right). \quad (85)$$

First, we will provide a generalized version of prox-grad inequality; Prox-grad inequality is originally used to prove the convergence of FISTA as below [11]:

$$F(y^{\oplus}) - L \langle y^{\oplus, 1} - y, x - y^{\oplus, 1} \rangle \leq F(x) + \frac{L}{2} \|y^{\oplus, 1} - y\|^2.$$

**Proposition 4** (Generalized version of prox-grad inequality). For every  $x, y \in \mathbb{R}^d$ , the following holds:

$$\langle x, y \rangle^\alpha := F(y^{\oplus, \alpha}) - F(x^{\oplus, \alpha}) - L\alpha \langle y^{\oplus, \alpha} - y, x^{\oplus, \alpha} - y^{\oplus, \alpha} \rangle - \frac{L}{2} \|y^{\oplus, \alpha} - y\|^2 \leq 0.$$

*Proof.* The optimality condition of (85) provides

$$L\alpha(y^{\oplus, \alpha} - y) + \nabla f(y) + u = 0, \quad u \in \partial g(y^{\oplus, \alpha}). \quad (86)$$

In addition,  $L$ -smoothness of  $f$  and convexity inequality of  $f$  and  $g$  give

$$\begin{aligned} F(y^{\oplus, \alpha}) &\leq f(y) + \langle \nabla f(y), y^{\oplus, \alpha} - y \rangle + \frac{L}{2} \|y^{\oplus, \alpha} - y\|^2 + g(y^{\oplus, \alpha}) \\ g(y^{\oplus, \alpha}) + \langle u, x^{\oplus, \alpha} - y^{\oplus, \alpha} \rangle &= g(y^{\oplus, \alpha}) + \langle -L\alpha(y^{\oplus, \alpha} - y) - \nabla f(y), x^{\oplus, \alpha} - y^{\oplus, \alpha} \rangle \leq g(x^{\oplus, \alpha}), \\ f(y) + \langle \nabla f(y), x^{\oplus, \alpha} - y \rangle &\leq f(x^{\oplus, \alpha}). \end{aligned}$$

Summing the above inequalities provides the generalized version of prox-grad inequality.  $\square$

**Two problem settings for the composite optimization problem.** For the composite optimization problem, we consider the following two problems.

(P1') Efficiently reduce  $F(x_N^{\oplus, \alpha}) - F_\star$  assuming  $x_\star$  exists and  $\|x_0 - x_\star\| \leq R$ .

(P2') Efficiently reduce  $\min_{v \in \partial F(y_N^{\oplus, \alpha})} \|v\|^2$  assuming  $F_\star > -\infty$  and  $F(y_0) - F_\star \leq R$ .

Note that when  $g = 0$ , (P1') and (P2') collapse to (P1) and (P2), respectively.

**Parameterized FSFOM that reduces the composite function value : GFPGM [27]** First define  $x_\star := \operatorname{argmin}_{x \in \mathbb{R}^d} F(x)$ . We recall the composite function minimization FSFOMs with  $x_i^{\oplus, \alpha}$  by a lower triangular matrix  $\{h_{k,i}\}_{0 \leq i < k \leq N}$  as follows:

$$x_{k+1} = x_k - \sum_{i=0}^k \alpha h_{k+1,i} (x_i - x_i^{\oplus, \alpha}), \quad \forall k = 0, \dots, N-1. \quad (87)$$

In the case  $g = 0$ , (87) collapses to (1) as  $\alpha(z - z^{\oplus, \alpha}) = (z - z^+) = \frac{1}{L} \nabla f(z)$ . The iterations of GFPGM are defined as

$$x_{k+1} = x_k^{\oplus, 1} + \frac{(T_k - t_k)t_{k+1}}{t_k T_{k+1}} (x_k^{\oplus, 1} - x_{k-1}^{\oplus, 1}) + \frac{(t_k^2 - T_k)t_{k+1}}{t_k T_{k+1}} (x_k^{\oplus, 1} - x_k), \quad k = 0, \dots, N-1 \quad (88)$$

where  $x_{-1}^{\oplus, 1} = x_0$  and  $[t_i > 0, T_i = \sum_{j=0}^i t_j \leq t_i^2 \text{ for } 0 \leq i \leq N]$ . (88) reduces the composite function value as follows:

$$F(x_N^{\oplus, 1}) - F_\star \leq \frac{1}{T_N} \frac{1}{2} \|x_0 - x_\star\|^2. \quad (89)$$

We can reconstruct the convergence proof of [27, Theorem 3.3] via our energy function scheme, which is defined as

$$\mathcal{U}_k = \frac{L}{2} \|x_0 - x_\star\|^2 + \sum_{i=0}^{k-1} u_i \langle x_i, x_{i+1} \rangle^1 + \sum_{i=0}^k (u_i - u_{i-1}) \langle x_\star, x_i \rangle^1 \quad (90)$$

for  $k = -1, \dots, N$ . The convergence rate result (89) is proved by using that  $\{\mathcal{U}_k\}_{k=-1}^N$  is dissipative, and

$$\mathcal{U}_N - T_N (F(x_N^{\oplus, 1}) - F_\star) = \frac{L}{2} \left\| x_0 - x_\star + \sum_{i=0}^N (u_i - u_{i-1}) (x_i - x_i^{\oplus, 1}) \right\|^2 + \sum_{i=0}^N \frac{L(T_i - t_i^2)}{2} \|x_i^{\oplus, 1} - x_i\|^2 \geq 0. \quad (91)$$

Here, we can make a crucial observation.

**The formulation of (90) is same with (2), which all  $[\cdot, \cdot]$  are changed into  $\langle \cdot, \cdot \rangle^1$  and  $u_i = T_i$ .** ( $\diamond$ )

Moreover, we can express the recursive formula of the  $H$  matrix of (88) as

$$h_{k+1,i} = \begin{cases} 1 + \frac{(t_k-1)t_{k+1}}{T_{k+1}} & i = k \\ \frac{(T_k-t_k)t_{k+1}}{t_k T_{k+1}} (h_{k,k-1} - 1) & i = k-1 \\ \frac{(T_k-t_k)t_{k+1}}{t_k T_{k+1}} h_{k,i} & i = 0, \dots, k-2 \end{cases}. \quad (92)$$

As we used a parameterized family in the convex function (see Section 2.5), we will use these convergence results to construct the method for minimizing  $\min_{v \in \partial F(y_N^{\oplus, \alpha})} \|v\|^2$ .

**Relationship between minimizing  $\frac{\alpha^2}{2L} \|y_N^{\oplus, \alpha} - y_N\|^2$  and minimizing  $\min_{v \in \partial F(y_N^{\oplus, \alpha})} \|v\|^2$ .**

Note that  $L\alpha(y_N^{\oplus, \alpha} - y_N) + \nabla f(y_N) + u = 0$ ,  $u \in \partial g(y_N^{\oplus, \alpha})$ , which gives

$$\min_{v \in \partial F(y_N^{\oplus, \alpha})} \|v\|^2 \stackrel{(\bullet)}{\leq} (\|\nabla f(y_N) - \nabla f(y_N^{\oplus, \alpha})\| + L\alpha \|y_N - y_N^{\oplus, \alpha}\|)^2 \stackrel{(\circ)}{\leq} L^2(\alpha + 1)^2 \|y_N - y_N^{\oplus, \alpha}\|^2. \quad (93)$$

( $\bullet$ ) is triangle inequality and ( $\circ$ ) comes from the  $L$ -smoothness of  $f$ .

## D.2 Proof of Theorem 3

In this section, we give the proof outline of Theorem 3 and discuss the construction of matrix  $C$ .

*Proof outline of Theorem 3.* We begin by proposing a parameterized family reduces  $\frac{1}{2L} \|y_N^{\oplus, \alpha} - y_N\|^2$  under a fixed value of  $a > 0$ . To construct this family, take  $\{t_i\}_{i=0}^N$ ,  $T_i = \sum_{j=0}^i t_j$  that satisfies (96). We define the  $H$  matrix of the family as  $(\frac{1}{\alpha}H_0 + \frac{1}{\alpha^2}C)^A$ , where  $H_0$  has the same formulation as the  $H$  matrix of GFGPM (88) and  $C$  follows the recursive formula (97). We refer to this family as (SFG-family).

We can prove that (SFG-family) exhibits the convergence rate

$$\frac{\alpha L}{2} \|y_N^{\oplus, \alpha} - y_N\|^2 \leq \frac{1}{T_N} (F(y_0) - F_\star), \quad (94)$$

which is motivated by the observation ( $\diamond$ ). In parallel with ( $\diamond$ ), we consider the energy function

$$\mathcal{V}_k = \frac{F(y_0) - F(y_N^{\oplus, \alpha}) + \langle y_{-1}, y_0 \rangle^\alpha}{T_N} + \sum_{i=0}^{k-1} \frac{1}{T_{N-i-1}} \langle y_i, y_{i+1} \rangle^\alpha + \sum_{i=0}^{k-1} \left( \frac{1}{T_{N-i-1}} - \frac{1}{T_{N-i}} \right) \langle y_N, y_i \rangle^\alpha$$

for  $k = 0, \dots, N$ , which  $\llbracket y_N, \star \rrbracket$  changed into  $\langle y_{-1}, y_0 \rangle^\alpha$ , all other  $\llbracket \cdot, \cdot \rrbracket$  terms changed into  $\langle \cdot, \cdot \rangle^\alpha$  and  $v_i = \frac{1}{T_{N-i}}$ .

(94) follows from the fact that  $\{\mathcal{V}_i\}_{i=0}^N$  is dissipative and

$$\frac{\alpha L}{2} \|y_N^{\oplus, \alpha} - y_N\|^2 \leq \mathcal{V}_N.$$

Detailed justification is given in Appendix D.3.

For the next step, we show that (SFG) is an instance of (SFG-family), under the choice of

$$\alpha = 4, \quad T_i = \frac{(i+2)(i+3)}{4}, \quad 0 \leq i \leq N.$$

The detailed derivation is given in Appendix D.4.

Finally, combining the convergence result

$$2L\|y_N^{\oplus,4} - y_N\|^2 \leq \mathcal{V}_N \leq \mathcal{V}_0 \leq \frac{4}{(N+2)(N+3)} (F(y_0) - F_*)$$

and (93) gives

$$\min_{v \in \partial F(y_N^{\oplus,4})} \|v\|^2 \leq 25L^2 \left\| y_N^{\oplus,4} - y_N \right\|^2 \leq \frac{50}{(N+2)(N+3)} (F(y_0) - F_*).$$

□

**Construction of matrix  $C$ .** Define

$$g := L\alpha [y_0 - y_0^{\oplus,\alpha} | \dots | y_N - y_N^{\oplus,\alpha}].$$

For an FSFOM (87) with matrix  $H_1$ , we have  $\mathcal{V}_N - \frac{\alpha L}{2} \|y_N^{\oplus,\alpha} - y_N\|^2 = \frac{1}{2L\alpha^2} \text{Tr} (g^\top g \mathcal{T}^\oplus)$  with

$$\begin{aligned} \mathcal{T}^\oplus := & \underbrace{\alpha^2 \left[ \sum_{i=0}^N \frac{1}{T_{N-i}} (\mathcal{H}_1^\top (\mathbf{e}_i + \dots + \mathbf{e}_N) (\mathbf{e}_i - \mathbf{e}_{i-1})^\top + (\mathbf{e}_i - \mathbf{e}_{i-1}) (\mathbf{e}_i + \dots + \mathbf{e}_N)^\top \mathcal{H}_1) \right]}_{\mathcal{A}_3} \\ & + \underbrace{\alpha \left[ \sum_{i=0}^N \frac{1}{T_{N-i}} ((\mathbf{e}_{i-1} - \mathbf{e}_i) (\mathbf{e}_{i-1} - \mathbf{e}_N)^\top + (\mathbf{e}_{i-1} - \mathbf{e}_N) (\mathbf{e}_{i-1} - \mathbf{e}_i)^\top) - \frac{1}{T_N} \mathbf{e}_0 \mathbf{e}_0^\top - \mathbf{e}_N \mathbf{e}_N^\top \right]}_{\mathcal{B}_3} \\ & + \left[ \frac{\alpha}{T_N} \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=0}^{N-1} \frac{1}{T_{N-1-i}} \mathbf{e}_i \mathbf{e}_i^\top - \frac{1}{T_0} \mathbf{e}_N \mathbf{e}_N^\top \right] \end{aligned}$$

where  $e_{-1} = \mathbf{0}$ ,  $e_i$  is a unit vector which  $(i+1)$ -th component is 1 and  $\mathbb{R}^{N+1}$  and  $\mathcal{H}_1 := \begin{bmatrix} 0 & 0 \\ H_1 & 0 \end{bmatrix}$ . If we take  $H_1$  that makes  $\mathcal{T}^\oplus \succeq 0$ ,  $\mathcal{V}_N \geq \frac{\alpha L}{2} \|y_N^{\oplus,\alpha} - y_N\|^2$  follows. Next we observe  $\frac{1}{\alpha^2} \mathcal{A}_3 + \frac{1}{\alpha} \mathcal{B}_3 = \mathcal{T}(H_1, v)$  for  $v_i = \frac{1}{T_{N-i}}$  which is defined as (22). By expanding (91) and defining  $g' := L[x_0 - x_0^{\oplus,\alpha} | \dots | x_N - x_N^{\oplus,\alpha}]$ ,

$$\mathcal{U}_N - T_N (F(x_N^{\oplus,1}) - F_*) - \frac{L}{2} \left\| x_0 - x_* + \sum_{i=0}^N (u_i - u_{i-1}) (x_i - x_i^{\oplus,1}) \right\|^2 = \text{Tr} (g' (g')^\top \mathcal{S}^\oplus)$$

follows where

$$\mathcal{S}^\oplus := \mathcal{S}(H_0, u) - \sum_{i=0}^{N-1} (T_i - t_i^2) \mathbf{e}_i \mathbf{e}_i^\top.$$

Here,  $u_i = T_i$ , and  $\mathcal{S}(H_0, u)$  is given by the formulation (19). Using (91), we obtain

$$\mathcal{S}(H_0, u) = \sum_{i=0}^{N-1} (2T_i - t_i^2) \mathbf{e}_i \mathbf{e}_i^\top + (T_N - t_N^2) \mathbf{e}_N \mathbf{e}_N^\top.$$

Now recall matrix  $\mathcal{M}(u)$  (26) and the result of Theorem 1:

$$\mathcal{S}(H_0, u) = \mathcal{M}(u)^\top \mathcal{T}(H_0^A, v) \mathcal{M}(u). \quad (95)$$

By substituting  $H_1 = \frac{1}{\alpha} H_0^A + X$  and using (95), the result is

$$\begin{aligned} \mathcal{M}(u)^\top \mathcal{T}^\oplus \mathcal{M}(u) = & \alpha \left[ \sum_{i=0}^{N-1} (2T_i - t_i^2) \mathbf{e}_i \mathbf{e}_i^\top + (T_N - t_N^2) \mathbf{e}_N \mathbf{e}_N^\top \right] \\ & + \underbrace{\alpha^2 \mathcal{M}(u)^\top \left[ \sum_{i=0}^N \frac{1}{T_{N-i}} (\mathcal{X}^\top (\mathbf{e}_i + \dots + \mathbf{e}_N) (\mathbf{e}_i - \mathbf{e}_{i-1})^\top + (\mathbf{e}_i - \mathbf{e}_{i-1}) (\mathbf{e}_i + \dots + \mathbf{e}_N)^\top \mathcal{X}) \right]}_{\mathcal{A}_4} \mathcal{M}(u) \\ & + \underbrace{\mathcal{M}(u)^\top \left[ \frac{\alpha}{T_N} \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=0}^{N-1} \frac{1}{T_{N-1-i}} \mathbf{e}_i \mathbf{e}_i^\top - \frac{1}{T_0} \mathbf{e}_N \mathbf{e}_N^\top \right]}_{\mathcal{C}_3} \mathcal{M}(u) \end{aligned}$$

where  $\mathcal{X} = \begin{bmatrix} 0 & 0 \\ X & 0 \end{bmatrix}$ .

Next, consider  $\mathcal{A}_4$  as a function of the lower triangular matrix  $\mathcal{X}$ . The key observation is that if we choose  $\mathcal{X}$  appropriately, all non-diagonal terms of  $\mathcal{C}_3$  can be eliminated. With this choice of  $\mathcal{X}$ , we have

$$\mathcal{T}^\oplus = \alpha \sum_{i=0}^N (2T_i - t_i^2) \mathbf{e}_i \mathbf{e}_i^\top - T_0 \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=1}^N \left( \frac{T_i^2}{T_{i-1}} + t_i^2 \sum_{j=0}^{i-2} \frac{1}{T_j} \right) \mathbf{e}_i \mathbf{e}_i^\top.$$

Note that  $\mathcal{T}^\oplus$  contains only diagonal terms. Therefore,  $\mathcal{T}^\oplus \succeq 0$  is equivalent to all coefficients of  $\mathbf{e}_i \mathbf{e}_i^\top$  being nonnegative, which can be formulated using  $t_i$  and  $T_i$ .

**Remark.** In the proof of Theorem 1 ( $\mathbf{B}_2$  and  $\mathbf{B}_1$  at (27) and (28)), we have shown that

$$\mathcal{A}_4 = \alpha^2 \left[ \sum_{i=0}^N (\mathcal{X}^A)^\top T_i (\mathbf{e}_i - \mathbf{e}_{i+1})(\mathbf{e}_0 + \cdots + \mathbf{e}_i)^\top + T_i (\mathbf{e}_0 + \cdots + \mathbf{e}_i)(\mathbf{e}_i - \mathbf{e}_{i+1})^\top \mathcal{X}^A \right].$$

Observe that  $(\mathbf{e}_i - \mathbf{e}_{i+1})(\mathbf{e}_0 + \cdots + \mathbf{e}_i)^\top$  is a lower triangular matrix and  $\mathcal{X}^A$  is a strictly lower triangular matrix, i.e., all diagonal component is 0. Thus it is worth noting that  $\mathcal{A}_4$  cannot induce any diagonal term  $\mathbf{e}_i \mathbf{e}_i^\top$ , choosing matrix  $C$  as the optimal one.

### D.3 Parameterized family reduces the gradient mapping norm

**FSFOM that reduces  $\frac{1}{2L} \|y_N^{\oplus, \alpha} - y_N\|^2$  in the composite minimization.** We propose the parameterized family that reduces  $\frac{1}{2L} \|y_N^{\oplus, \alpha} - y_N\|^2$ . For fixed  $a > 0$ , take  $\{t_i\}_{i=0}^N$ ,  $T_i = \sum_{j=0}^i t_j$  that satisfies the following conditions:

$$\begin{aligned} \alpha(2T_0 - t_0^2) &\geq T_0, \\ \alpha(2T_k - t_k^2) &\geq \frac{T_k^2}{T_{k-1}} + t_k^2 \left( \frac{1}{T_0} + \sum_{i=0}^{k-2} \frac{1}{T_i} \right) \quad k = 1, \dots, N \end{aligned} \quad (96)$$

where we define  $\sum_{i=0}^{k-2} \frac{1}{T_i} = 0$  for  $k = 1$  case. We define an lower triangular matrix  $C = \{c_{k,i}\}_{0 \leq i < k \leq N}$  as

$$c_{k+1,i} = \begin{cases} \frac{t_1}{T_1} & i = 0, k = 0 \\ \frac{t_{k+1}}{T_{k+1}} \left( \frac{t_k}{T_0} + t_k \sum_{j=0}^{k-2} \frac{1}{T_j} + \frac{T_k}{T_{k-1}} \right) & i = k, k = 1, \dots, N-1 \\ \frac{t_{k+1}(T_k - t_k)}{t_k T_{k+1}} c_{k,i} & i = 0, \dots, k-1, i = 2, \dots, N-1 \end{cases} \quad (97)$$

Matrix  $C$  has a crucial role in the correspondence between composite function value minimization and composite function gradient norm minimization with  $H \rightarrow (\frac{1}{\alpha} H + \frac{1}{\alpha^2} C)^A$ .

**Proposition 5** (SFG family). Assume that we choose  $H_0$  matrix of (88), and define  $C$  as (97). Consider an FSFOM (87) which  $H$  matrix is  $(\frac{1}{\alpha} H_0 + \frac{1}{\alpha^2} C)^A$ . Such FSFOM can be expressed as

$$\begin{aligned} y_{k+1} &= y_k^{\oplus, \alpha} + \beta'_k (y_k^{\oplus, \alpha} - y_{k-1}^{\oplus, \alpha}) + \gamma'_k (y_k^{\oplus, \alpha} - y_k), \quad k = 0, \dots, N-1, \\ \beta'_k &= \frac{\beta_{N-k}(\beta_{N-1-k} + \gamma_{N-1-k})}{\beta_{N-k} + \gamma_{N-k}}, \quad \gamma'_k = \frac{\gamma_{N-k}(\beta_{N-1-k} + \gamma_{N-1-k})}{\beta_{N-k} + \gamma_{N-k}}, \quad (\text{SFG-family}) \\ \beta_k &= \frac{t_{k+1}(T_k - t_k)}{t_k T_{k+1}}, \quad \gamma_k = \frac{t_{k+1}(t_k^2 - T_k)}{t_k T_{k+1}} + \frac{1}{\alpha} c_{k+1,k}. \end{aligned}$$

and  $y_{-1}^{\oplus, \alpha} = y_0$ . Then it exhibits the convergence rate

$$\frac{\alpha L}{2} \|y_N^{\oplus} - y_N\|^2 \leq \frac{1}{T_N} (F(y_0) - F_\star).$$

*Proof of Proposition 5.* To start, we give the claim about matrix  $C$ .

**Claim 1.**  $C := \begin{bmatrix} 0 & 0 \\ C & 0 \end{bmatrix}$  satisfies the following equality.

$$\begin{aligned} & 2 \left( \sum_{i=0}^N T_i (\mathbf{e}_i - \mathbf{e}_{i+1}) (\mathbf{e}_0 + \cdots + \mathbf{e}_i)^\top \right) C \\ &= \sum_{i=1}^N \frac{1}{T_{i-1}} \mathbf{f}_{N-i} \mathbf{f}_{N-i}^\top + \frac{1}{T_0} \mathbf{f}_N \mathbf{f}_N^\top - T_0 \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=1}^N \left( \frac{T_i^2}{T_{i-1}} + t_i^2 \sum_{j=0}^{i-2} \frac{1}{T_j} \right) \mathbf{e}_i \mathbf{e}_i^\top \end{aligned} \quad (98)$$

where  $\{\mathbf{e}_i\}_{i=0}^N$  is a any basis of  $\mathbb{R}^{N+1}$ ,  $\mathbf{e}_{N+1}$ ,  $\mathbf{f}_{-1}$  are zero vectors and  $\{\mathbf{f}_i\}_{i=0}^N$  are another basis of  $\mathbb{R}^{N+1}$  which is defined as

$$T_i (\mathbf{e}_i - \mathbf{e}_{i+1}) = \mathbf{f}_{N-i} - \mathbf{f}_{N-i-1}, \quad i = 0, 1, \dots, N.$$

The proof of Claim 1 will be provided after the proof of Proposition 5. To begin with, for the FSFOM with matrix  $(\frac{1}{\alpha} H_0 + \frac{1}{\alpha^2} C)^A$ , define the energy function

$$\mathcal{V}_k = \frac{F(y_0) - F(y_N^{\oplus, \alpha}) + \langle y_{-1}, y_0 \rangle^\alpha}{T_N} + \sum_{i=0}^{k-1} \frac{1}{T_{N-i-1}} \langle y_i, y_{i+1} \rangle^\alpha + \sum_{i=0}^{k-1} \left( \frac{1}{T_{N-i-1}} - \frac{1}{T_{N-i}} \right) \langle y_N, y_i \rangle^\alpha,$$

for  $k = 0, \dots, N$ .  $\{\mathcal{V}_k\}_{k=0}^N$  is dissipative since  $\langle \cdot, \cdot \rangle^\alpha \leq 0$ . This energy function is inspired by (3). Note that if we can prove

$$\frac{\alpha L}{2} \|y_N^{\oplus, \alpha} - y_N\|^2 \leq \mathcal{V}_N, \quad (99)$$

then

$$\frac{\alpha L}{2} \|y_N^{\oplus, \alpha} - y_N\|^2 \leq \mathcal{V}_N \leq \mathcal{V}_0 \leq \frac{1}{T_N} (F(y_0) - F(y_N^{\oplus, \alpha})) \leq \frac{1}{T_N} (F(y_0) - F_\star).$$

Thus, it is enough to show (99). Defining  $g_i := L\alpha(y_i - y_i^{\oplus, \alpha})$  for  $0 \leq i \leq N$  gives

$$\begin{aligned} \langle y_i, y_j \rangle^\alpha &= F(y_j^{\oplus, \alpha}) - F(y_i^{\oplus, \alpha}) + \langle g_j, y_i - y_j \rangle + \frac{1}{La} \langle g_j, g_j - g_i \rangle - \frac{1}{2L\alpha^2} \|g_j\|^2, \\ \langle y_{-1}, y_0 \rangle^\alpha &= F(y_0^{\oplus, \alpha}) - F(y_0) + \frac{1}{L} \frac{2\alpha - 1}{2\alpha^2} \|g_0\|^2. \end{aligned}$$

Plugging the above equalities provides

$$\begin{aligned}
& 2L\alpha^2 \left( \mathcal{V}_N - \frac{\alpha L}{2} \|y_N^{\oplus, \alpha} - y_N\|^2 \right) \\
&= 2L\alpha^2 \left[ \sum_{i=0}^{N-1} \frac{1}{T_{N-i-1}} \langle g_{i+1}, y_i - y_{i+1} \rangle + \sum_{i=0}^{N-1} \left( \frac{1}{T_{N-i-1}} - \frac{1}{T_{N-i}} \right) \langle g_i, y_N - y_i \rangle \right] + \frac{2\alpha-1}{T_N} \|g_0\|^2 - \alpha \|g_N\|^2 \\
&\quad + \sum_{i=0}^{N-1} \frac{1}{T_{N-i-1}} (2\alpha \langle g_{i+1}, g_{i+1} - g_i \rangle - \|g_{i+1}\|^2) + \sum_{i=0}^{N-1} \left( \frac{1}{T_{N-i-1}} - \frac{1}{T_{N-i}} \right) (2\alpha \langle g_i, g_i - g_N \rangle - \|g_i\|^2) \\
&= 2L\alpha^2 \left( \sum_{i=0}^{N-1} \frac{1}{T_{N-i-1}} \langle g_{i+1}, y_i - y_{i+1} \rangle + \sum_{i=0}^{N-1} \left( \frac{1}{T_{N-i-1}} - \frac{1}{T_{N-i}} \right) \langle g_i, y_N - y_i \rangle \right) \\
&\quad + \left[ \sum_{i=0}^{N-1} \frac{\alpha}{T_{N-i-1}} \|g_{i+1} - g_i\|^2 + \sum_{i=0}^{N-1} \left( \frac{\alpha}{T_{N-i-1}} - \frac{\alpha}{T_{N-i}} \right) \|g_i - g_N\|^2 - \alpha \|g_N\|^2 + \frac{\alpha}{T_N} \|g_N\|^2 \right] \\
&\quad + \sum_{i=0}^{N-1} \frac{1}{T_{N-i-1}} (-\alpha \|g_i\|^2 + (a-1) \|g_{i+1}\|^2) + \sum_{i=0}^{N-1} \left( \frac{1}{T_{N-i-1}} - \frac{1}{T_{N-i}} \right) (-\alpha \|g_N\|^2 + (\alpha-1) \|g_i\|^2) \\
&\quad + \frac{2\alpha-1}{T_N} \|g_0\|^2 - \frac{\alpha}{T_N} \|g_N\|^2 \\
&= 2L\alpha^2 \left( \sum_{i=0}^{N-1} \frac{1}{T_{N-i-1}} \langle g_{i+1}, y_i - y_{i+1} \rangle + \sum_{i=0}^{N-1} \left( \frac{1}{T_{N-i-1}} - \frac{1}{T_{N-i}} \right) \langle g_i, y_N - y_i \rangle \right) \\
&\quad + \left[ \sum_{i=0}^{N-1} \frac{\alpha}{T_{N-i-1}} \|g_{i+1} - g_i\|^2 + \sum_{i=0}^{N-1} \left( \frac{\alpha}{T_{N-i-1}} - \frac{\alpha}{T_{N-i}} \right) \|g_i - g_N\|^2 - \alpha \|g_N\|^2 + \frac{\alpha}{T_N} \|g_N\|^2 \right] \\
&\quad + \frac{\alpha}{T_N} \|g_0\|^2 - \sum_{i=0}^{N-1} \frac{1}{T_{N-1-i}} \|g_i\|^2 - \frac{1}{T_0} \|g_N\|^2.
\end{aligned}$$

Next, define  $g := [g_0 | g_1 \dots | g_N]$  and

$$\mathcal{H}_1 := \begin{bmatrix} 0 & 0 \\ (\frac{1}{\alpha} H_0 + \frac{1}{\alpha^2} C)^A & 0 \end{bmatrix}, \quad \mathcal{H}_0 := \begin{bmatrix} 0 & 0 \\ H_0 & 0 \end{bmatrix}.$$

By the same procedure as the proof of Theorem 1, we obtain  $2L\alpha^2 \left( \mathcal{V}_N - \frac{\alpha L}{2} \|y_N^{\oplus, \alpha} - y_N\|^2 \right) = \text{Tr}(g^\top g \mathcal{T}^\oplus)$  where

$$\begin{aligned}
\mathcal{T}^\oplus &= \alpha^2 \left[ \sum_{i=0}^N \frac{1}{T_{N-i}} (\mathcal{H}_1^\top (\mathbf{f}_i + \dots + \mathbf{f}_N) (\mathbf{f}_i - \mathbf{f}_{i-1})^\top + (\mathbf{f}_i - \mathbf{f}_{i-1}) (\mathbf{f}_i + \dots + \mathbf{f}_N)^\top \mathcal{H}_1) \right] \\
&\quad + \alpha \left[ \sum_{i=0}^N \frac{1}{T_{N-i}} ((\mathbf{f}_{i-1} - \mathbf{f}_i) (\mathbf{f}_{i-1} - \mathbf{f}_N)^\top + (\mathbf{f}_{i-1} - \mathbf{f}_N) (\mathbf{f}_{i-1} - \mathbf{f}_i)^\top) - \frac{1}{T_N} \mathbf{f}_0 \mathbf{f}_0^\top - \mathbf{f}_N \mathbf{f}_N^\top \right] \\
&\quad + \left[ \frac{\alpha}{T_N} \mathbf{f}_0 \mathbf{f}_0^\top - \sum_{i=0}^{N-1} \frac{1}{T_{N-1-i}} \mathbf{f}_i \mathbf{f}_i^\top - \frac{1}{T_0} \mathbf{f}_N \mathbf{f}_N^\top \right]
\end{aligned}$$

And consider the vector  $\mathbf{e}_i \in \mathbb{R}^{(N+1) \times 1}$  which is same with (29).

In the proof of Theorem 1 ((27), (28)), we have shown that

$$\begin{aligned}
& \frac{1}{2L} \sum_{i=0}^N \frac{1}{T_{N-i}} (\mathcal{H}_1^\top (\mathbf{f}_i + \dots + \mathbf{f}_N) (\mathbf{f}_i - \mathbf{f}_{i-1})^\top + (\mathbf{f}_i - \mathbf{f}_{i-1}) (\mathbf{f}_i + \dots + \mathbf{f}_N)^\top \mathcal{H}_1) \\
&= \frac{1}{2L} \left( \frac{\mathcal{H}_0^\top}{\alpha} + \frac{C^\top}{\alpha^2} \right) \left[ \sum_{i=0}^N T_i (\mathbf{e}_0 + \dots + \mathbf{e}_i) (\mathbf{e}_i - \mathbf{e}_{i+1})^\top \right] + \frac{1}{2L} \left[ \sum_{i=0}^N T_i (\mathbf{e}_i - \mathbf{e}_{i+1}) (\mathbf{e}_0 + \dots + \mathbf{e}_i)^\top \right] \left( \frac{\mathcal{H}_0}{\alpha} + \frac{C}{\alpha^2} \right)
\end{aligned}$$

and

$$\begin{aligned}
& \frac{1}{2L} \sum_{i=0}^N \frac{1}{T_{N-i}} ((\mathbf{f}_{i-1} - \mathbf{f}_i)(\mathbf{f}_{i-1} - \mathbf{f}_N)^\top + (\mathbf{f}_{i-1} - \mathbf{f}_N)(\mathbf{f}_{i-1} - \mathbf{f}_i)^\top) - \frac{1}{2T_N L} \mathbf{f}_0 \mathbf{f}_0^\top - \frac{1}{2L} \mathbf{f}_N \mathbf{f}_N^\top \\
&= -\frac{1}{2L} \left( \sum_{i=0}^N (T_i - T_{i-1}) \mathbf{e}_i \right) \left( \sum_{i=0}^N (T_i - T_{i-1}) \mathbf{e}_i \right)^\top \\
&\quad + \frac{1}{2L} \left[ \sum_{i=0}^N T_i ((\mathbf{e}_i - \mathbf{e}_{i+1}) \mathbf{e}_i^\top + \mathbf{e}_i (\mathbf{e}_i - \mathbf{e}_{i+1})^\top) - T_N \mathbf{e}_N \mathbf{e}_N^\top \right]
\end{aligned}$$

under the above transformation. Therefore,  $\mathcal{T}^\oplus$  can be expressed as

$$\begin{aligned}
\mathcal{T}^\oplus &= \alpha^2 \left( \left( \frac{\mathcal{H}_0^\top}{\alpha} + \frac{\mathcal{C}^\top}{\alpha^2} \right) \left[ \sum_{i=0}^N T_i (\mathbf{e}_0 + \cdots + \mathbf{e}_i) (\mathbf{e}_i - \mathbf{e}_{i+1})^\top \right] + \left[ \sum_{i=0}^N T_i (\mathbf{e}_i - \mathbf{e}_{i+1}) (\mathbf{e}_0 + \cdots + \mathbf{e}_i)^\top \right] \left( \frac{\mathcal{H}_0}{\alpha} + \frac{\mathcal{C}}{\alpha^2} \right) \right) \\
&\quad + \alpha \left( - \left( \sum_{i=0}^N (T_i - T_{i-1}) \mathbf{e}_i \right) \left( \sum_{i=0}^N (T_i - T_{i-1}) \mathbf{e}_i \right)^\top + \left[ \sum_{i=0}^N T_i ((\mathbf{e}_i - \mathbf{e}_{i+1}) \mathbf{e}_i^\top + \mathbf{e}_i (\mathbf{e}_i - \mathbf{e}_{i+1})^\top) - T_N \mathbf{e}_N \mathbf{e}_N^\top \right] \right) \\
&\quad + \left[ \frac{\alpha}{T_N} \mathbf{f}_0 \mathbf{f}_0^\top - \sum_{i=0}^{N-1} \frac{1}{T_{N-1-i}} \mathbf{f}_i \mathbf{f}_i^\top - \frac{1}{T_0} \mathbf{f}_N \mathbf{f}_N^\top \right] \\
&= \alpha \mathcal{A} + \mathcal{B} + \frac{\alpha}{T_N} \mathbf{f}_0 \mathbf{f}_0^\top
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{A} &= \mathcal{H}_0^\top \left[ \sum_{i=0}^N T_i (\mathbf{e}_0 + \cdots + \mathbf{e}_i) (\mathbf{e}_i - \mathbf{e}_{i+1})^\top \right] + \left[ \sum_{i=0}^N T_i (\mathbf{e}_i - \mathbf{e}_{i+1}) (\mathbf{e}_0 + \cdots + \mathbf{e}_i)^\top \right] \mathcal{H}_0 \\
&\quad - \left( \sum_{i=0}^N (T_i - T_{i-1}) \mathbf{e}_i \right) \left( \sum_{i=0}^N (T_i - T_{i-1}) \mathbf{e}_i \right)^\top + \left[ \sum_{i=0}^N T_i ((\mathbf{e}_i - \mathbf{e}_{i+1}) \mathbf{e}_i^\top + \mathbf{e}_i (\mathbf{e}_i - \mathbf{e}_{i+1})^\top) - T_N \mathbf{e}_N \mathbf{e}_N^\top \right]
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{B} &= \mathcal{C}^\top \left[ \sum_{i=0}^N T_i (\mathbf{e}_0 + \cdots + \mathbf{e}_i) (\mathbf{e}_i - \mathbf{e}_{i+1})^\top \right] + \left[ \sum_{i=0}^N T_i (\mathbf{e}_i - \mathbf{e}_{i+1}) (\mathbf{e}_0 + \cdots + \mathbf{e}_i)^\top \right] \mathcal{C} \\
&\quad - \sum_{i=0}^{N-1} \frac{1}{T_{N-1-i}} \mathbf{f}_i \mathbf{f}_i^\top - \frac{1}{T_0} \mathbf{f}_N \mathbf{f}_N^\top.
\end{aligned}$$

To calculate  $\mathcal{A}$ , expand the energy function (91). Then we obtain

$$\mathcal{A} = \sum_{i=0}^{N-1} (2T_i - t_i^2) \mathbf{e}_i \mathbf{e}_i^\top + (T_N - t_N^2) \mathbf{e}_N \mathbf{e}_N^\top.$$

Moreover, by Claim 1,

$$\mathcal{B} = -T_0 \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=1}^N \left( \frac{T_i^2}{T_{i-1}} + t_i^2 \sum_{j=0}^{i-2} \frac{1}{T_j} \right) \mathbf{e}_i \mathbf{e}_i^\top.$$

Combining above results and  $T_N \mathbf{e}_N = \mathbf{f}_0$ , we achieve

$$\mathcal{T}^\oplus = \alpha \sum_{i=0}^N (2T_i - t_i^2) \mathbf{e}_i \mathbf{e}_i^\top - T_0 \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=1}^N \left( \frac{T_i^2}{T_{i-1}} + t_i^2 \sum_{j=0}^{i-2} \frac{1}{T_j} \right) \mathbf{e}_i \mathbf{e}_i^\top$$

and the condition (96) makes each coefficient of  $\mathbf{e}_i \mathbf{e}_i^\top$  nonnegative, which gives  $\mathcal{T} \succeq 0$ . To achieve the iteration formula (SFG-family), consider the FSFOM (87) with  $\frac{1}{\alpha} H_0 + \frac{1}{\alpha^2} C$ , which can be

expressed as

$$x_{k+1} = x_k^{\oplus, \alpha} + \beta_k (x_k^{\oplus, \alpha} - x_{k-1}^{\oplus, \alpha}) + \gamma_k (x_k^{\oplus, \alpha} - x_k), \quad k = 0, \dots, N-1,$$

$$\beta_k = \frac{t_{k+1}(T_k - t_k)}{t_k T_{k+1}}, \quad \gamma_k = \frac{t_{k+1}(t_k^2 - T_k)}{t_k T_{k+1}} + \frac{1}{\alpha} c_{k+1, k}.$$

For last step, We make an observation that Proposition 1 can be applied when all  $z^+$  terms changed into  $z^{\oplus, \alpha}$ . Proposition 1 gives the H-dual.  $\square$

*Proof of Claim 1.* The left hand side of (98) is

$$\begin{aligned} & 2 \left( \sum_{i=0}^N T_i (\mathbf{e}_i - \mathbf{e}_{i+1}) (\mathbf{e}_0 + \dots + \mathbf{e}_i)^\top \right) \mathcal{C} \\ &= 2 \left( \sum_{i=0}^N T_i (\mathbf{e}_i - \mathbf{e}_{i+1}) (\mathbf{e}_0 + \dots + \mathbf{e}_i)^\top \right) \left( \sum_{k>i} c_{k,i} \mathbf{e}_k \mathbf{e}_i^\top \right) \\ &= 2 \left( \sum_{i=0}^N (\mathbf{f}_{N-i} - \mathbf{f}_{N-i+1}) (\mathbf{e}_0 + \dots + \mathbf{e}_i)^\top \right) \left( \sum_{k>i} c_{k,i} \mathbf{e}_k \mathbf{e}_i^\top \right) \\ &= 2 \left( \sum_{i=0}^N \mathbf{f}_{N-i} \mathbf{e}_i^\top \right) \left( \sum_{k>i} c_{k,i} \mathbf{e}_k \mathbf{e}_i^\top \right) \\ &= 2 \sum_{i=0}^N \left( \sum_{k=i+1}^N c_{k,i} \mathbf{f}_{N-k} \right) \mathbf{e}_i^\top \end{aligned}$$

Now we calculate the right-hand side of (98). By plugging

$$\mathbf{f}_{N-i} = T_i \mathbf{e}_i + t_{i+1} \mathbf{e}_{i+1} + \dots + t_N \mathbf{e}_N,$$

$$\sum_{i=1}^N \frac{1}{T_{i-1}} \mathbf{f}_{N-i} \mathbf{f}_{N-i}^\top + \frac{1}{T_0} \mathbf{f}_N \mathbf{f}_N^\top - T_0 \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=1}^N \left( \frac{T_i^2}{T_{i-1}} + t_i^2 \sum_{j=0}^{i-2} \frac{1}{T_j} \right) \mathbf{e}_i \mathbf{e}_i^\top = 2 \sum_{i=0}^{N-1} \mathbf{a}_i \mathbf{e}_i^\top$$

where

$$\begin{aligned} \mathbf{a}_0 &= t_1 \mathbf{e}_1 + \dots + t_N \mathbf{e}_N, \\ \mathbf{a}_i &= \left( \frac{t_i}{T_0} + \sum_{j=0}^{i-2} \frac{t_i}{T_j} + \frac{T_i}{T_{i-1}} \right) (t_{i+1} \mathbf{e}_{i+1} + \dots + t_N \mathbf{e}_N), \quad i = 1, \dots, N-1. \end{aligned}$$

Now we claim that  $\mathbf{a}_i = \sum_{k=i+1}^N c_{k,i} \mathbf{f}_{N-k}$  for  $i = 0, \dots, N-1$ . Note that

$$\begin{aligned} \sum_{k=i+1}^N c_{k,i} \mathbf{f}_{N-k} &= \sum_{k=i+1}^N c_{k,i} (T_k \mathbf{e}_k + t_{k+1} \mathbf{e}_{k+1} + \dots + t_N \mathbf{e}_N) \\ &= \sum_{k=i+1}^N (c_{k,i} T_k + c_{k-1,i} t_k + \dots + c_{i+1,i} t_k) \mathbf{e}_k. \end{aligned}$$

Coefficients of  $\mathbf{e}_{i+1}$  are coincides since

$$\begin{aligned} c_{1,0} T_1 &= t_1, \\ c_{i+1,i} T_{i+1} &= t_{i+1} \left( \frac{t_i}{T_0} + \sum_{j=0}^{i-2} \frac{t_i}{T_j} + \frac{T_i}{T_{i-1}} \right) \quad i = 1, \dots, N-1. \end{aligned}$$

Now assume

$$\frac{t_j}{t_{j+1}} = \frac{c_{j,i} T_j + c_{j-1,i} t_j + \dots + c_{i+1,i} t_j}{c_{j+1,i} T_{j+1} + c_{j,i} t_{j+1} + \dots + c_{i+1,i} t_{j+1}}. \quad (100)$$

By multiplying the above equation recursively, we obtain

$$\frac{t_{i+1}}{t_{j+1}} = \frac{c_{i+1,i}T_i}{c_{j+1,i}T_{j+1} + c_{j,i}t_{j+1} + \cdots + c_{i+1,i}t_{j+1}}$$

for  $1 \leq j \leq N-1$ , which implies  $\mathbf{a}_i = \sum_{k=i+1}^N c_{k,i} \mathbf{f}_{N-k}$ .

To prove (100), expand and obtain

$$t_{j+1}c_{j,i}T_j = t_j(c_{j+1,i}T_{j+1} + c_{j,i}t_{j+1}).$$

Above equation holds due to the recursive formula of  $C$  (97).  $\square$

#### D.4 Instances of SFG family

**Derivation of (SFG)** Here, we will show that (SFG) is the instance of SFG family, under the choice of

$$\alpha = 4, \quad T_i = \frac{(i+2)(i+3)}{4}, \quad 0 \leq i \leq N.$$

First,  $t_i = \frac{i+2}{2}$  for  $i \geq 1$  and  $t_0 = \frac{3}{2}$ . Also (96) holds since

$$4 \left( 3 - \frac{9}{4} \right) \geq \frac{3}{2}$$

$$4 \left( \frac{(k+2)(k+3)}{2} - \frac{(k+2)^2}{4} \right) \geq \frac{(k+2)^2(k+3)^2}{\frac{16}{(k+1)(k+2)}} + \frac{(k+2)^2}{4} \left( \frac{2}{3} + 4 \left( \frac{1}{2} - \frac{1}{k+1} \right) \right) \quad k = 1, \dots, N.$$

Plugging above values into (92) and (97), we obtain

$$h_{k+1,i} = \begin{cases} 1 + \frac{1}{4} & i = k, k = 0 \\ 1 + \frac{k}{k+4} & i = k, k = 1, \dots, N-1 \\ \frac{k+1}{k+4} (h_{k,k-1} - 1) & i = k-1 \\ \frac{k+1}{k+4} h_{k,i} & i = 0, \dots, k-2 \end{cases}.$$

and

$$c_{k+1,i} = \begin{cases} \frac{1}{2} & i = 0, k = 0 \\ \frac{2(4k+5)}{3(k+4)} & i = k, k = 1, \dots, N-1 \\ \frac{k+1}{k+4} c_{k,i} & i = 0, \dots, k-1, i = 2, \dots, N-1 \end{cases}.$$

Other terms come directly, and

$$\begin{aligned} c_{k+1,k} &= \frac{t_{k+1}}{T_{k+1}} \left( \frac{t_k}{T_0} + t_k \sum_{j=0}^{k-2} \frac{1}{T_j} + \frac{T_k}{T_{k-1}} \right) \\ &= \frac{t_{k+1}}{T_{k+1}} \left( \frac{t_k}{T_0} + t_k \sum_{j=0}^{k-1} \frac{1}{T_j} + 1 \right) \\ &= \frac{\frac{k+3}{2}}{\frac{(k+3)(k+4)}{4}} \left( \frac{k+2}{2} \times \frac{2}{3} + \frac{k+2}{2} \left( \frac{4}{2 \times 3} + \cdots + \frac{4}{(k+1)(k+2)} \right) + 1 \right) \\ &= \frac{2}{k+4} \left( \frac{k+2}{3} + \frac{k+2}{2} \frac{2k}{k+2} + 1 \right) \\ &= \frac{2(4k+5)}{3(k+4)}. \end{aligned}$$

To sum up, the matrix  $\{g_{k,i}\}_{0 \leq i < k \leq N} := \frac{1}{4}H_0 + \frac{1}{16}C$  satisfies the following recursive formula.

$$g_{k+1,i} = \begin{cases} \frac{1}{4} \left(1 + \frac{3}{8}\right) & i = 0, k = 0 \\ \frac{1}{4} \left(1 + \frac{10k+5}{6(k+4)}\right) & i = k, k = 1, \dots, N-1 \\ \frac{k+1}{k+4} \left(g_{k,i} - \frac{1}{4}\right) & i = k-1 \\ \frac{k+1}{k+4} g_{k,i} & i = 0, \dots, k-2 \end{cases}.$$

Note that

$$\begin{aligned} g_{k+1,k} &= \frac{1}{4} \left(1 + \frac{k}{k+4}\right) + \frac{1}{16} \left(\frac{2(4k+5)}{3(k+4)}\right) \\ &= \frac{1}{4} \left(1 + \frac{k}{k+4} + \frac{4k+5}{6(k+4)}\right) \\ &= \frac{1}{4} \left(1 + \frac{10k+5}{6(k+4)}\right). \end{aligned}$$

Therefore, the FSFOM with  $\{g_{k,i}\}_{0 \leq i < k \leq N}$  can be expressed as

$$\begin{aligned} x_1 &= x_0^{\oplus,4} + \frac{1}{4} \cdot (x_0^{\oplus,4} - x_{-1}^{\oplus,4}) + \frac{1}{8} (x_0^{\oplus,4} - x_0), \\ x_{k+1} &= x_k^{\oplus,4} + \frac{k+1}{k+4} (x_k^{\oplus,4} - x_{k-1}^{\oplus,4}) + \frac{4k-1}{6(k+4)} (x_k^{\oplus,4} - x_k), \quad k = 1, \dots, N-1 \end{aligned}$$

where  $x_{-1}^{\oplus,4} = x_0$ . To obtain the H-dual, we apply Proposition 1.

$$\begin{aligned} y_{k+1} &= y_k^{\oplus,4} + \frac{(N-k+1)(2N-2k-1)}{(N-k+3)(2N-2k+1)} (y_k^{\oplus,4} - y_{k-1}^{\oplus,4}) + \frac{(4N-4k-1)(2N-2k-1)}{6(N-k+3)(2N-2k+1)} (y_k^{\oplus,4} - y_k) \\ y_N &= y_{N-1}^{\oplus,4} + \frac{3}{10} (y_{N-1}^{\oplus,4} - y_{N-2}^{\oplus,4}) + \frac{3}{40} (y_{N-1}^{\oplus,4} - y_{N-1}) \end{aligned}$$

where  $k = 0, \dots, N-2$  and  $y_{-1}^{\oplus,4} = y_0$ .

**Fastest method among the SFG family via a numerical choice of  $a$**  In this section, we give simple form of SFG, when all inequality conditions in (96) holds as equalities.  $\{g_{k,i}\}_{0 \leq i < k \leq N}$  becomes

$$g_{1,0} = \frac{1}{\alpha} \left(1 + \frac{(t_0-1)t_1}{T_1}\right) + \frac{1}{\alpha^2} \left(\frac{t_1}{T_1}\right)$$

and for  $k > 0$ ,

$$\begin{aligned} g_{k+1,k} &= \frac{1}{\alpha} \left(1 + \frac{(t_k-1)t_{k+1}}{T_{k+1}}\right) + \frac{1}{\alpha^2} \frac{t_{k+1}}{T_{k+1}} \left(t_k \left(\frac{1}{T_0} + \frac{1}{T_0} + \dots + \frac{1}{T_{k-2}}\right) + \frac{T_k}{T_{k-1}}\right) \\ &= \frac{1}{\alpha} \left(1 + \frac{(t_k-1)t_{k+1}}{T_{k+1}}\right) + \frac{1}{\alpha^2} \frac{t_{k+1}}{T_{k+1}} \left(\frac{\alpha(2T_k - t_k^2) - \frac{T_k^2}{T_{k-1}}}{t_k} + \frac{T_k}{T_{k-1}}\right) \\ &= \frac{1}{\alpha} \left(1 + \frac{(t_k-1)t_{k+1}}{T_{k+1}}\right) + \frac{1}{\alpha^2} \frac{t_{k+1}}{T_{k+1}} \left(\frac{\alpha(2T_k - t_k^2)}{t_k} - \frac{T_k}{t_k}\right) \\ &= \frac{1}{\alpha} + \frac{1}{\alpha} \frac{t_{k+1}}{T_{k+1}} \left(t_k - 1 + \frac{2T_k - t_k^2}{t_k} - \frac{T_k}{t_k}\right) - \left(\frac{1}{\alpha^2} - \frac{1}{\alpha}\right) \frac{t_{k+1}T_k}{T_{k+1}t_k} \\ &= \frac{1}{\alpha} + \frac{1}{\alpha} \frac{(T_k - t_k)t_{k+1}}{t_k T_{k+1}} + \left(\frac{1}{\alpha} - \frac{1}{\alpha^2}\right) \frac{t_{k+1}T_k}{T_{k+1}t_k}, \\ g_{k+1,k-1} &= \frac{t_{k+1}(T_k - t_k)}{t_k T_{k+1}} \left(g_{k,k-1} - \frac{1}{\alpha}\right), \\ g_{k+1,i} &= \frac{t_{k+1}(T_k - t_k)}{t_k T_{k+1}} g_{k,i}, \quad i = k-2, \dots, 0. \end{aligned}$$

Hence FSFOM with  $\frac{1}{\alpha}H_0 + \frac{1}{\alpha^2}C$  is

$$x_{k+1} = x_k^{\oplus, \alpha} + \frac{T_{k-1}t_{k+1}}{t_k T_{k+1}}(x_k^{\oplus, \alpha} - x_{k-1}^{\oplus, \alpha}) + \left(1 - \frac{1}{\alpha}\right) \frac{t_{k+1}T_k}{T_{k+1}t_k}(x_k^{\oplus, \alpha} - x_k)$$

where  $T_{-1} = \frac{2a-1}{\alpha^2}$ . By using Proposition 1, we obtain H-dual.

$$y_{k+1} = y_k^{\oplus, \alpha} + \beta'_k(y_k^{\oplus, \alpha} - y_{k-1}^{\oplus, \alpha}) + \left(1 - \frac{1}{\alpha}\right) \frac{T_{N-k}}{T_{N-k+1}}\beta_{N-k}(y_k^{\oplus, \alpha} - y_k)$$

$$\beta'_k = \frac{T_{N-k-1}t_{N-k}(T_{N-k-2} + (1 - \frac{1}{\alpha})T_{N-k-1})}{t_{N-k-1}T_{N-k}(T_{N-k-1} + (1 - \frac{1}{\alpha})T_{N-k})}, \quad k = 0, \dots, N-1.$$

Since all equality holds at (96), the above FSFOM achieves the fastest rate among the SFG family under fixed  $\alpha$ .<sup>7</sup> Now we optimize  $a$  to achieve the fastest convergence rate. Combine (93) and the result of Proposition 5 to obtain

$$\min_{v \in \partial F(y_N^{\oplus})} \|v\|^2 \leq L^2(\alpha + 1)^2 \|y_N - y_N^{\oplus, \alpha}\|^2 \leq L^2 \frac{2(\alpha + 1)^2}{\alpha T_N} (F(y_0) - F_\star).$$

To achieve the tightest convergence guarantee, we solve the following optimization problem under the fixed  $N$ .

$$\begin{aligned} & \underset{\alpha}{\text{minimize}} && \frac{2(\alpha + 1)^2}{\alpha T_N} \\ & \text{subject to.} && \alpha(2T_0 - t_0^2) = T_0, \\ & && \alpha(2T_k - t_k^2) = \frac{T_k^2}{T_{k-1}} + t_k^2 \left( \frac{1}{T_0} + \sum_{i=0}^{k-2} \frac{1}{T_i} \right) \quad k = 1, \dots, N. \end{aligned}$$

Denote the solution of the above optimization problem as  $R(\alpha, N)$ . Since  $R(\alpha, N)$  depends on  $a$  and cannot achieve a closed-form solution, we numerically choose  $a$  and observe an asymptotic behavior of  $R(\alpha, N)$ . By choosing  $\alpha = 3.8$ , asymptotic rate of  $R(3.8, N)$  is about  $\frac{46}{N^2}$ .

## E Broader Impacts

Our work focuses on the theoretical aspects of convex optimization algorithms. There are no negative social impacts that we anticipate from our theoretical results.

## F Limitations

Our analysis concerns  $L$ -smooth convex functions. Although this assumption is standard in optimization theory, many functions that arise in machine learning practice are neither smooth nor convex.

---

<sup>7</sup>In fact, when  $a = 1$ , the FSFOM becomes FISTA-G [31].