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

Computer-Assisted Design of Accelerated Composite Optimization Methods: OptISTA

Uijeong Jang · Shuvomoy Das Gupta · Ernest K. Ryu

Received: date / Accepted: date

Abstract The accelerated composite optimization method FISTA (Beck, Teboulle 2009) is suboptimal by a constant factor, and we present a new method OptISTA that improves FISTA by a constant factor of 2. The performance estimation problem (PEP) has recently been introduced as a new computer-assisted paradigm for designing optimal first-order methods. In this work, we present a double-function stepsize-optimization PEP methodology that poses the optimization over fixed-step first-order methods for composite optimization as a finite-dimensional nonconvex QCQP, which can be practically solved through spatial branch-and-bound algorithms, and use it to design the exact optimal method OptISTA for the composite optimization setup. We then establish the exact optimality of OptISTA with a lower-bound construction that extends the semi-interpolated zero-chain construction (Drori, Taylor 2022) to the double-function setup of composite optimization. By establishing exact optimality, our work concludes the search for the fastest first-order methods, with respect to the performance measure of worst-case function value suboptimality, for the proximal, projected-gradient, and proximal-gradient setups involving a smooth convex function and a closed proper convex function.

1 Introduction

Since the seminal work by Nesterov on accelerated gradient methods [57] and by Nemirovsky on matching complexity lower bounds [56, 55, 53], accelerated first-order methods have been central to the theory and practice of large-scale convex optimization. Recently, the performance estimation problem (PEP) [26, 69] has been introduced as a new computer-assisted paradigm for designing optimal first-order methods and was used to discover OGM, which surprisingly achieves a factor-2 speedup over Nesterov's method [26, 40], and several other new accelerated methods such as OGM-G [43], APPM [39], and EAG [73].

However, the PEP methodology of searching over first-order methods to find globally optimal ones has so far been limited to the unconstrained optimization setup with a single function or monotone operator; it has not been extended to the composite optimization setup, where the objective is a

Uijeong Jang
Seoul National University
E-mail: uijeong97@snu.ac.kr

Shuvomoy Das Gupta
Columbia University
E-mail: sd3871@columbia.edu

Ernest K. Ryu
University of California, Los Angeles
E-mail: eryu@math.ucla.edu

sum of two functions, or in the constrained optimization setup, where methods use projections onto the constraint. In particular, whether it is possible to improve FISTA [6] or projected-gradient-type methods and achieve a speedup similar to OGM was an open problem.

In this work, we present a PEP methodology that poses the optimization over fixed-step first-order methods for composite optimization as a finite-dimensional nonconvex QCQP, which can be practically solved through spatial branch-and-bound algorithms [1, 48, 38, 16]. Using this methodology, we obtain OptISTA, a new composite optimization method that improves the prior state-of-the-art rates by a constant factor, including that of FISTA [6].

At the same time, improving the prior lower bound constructions of Nesterov and Nemirovsky [58, 55, 54] has been an active area of research that parallels the PEP line of work. New lower bounds have certified the exact optimality of methods such as a variant of Kelley's cutting plane method [27], OGM [26, 40], and ITEM [67]. However, these prior lower bounds do not exactly match the upper bound of OptISTA. In this work, we provide a lower-bound construction that establishes the exact optimality of OptISTA for both the proximal-gradient and projected-gradient setups. By exact optimality, we mean the upper and lower bounds are exactly equal (rather than matching only up to a constant). The key insight is to extend the semi-interpolated zero-chain construction to the double-function setup of composite optimization.

Preliminaries and notation. Write $\mathbb{R}^d$ for the underlying Euclidean space. Write $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ to denote the standard inner product and norm on $\mathbb{R}^d$. For $a, b \in \mathbb{Z}$, denote

$$[a : b] = \{a, a+1, a+2, \dots, b-1, b\} \subset \mathbb{Z}.$$

We follow standard convex-analytical definitions [8, 58]. A set $S \subseteq \mathbb{R}^d$ is convex if

$$\theta x + (1-\theta) y \in S \text{ for any } x, y \in S \text{ and } \theta \in [0,1].$$

A function $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex if

$$f\left(\theta x + (1-\theta) y\right) \leq \theta f(x) + (1-\theta) f(y) \text{ for all } x, y \in \mathbb{R}^d \text{ and } \theta \in (0,1).$$

The subdifferential of $f \colon \mathbb{R}^d \to \mathbb{R}$ at x, denoted by $\partial f(x)$, is

$$\partial f(x) = \{g \in \mathbb{R}^d \,|\, f(y) \geq f(x) + \langle g, y - x \rangle, \, \forall \, y \in \mathbb{R}^d\}.$$

An optimization method is usually designed for a specific class of functions. The class of L-smooth convex functions is denoted by $\mathcal{F}_{0,L}$, and any function f in this class is defined by

$$f(y) \geq f(x) + \langle \nabla f(x), \, y - x \rangle + (1/2L) \|\nabla f(x) - \nabla f(y)\|^2$$

for all $x, y \in \mathbb{R}^d$ [58, Theorem 2.1.5, (2.1.10)]. The class of closed, convex, and proper functions that are potentially nonsmooth is denoted by $\mathcal{F}_{0,\infty}$, and any function $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ in this class is defined by

$$h(y) \geq h(x) + \langle u, \, y - x \rangle$$

for all $x, y \in \mathbb{R}^d$ and for all $u \in \partial h(x)$ [69, Definition 3.1]. For a closed, convex, and proper function h and $\gamma > 0$, we define the proximal operator $\mathbf{prox}_{\gamma h} \colon \mathbb{R}^d \to \mathbb{R}^d$ as

$$\mathbf{prox}_{\gamma h}(x) = \operatorname*{argmin}_{z \in \mathbb{R}^d} \left\{ h(z) + \frac{1}{2\gamma} \|z - x\|^2 \right\}.$$

1.1 Prior work

FISTA and its variants. To solve composite convex optimization problems, one of the first methods is the proximal gradient method that can be traced back to the works by [11,62,47] and has a convergence rate of $\mathcal{O}(1/N)$ in functions value suboptimality. When the nonsmooth function in the composite setup is the ℓ^1-norm, the proximal gradient method is called iterative shrinkage thresholding algorithm (ISTA) [17,36,72,28]. The celebrated FISTA method due to Beck and Teboulle [6] accelerates the convergence rate to $\mathcal{O}(1/N^2)$ in function values. MFISTA is a variant of FISTA that maintains a monotonically nonincreasing sequence of function values and preserves the same convergence rate of FISTA [5]. A comprehensive review of FISTA and FISTA-based methods can found in the monograph [4, Chapter 10]. In [42], Kim and Fessler propose two FISTA-variants called GFPGM and FPGM-OCG, both having same convergence rate $\mathcal{O}(1/N^2)$ but a worse constant compared to FISTA. Finally, in [68, Section 4.2], a new method called FPGM2 is proposed which has worse convergence rate than OptISTA approximately by a factor of 2.

Proximal point method and its variants. The proximal point method, which traces its origins back to the 1970s, has been the subject of extensive research in the field of optimization [50,51,63, 10]. Güler's 1991 work [34] studies the convergence rate for the proximal point method in terms of reducing the function values of convex functions. An accelerated proximal point method for maximally monotone operators through the lens of monotone operator theory was proposed independently in [46, 39].

In Section 4.4, we describe a proximal method OPPA, parameterized by $\gamma_0, \gamma_1, \ldots, \gamma_{N-1}$ and establish its exact optimality by providing an exact matching lower bound. In [35], Güler proposed the so-called Güler's second method and provided a bound that improves upon Güler's prior work [34]. Güler's second method turns out to be an instance of OPPA with $\gamma_0 = \gamma_1 = \cdots = \gamma_{N-1}$. Monteiro and Svaiter proposed Accelerated Hybrid Proximal Extragradient (A-HPE) [52] as an inexact accelerated proximal method that generalizes Güler's second method. A-HPE with its parameter $\sigma = 0$ turns out to include OPPA when the proximal operators are evaluated exactly. Barré, Taylor, and Bach [2] proposed Optimized Relatively Inexact Proximal Point Algorithm (ORI-PPA) also as an inexact accelerated proximal method generalizing Güler's second method. ORI-PPA with its parameter $\sigma = 0$ also turns out to be equivalent to OPPA when proximal operators are evaluated exactly. Although A-HPE and ORI-PPA reduce to the same method OPPA under the aforementioned conditions, their analyses are slightly different. In Section 4.4, we reference the analysis of [2] as it is tighter than the analysis of [52] by a factor of 2 in the case of exact proximal evaluations.

Lower bounds. Leveraging the information-based complexity framework [56,55], iteration complexity lower bounds have been thoroughly explored for single-function first-order convex optimization methods [58,55,23,12,24,13,22,25]. In this study, we extend the semi-interpolated zero-chain construction introduced by [25] to the double-function scenario of composite optimization.

Performance estimation problem: A foundation for optimal methods. Our primary methodology for constructing optimal methods in this paper is rooted in the well-established performance estimation problem (PEP) framework, a computer-assisted approach for analyzing and designing optimization methods, with a particular focus on first-order methods. Since the pioneering work of [26], PEP has been extensively employed to discover and analyze numerous first-order and operator splitting methods [69,68,66,2,3,18,19,22,64,40,41,42,39,43,16,15]. The majority of these works employ the SDP-based methodology introduced by [26,69,68], demonstrating that computing the worst-case performance measure for a known fixed-step first-order method is equivalent to solving a convex SDP. The work [16] developed a nonconvex QCQP framework to determine optimal fixed-step first-order methods for a single-function setup.

1.2 Contributions

This paper presents two major contributions: one concrete and one conceptual. The first contribution is the exact optimal accelerated composite optimization method OptISTA, the combination of the upper and lower bound results. The second major contribution is the methodologies: the double-function stepsize-optimization PEP, used to discover OptISTA, and the double-function semi-interpolated zero-chain construction, used to establish the exact matching lower bound. In a sense, the upper and lower bounds of OptISTA is merely one application demonstrating the strength of the presented methodology.

Finally, as an auxiliary contribution, we consider the proximal minimization setup, where the proximal oracle but not the gradient oracle is used. We show that the prior method OPPA [52, 2], which we review in Section 4.4, is exactly optimal by adapting the semi-interpolated zero-chain construction to proximal setup and providing an exact matching lower bound.

2 New exact optimal method

We now present the main result: OptISTA. In this section, we first state the method and describe its convergence result. Later in Section 3, we will present the methodology used to discover the method. Later in Section 4, we will provide complexity lower bounds establishing exact optimality of OptISTA and OPPA.

2.1 OptISTA: Exact optimal composite optimization method

Consider the composite minimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x)+h(x), \tag{$\mathcal{P}$}$$

where $f\colon \mathbb{R}^d\to\mathbb{R}$ is L-smooth convex and $h\colon \mathbb{R}^d\to\mathbb{R}\cup\{\infty\}$ is closed, convex, and proper (possibly nonsmooth). Assume $(\mathcal{P})$ has a minimizer $x_\star$ (not necessarily unique).

Let $N>0$ be the total iteration count and $L>0$. We propose **Opt**imal **I**terative **S**hrinkage **T**hresholding **A**lgorithm

$$\begin{aligned}
y_{i+1} &= \mathbf{prox}_{\frac{\gamma_i}{L}h}\Big(y_i-\frac{\gamma_i}{L}\nabla f(x_i)\Big)\\
z_{i+1} &= x_i+\frac{1}{\gamma_i}\left(y_{i+1}-y_i\right)\\
x_{i+1} &= z_{i+1}+\frac{\theta_i-1}{\theta_{i+1}}\left(z_{i+1}-z_i\right)+\frac{\theta_i}{\theta_{i+1}}\left(z_{i+1}-x_i\right)
\end{aligned} \tag{OptISTA}$$

for $i=0,\dots,N-1$, where $z_0=y_0=x_0\in\mathbb{R}^d$ is a starting point and

$$\gamma_i=\frac{2\theta_i}{\theta_N^2}\left(\theta_N^2-2\theta_i^2+\theta_i\right)\ \text{ for } i=0,\dots,N-1,\quad \theta_i=\begin{cases}1 & \text{if } i=0,\\ \frac{1+\sqrt{1+4\theta_{i-1}^2}}{2} & \text{if } 1\le i\le N-1,\\ \frac{1+\sqrt{1+8\theta_{N-1}^2}}{2} & \text{if } i=N.\end{cases}$$

Since the γ-coefficients, the proximal stepsizes, depend on θ_N, the total iteration count N must be chosen prior to the start of the method. To practically implement OptISTA, one should pre-compute θ_N in one for-loop and then perform the main iteration in a second for-loop. (The for-loops are not nested.)

Theorem 1 *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be L-smooth convex and $h\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ be closed, convex, and proper. Assume $x_\star \in \operatorname{argmin}(f+h)$ exists. Let $N > 0$. Then,* OptISTA *exhibits the rate*

$$f(y_N) + h(y_N) - f(x_\star) - h(x_\star) \leq \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2 - 1)} \leq \frac{L\|x_0 - x_\star\|^2}{(N+1)^2},$$

where θ_N is as defined for OptISTA*.*

A noteworthy prior work that we compare Theorem 1 to is OGM [26,40], which achieves a factor-2 speedup over the classical Nesterov's accelerated gradient method. Let $N > 0$ be the total iteration count and $L > 0$. The **O**ptimized **G**radient **M**ethod is

$$\begin{aligned} y_{i+1} &= x_i - \frac{1}{L}\nabla f(x_i) \\ x_{i+1} &= y_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}(y_{i+1} - y_i) + \frac{\theta_i}{\theta_{i+1}}(y_{i+1} - x_i) \end{aligned} \tag{OGM}$$

with starting point $x_0 = y_0$, f is L-smooth and convex, and with the sequence θ_i satisfying

$$\theta_i = \begin{cases} 1 & \text{if } i = 0, \\ \frac{1+\sqrt{1+4\theta_{i-1}^2}}{2} & \text{if } 1 \leq i \leq N-1, \\ \frac{1+\sqrt{1+8\theta_{N-1}^2}}{2} & \text{if } i = N. \end{cases}$$

Its convergence rate is

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

The θ-coefficients of OptISTA are identical to the θ-coefficients of OGM [26,40], which are in turn equal to the standard θ-coefficients of Nesterov's FGM [57] for $i = 0, 1, \ldots, N-1$, but the last coefficient θ_N differs. (Roughly, $\theta_i \sim i/2$ for $i < N$ and $\theta_N \sim N/\sqrt{2}$ [59, Theorem 7].)

Interestingly, OptISTA reduces to OGM when $h = 0$. To see this, note that $\mathbf{prox}_{\alpha h \equiv 0}(\cdot)$ is an identity operator for any scalar $\alpha > 0$. Hence, the iterates of OptISTA reduces to

$$\begin{aligned} y_{i+1} &= y_i - \frac{\gamma_i}{L}\nabla f(x_i) \\ z_{i+1} &= x_i + \frac{1}{\gamma_i}(y_{i+1} - y_i) = x_i - \frac{1}{L}\nabla f(x_i) \\ x_{i+1} &= z_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}(z_{i+1} - z_i) + \frac{\theta_i}{\theta_{i+1}}(z_{i+1} - x_i), \end{aligned}$$

where z-iterates are now identical to the y-iterates of OGM.

The new rate of Theorem 1 improves the prior rates of FISTA $L\|x_0 - x_\star\|^2/(2\theta_{N-1}^2) \sim 2L\|x_0 - x_\star\|^2/N^2$ [6] and FPGM2 $2L\|x_0 - x_\star\|^2/(N^2 + 7N)$ [66] by a factor of 2 in the leading $\mathcal{O}(1/N^2)$-term. We note that a similar factor-two improvement appears in OGM as well.

In terms of solving ($\mathcal{O}^{\text{outer}}$), given the information that N is the total iteration count, we believe that not performing a gradient evaluation in the last step, and the absence of constraints for the next step, allows for finding a slightly more optimal step size. However, even for the authors, it is difficult to pinpoint the exact reason why this improvement specifically amounts to a factor of two. Furthermore, the rate of Theorem 1 exactly matches the lower bound we later present in Theorem 2 of Section 4 and, therefore, provably cannot be improved without further assumptions. We say OptISTA is exactly optimal in this sense.

The rate of Nesterov's FGM [57] is identical to that of FISTA. This rate was improved by a factor of 2 with OGM, which has the rate $L\|x_0 - x_\star\|^2/(2\theta_N^2) \sim L\|x_0 - x_\star\|^2/N^2$ [26,40]. The rate of OGM is slightly better than that of OptISTA by a factor of $\theta_N^2/(\theta_N^2 - 1)$, but this difference does not manifest in the leading $\mathcal{O}(1/N^2)$-term.

Also, it is worth noting that there is a difference of 2 in the denominators of the worst-case rate of OGM ($2\theta_N^2$) and OptISTA ($2\theta_N^2 - 2$) is reminiscent of some previous results. In prior work, the denominators of the convergence rates of gradient descent on a single function ($4N + 2$) [26] and that of projected/proximal gradient on composite function ($4N$) [71] also had a difference of 2.

2.2 OptISTA proof outline

The proof of Theorem 1 utilizes the Lyapunov analysis, which is commonly used in analyzing the convergence of first-order methods [70,59,45,44]. The main structure of the proof is usually twofold. First, define an equivalent form of the given method. Usually, this equivalent form introduces an "auxiliary sequence" that may not be present in the original form. Then, by constructing a non-increasing Lyapunov sequence using the auxiliary sequence, we conclude the convergence result. For example, the $\mathcal{O}(1/N^2)$ rate of Nesterov's method [57] can be proved by defining

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

where $x_{k-1}^+ = x_{k-1} - \frac{1}{L}\nabla f(x_{k-1})$ with $z_0 = x_0$ and $z_{k+1} = z_k - \frac{k+1}{2L}\nabla f(x_k)$, and showing $U_k \le \cdots \le U_0$.

However, the most challenging part of this style of proof is identifying the auxiliary sequence and the nonincreasing Lyapunov sequence. This process is far from straightforward, and in the case of composite functions, as in our paper, it is inevitably more complex and lengthy than for single function setups. In this section, we will introduce only the key ideas, leaving the detailed calculations to the appendix.

Define the following sequences, which will be soon shown as an alternative representation of (OptISTA) in Lemma 1:

$$\begin{aligned}
y_{i+1} &= \mathbf{prox}_{\frac{\gamma_i}{L}h}\big(y_i - \frac{\gamma_i}{L}\nabla f(x_i)\big)\\
z_{i+1} &= x_i + \frac{1}{\gamma_i}\left(y_{i+1} - y_i\right)\\
w_{i+1} &= w_i - \frac{2\theta_i}{L}\nabla f(x_i) - \frac{2\theta_i}{L}h'(y_{i+1})\\
x_{i+1} &= \left(1 - \frac{1}{\theta_{i+1}}\right)z_{i+1} + \frac{1}{\theta_{i+1}}w_{i+1}
\end{aligned} \tag{OptISTA-A}$$

for $i = 1, \dots, N-1$, where $w_0 = x_0$ and $h'(y_{i+1}) = \frac{L}{\gamma_i}\left(y_i - \frac{\gamma_i}{L}\nabla f(x_i) - y_{i+1}\right) \in \partial h(y_{i+1})$. Notably, OptISTA-A has the auxilary sequence $\{w_i\}_{i\in[0:N]}$ that was not explicitly present in OptISTA. Also we define the sequence $\{\tilde{\theta}_i\}_{i=0,\dots,N-1}$, which is similar to θ–sequence with slight modification at the last step, defined as

$$\tilde{\theta}_i = \begin{cases} \theta_i & \text{if } i \in [0 : N-2],\\ \frac{2\theta_{N-1}+\theta_N - 1}{2}, & \text{if } i = N-1. \end{cases}$$

Then we have the following two lemmas.

Lemma 1 *Denote x and y sequences generated by (OptISTA) and (OptIST-A) by $\{x_k, y_k\}_{k=0}^N$ and $\{\hat{x}_k, \hat{y}_k\}_{k=0}^N$, respectively, then*

$$\hat{x}_k = x_k, \quad \hat{y}_k = y_k$$

for every $k = 0, \dots, N$. Hence, (OptISTA-A) *is indeed equivalent to* (OptISTA).

Proof Note that $y_0 = x_0 = \hat{y}_0 = \hat{x}_0$ by definition. Assume $x_i = \hat{x}_i$ and $y_i = \hat{y}_i$ for $i \in [0:k]$. Then, $\hat{y}_{k+1} = y_{k+1}$ is immediate. To prove $\hat{x}_{k+1} = x_{k+1}$, observe that

$$\begin{aligned}
\hat{x}_{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) z_{k+1} + \frac{1}{\theta_{k+1}} w_{k+1} \\
&\overset{(\star)}{=} \left(1 - \frac{1}{\theta_{k+1}}\right) z_{k+1} + \frac{1}{\theta_{k+1}} \left(\theta_k x_k - (\theta_k - 1) z_k - \frac{2\theta_k}{L} \nabla f(x_k) - \frac{2\theta_k}{L} h'(y_{k+1})\right) \\
&= \left(1 - \frac{1}{\theta_{k+1}}\right) z_{k+1} - \frac{\theta_k - 1}{\theta_{k+1}} z_k + \frac{\theta_k}{\theta_{k+1}} \left(x_k - \frac{2}{L} \nabla f(x_k) - \frac{2}{L} h'(y_{k+1})\right) \\
&= \left(\frac{\theta_k - 1}{\theta_{k+1}}\right) (z_{k+1} - z_k) + z_{k+1} - \frac{\theta_k}{\theta_{k+1}} z_{k+1} + \frac{\theta_k}{\theta_{k+1}} \left(x_k - \frac{2}{L} \nabla f(x_k) - \frac{2}{L} h'(y_{k+1})\right) \\
&= z_{k+1} + \left(\frac{\theta_k - 1}{\theta_{k+1}}\right) (z_{k+1} - z_k) + \frac{\theta_k}{\theta_{k+1}} \left(x_k - z_{k+1} - \frac{2}{L} \nabla f(x_k) - \frac{2}{L} h'(y_{k+1})\right) \\
&\overset{(\circ)}{=} z_{k+1} + \left(\frac{\theta_k - 1}{\theta_{k+1}}\right) (z_{k+1} - z_k) + \frac{\theta_k}{\theta_{k+1}} \left(x_k - z_{k+1} + \frac{2}{\gamma_k} (y_{k+1} - y_k)\right) \\
&= z_{k+1} + \left(\frac{\theta_k - 1}{\theta_{k+1}}\right) (z_{k+1} - z_k) + \frac{\theta_k}{\theta_{k+1}} (x_k - z_{k+1} + 2 (z_{k+1} - x_k)) \\
&= z_{k+1} + \left(\frac{\theta_k - 1}{\theta_{k+1}}\right) (z_{k+1} - z_k) + \frac{\theta_k}{\theta_{k+1}} (z_{k+1} - x_k) = x_{k+1}
\end{aligned}$$

For $(\star)$, we used the fact that

$$w_{k+1} = w_k - \frac{2\theta_k}{L} \nabla f(x_k) - \frac{2\theta_k}{L} h'(y_{k+1})$$

and

$$x_k = \left(1 - \frac{1}{\theta_k}\right) z_k + \frac{1}{\theta_k} w_k \iff w_k = \theta_k x_k - (\theta_k - 1) z_k.$$

For $(\circ)$, we used that

$$y_{i+1} - y_i = -\frac{\gamma_k}{L} \nabla f(x_k) - \frac{\gamma_k}{L} h'(y_{k+1})$$

by the definition of $h'(y_{k+1})$ and that

$$z_{k+1} - x_i = \frac{1}{\gamma_k} (y_{k+1} - y_k) = -\frac{1}{L} \nabla f(x_k) - \frac{1}{L} h'(y_{k+1}).$$

Therefore,

$$x_k - z_{k+1} - \frac{2}{L} \nabla f(x_k) - \frac{2}{L} h'(y_{k+1}) = x_k - z_{k+1} + 2(z_{k+1} - x_i) = 2z_{k+1} - x_k - z_{k+1}.$$

□

Lemma 2 *In* (OptISTA)*, we have* $x_N = y_N$.

Proof The proof is quite techniqcal and therefore we defer its proof to Appendix B. □

Now we define the Lyapunov sequence $\{\mathcal{U}_k\}_{k\in[-1:N]}$. Explicit form of the sequence is quite cumbersome, therefore we introduce $k=-1,N$ cases only. We defer the details to the Appendix C.

$$\begin{aligned}\mathcal{U}_N &= f(x_N)-f(x_\star)+h(y_N)-h(x_\star)\\ &+\frac{L}{2\theta_N^2}\left\|w_N-x_\star+\frac{1}{L}\nabla f(x_\star)+\frac{2\theta_{N-1}}{L}h'(y_N)-\frac{\theta_N}{L}\nabla f(x_N)-\frac{2\tilde{\theta}_{N-1}}{L}h'(y_N)\right\|^2\\ &+\frac{L}{2\theta_N^2(\theta_N^2-1)}\left\|x_0-x_\star-\frac{\theta_N^2-1}{L}\nabla f(x_\star)-\sum_{i=0}^{N-1}\frac{2\tilde{\theta}_i}{L}h'(y_{i+1})\right\|^2\\ &+\sum_{i\neq j,i,j\in[1:N]}\frac{\tilde{\theta}_{i-1}\tilde{\theta}_{j-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)-h'(y_j)\|^2+\sum_{i=1}^{N-1}\frac{\tilde{\theta}_{i-1}^2}{L\theta_N^2}\|h'(y_i)-h'(y_{i+1})\|^2,\end{aligned}$$

$$\mathcal{U}_{-1}=\frac{L\|x_0-x_\star\|^2}{2(\theta_N^2-1)}.$$

Then, we show $\mathcal{U}_N\leq\mathcal{U}_{N-1}\leq\cdots\leq\mathcal{U}_1\leq\mathcal{U}_0\leq\mathcal{U}_{-1}$ to get

$$f(x_N)-f(x_\star)+h(y_N)-h(x_\star)\leq\mathcal{U}_N\leq\cdots\leq\mathcal{U}_{-1}=\frac{L\|x_0-x_\star\|^2}{2(\theta_N^2-1)}.$$

Finally, we use the fact $x_N=y_N$ by Lemma 2 to conclude that

$$f(y_N)+h(y_N)-f(x_\star)-h(x_\star)\leq\frac{L\|x_0-x_\star\|^2}{2(\theta_N^2-1)}.$$

The main challenge here is to show that $\{\mathcal{U}_k\}_{k\in[-1:N]}$ is nonincreasing, and the rigorous proof is provided in Appendix C.

3 Computer-assisted algorithmic design via PEP

In this section, we present the double-function stepsize-optimization PEP methodology that we used to design OptISTA.

3.1 Double-function fixed-step first-order methods

First, we concretely specify and parameterize the class of first-order methods for solving $(\mathcal{P})$: we consider *N-step double-function fixed-step first-order methods* (N-DF-FSFOM) defined as

$$\begin{aligned}y_{i+1}&=x_0-\sum_{j\in[0:i]}\frac{\phi_{i+1,j}}{L}\nabla f(x_j)-\sum_{j\in[0:i]}\frac{\psi_{i+1,j}}{L}h'(y_{j+1})\\ x_{i+1}&=x_0-\sum_{j\in[0:i]}\frac{\alpha_{i+1,j}}{L}\nabla f(x_j)-\sum_{j\in[0:i]}\frac{\beta_{i+1,j}}{L}h'(y_{j+1})\end{aligned}\qquad(N\text{-DF-FSFOM})$$

for $i\in[0:N-1]$, where $h'(y_{i+1})\in\partial h(y_{i+1})$ is a subgradient h at y_{i+1} defined by

$$\begin{aligned}\tilde{y}_{i+1}&=x_0-\sum_{j\in[0:i]}\frac{\phi_{i+1,j}}{L}f'(x_j)-\sum_{j\in[0:i-1]}\frac{\psi_{i+1,j}}{L}h'(y_{j+1})\\ y_{i+1}&=\mathbf{prox}_{\frac{\psi_{i+1,i}}{L}h}(\tilde{y}_{i+1})\\ h'(y_{i+1})&=\frac{L}{\psi_{i+1,j}}(\tilde{y}_{i+1}-y_{i+1})\end{aligned}$$

for $i \in [0 : N-1]$. The values of the stepsizes $\{\phi_{i,j}\}$, $\{\psi_{i,j}\}$, $\{\alpha_{i,j}\}$, and $\{\beta_{i,j}\}$, where i, j have the range $1 \leq i \leq N$ and $0 \leq j < i$, determine the specific instance of (N-DF-FSFOM). We consider x_N to be the output of (N-DF-FSFOM).

The class of (N-DF-FSFOM) includes the proximal point method [50,63], accelerated proximal point method [35], Nesterov's FGM [57], OGM [26,40], ISTA [17,36,72,28], FISTA [6], and other variants for the composite setup [66,68,42]. The class essentially includes all conceivable first-order methods that (i) use N evaluations of ∇f, (ii) N evaluations of $\mathbf{prox}_{\gamma h}$ (with the value of $\gamma > 0$ chosen freely every use), (iii) interleave the evaluations of ∇f and $\mathbf{prox}_{\gamma h}$ in any arbitrary order, and (iv) do not utilize function values or linesearches.

Discussion on stepsize dependence. For problem ($\mathcal{P}$), let L be the smoothness coefficient of f and $R \geq \|x_0 - x_\star\|$ be an upper bound on the distance to solution. Let N be total iteration count of the (N-DF-FSFOM). The stepsizes $\{\phi_{i,j}\}$, $\{\psi_{i,j}\}$, $\{\alpha_{i,j}\}$, and $\{\beta_{i,j}\}$ may depend on L, R, and N but are otherwise predetermined.

Methods like gradient descent, Nesterov FGM, and FISTA utilize L in their stepsizes. In fact, most smooth minimization methods without linesearch do so (with the notable recent exception of [49]). Subgradient methods often utilize R in their stepsizes [58, Section 3.2.3]. Stochastic first-order methods often utilize N in their stepsizes [32, Corollary 5.6] and recent methods such as OGM-G [43] and FISTA-G [45] also do so. Our exact optimal method OptISTA uses N and L, but not R in its stepsizes.

3.2 Double-function stepsize-optimization PEP (DF-SO-PEP)

Next, we introduce the double-function stepsize-optimization performance estimation problem (DF-SO-PEP), a computer-assisted methodology for finding the optimal (N-DF-FSFOM).

For the sake of convenience, assume $L = 1$ to eliminate the dependence on L. Note that we can easily extend to an arbitrary smoothness parameter $L > 0$ by mere scaling. Write $\mathcal{M}_N$ to denote the set of all (N-DF-FSFOM). Define the worst-case performance (risk) of $M \in \mathcal{M}_N$ as

$$\mathcal{R}(M) = \left(\begin{array}{ll} \text{maximize} & f(x_N) + h(x_N) - f(x_\star) - h(x_\star) \\ \text{subject to} & f \text{ is } L\text{-smooth convex, } h \text{ is closed convex proper} \\ & \text{there is an } x_\star \text{ minimizing } f + h \text{ such that } \|x_0 - x_\star\| \leq R \\ & \{(x_i, y_i)\}_{i \in [1:N]} \text{ generated by } M \text{ with initial point } x_0 = y_0 \end{array} \right), \qquad (\mathcal{O}^{\text{inner}})$$

where x_0, f, and h are the decision variables of the maximization problem. (f and h are infinite-dimensional variables.) The optimal method $M_N^\star \in \mathcal{M}_N$ is a solution to the following minimax optimization problem

$$\mathcal{R}^\star(\mathcal{M}_N) = \underset{M \in \mathcal{M}_N}{\text{minimize}}\ \mathcal{R}(M). \qquad (\mathcal{O}^{\text{outer}})$$

Later in this section, we demonstrate that ($\mathcal{O}^{\text{outer}}$) can be expressed as a finite-dimensional, nonconvex yet practically solvable quadratically constrained quadratic program (QCQP), which can be globally optimized using spatial branch-and-bound optimization solvers [1]. In essence, the conversion to QCQP consists of two primary steps: (1) formulating the inner maximization problem in ($\mathcal{O}^{\text{inner}}$) as a convex semidefinite programming minimization problem (SDP), and (2) transforming ($\mathcal{O}^{\text{outer}}$) into a QCQP using the minimization SDP from the first step. The dual variables of the aforementioned minimization SDP are referred to as the *inner-dual variables*. In the resulting QCQP form of ($\mathcal{O}^{\text{outer}}$), the decision variables consist of the stepsizes from (N-DF-FSFOM) and the inner-dual variables.

Comparison with prior approaches. Our work is the first instance of a stepsize-optimization performance estimation problem (SO-PEP) framework optimizing over methods in a double-function setup. However, there have been prior work [68,64] using the PEP to evaluate the performance of a given fixed method with two or more functions. The stepsize-optimization introduces nonconvexity and makes the resulting finite-dimensional optimization problem a nonconvex QCQP rather than a convex

SDP as in prior SO-PEP work [26,40]. We overcome such nonconvexity using spatial branch-and-bound algorithms [1,48,38,16].

3.3 Nonconvex QCQP formulation for $(\mathcal{O}^{\text{outer}})$

As mentioned in the previous section, we transform $(\mathcal{O}^{\text{outer}})$ into a (nonconvex) QCQP through two distinct steps. First, we reformulate the inner problem $(\mathcal{O}^{\text{inner}})$ as a convex SDP. This initial step adopts a similar approach as described in [68]. Second, we express the outer problem $(\mathcal{O}^{\text{outer}})$ as a QCQP. This second step is conceptually aligned with the method presented in [16], which was developed to design optimal first-order methods for minimizing a single function.

We use the following notations. Write $e_i \in \mathbb{R}^d$ for the standard i-th unit vector for $i \in [0:d-1]$. Write $\mathbb{R}^{m\times n}$ for the set of $m \times n$ matrices, $\mathbb{S}^n$ for the set of $n \times n$ symmetric matrices, and $\mathbb{S}^n_+$ for the set of $n \times n$ positive-semidefinite matrices. Write $(\cdot \odot \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ to denote the symmetric outer product, that is, for any $x, y \in \mathbb{R}^d$: $x \odot y = {}^{(xy^\intercal + yx^\intercal)}/_2$.

3.3.1 Formulating the inner problem as a convex SDP

Let $d > 0$, $L > 0$, and $R > 0$. Let $\mathcal{F}_{0,L}$ denote the set of L-smooth convex functions on $\mathbb{R}^d$ and $\mathcal{F}_{0,\infty}$ the set of closed convex proper functions on $\mathbb{R}^d$.

Infinite-dimensional inner optimization problem. Write $(\mathcal{O}^{\text{inner}})$ as

$$\mathcal{R}(M) = \left(\begin{array}{l} \text{maximize} \quad f(x_N) + h(x_N) - f(x_\star) - h(x_\star) \\ \text{subject to} \\ f \in \mathcal{F}_{0,L},\ h \in \mathcal{F}_{0,\infty}, \quad \triangleright \text{ function class} \\ \left.\begin{array}{l} y_{i+1} = x_0 - \displaystyle\sum_{j\in[0,i]} \frac{\phi_{i+1,j}}{L}\nabla f(x_j) - \sum_{j\in[0,i]} \frac{\psi_{i+1,j}}{L} h'(y_{j+1}), \quad i \in [0:N-1], \\ x_{i+1} = x_0 - \displaystyle\sum_{j\in[0,i]} \frac{\alpha_{i+1,j}}{L}\nabla f(x_j) - \sum_{j\in[0,i]} \frac{\beta_{i+1,j}}{L} h'(y_{j+1}), \quad i \in [0:N-1], \end{array}\right\} \quad \triangleright \text{ method } M \\ \nabla f(x_\star) + h'(x_\star) = 0,\ x_\star = 0, \quad \triangleright \text{ optimal solution} \\ \|x_0 - x_\star\|^2 \le R^2, \quad \triangleright \text{ initial condition} \end{array}\right)$$

where f, h, $x_0, \dots, x_N$, and $y_1, \dots, y_N$ are the decision variables. Furthermore, we set $x_\star = 0$ without loss of generality. As is, both f and h are infinite-dimensional decision variables.

Interpolation argument. We now convert the infinite-dimensional optimization problem into a finite-dimensional one with the following interpolation result.

Lemma 3 ($\mathcal{F}_{0,L}$- and $\mathcal{F}_{0,\infty}$-interpolation [69, Theorem 4]) *Let I be an index set, and let $\{(x_i, g_i, f_i)\}_{i\in I} \subseteq \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$. Let $0 < L \le \infty$. There exists $f \in \mathcal{F}_{0,L}$ satisfying $f(x_i) = f_i$ and $g_i \in \partial f(x_i)$ for all $i \in I$ if and only if* [1]

$$f_i \ge f_j + \langle g_j,\, x_i - x_j\rangle + \frac{1}{2L}\|g_i - g_j\|^2, \quad \forall\, i, j \in I.$$

When $L = \infty$, we mean $\frac{1}{2L}\|g_i - g_j\|^2 = 0$.

[1] This can be viewed as a discretization of the following condition [58, Theorem 2.1.5, Equation (2.1.10)]: $f \in \mathcal{F}_{0,L}$ if and only if

$$f(y) \ge f(x) + \langle \nabla f(x);\, y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

For notation convenience define

$$\begin{aligned}
&\text{index } \star \text{ is denoted by } -1,\\
&w_{-1} = x_\star = 0,\ w_i = x_i \text{ for } i \in [0:N],\ w_{N+i} = y_i \text{ for } i \in [1:N],\\
&I_f = [-1:N],\\
&I_h = [N:2N] \cup \{-1\},\\
&f(w_i) = f_i,\ \nabla f(w_i) = f_i' \text{ for } i \in I_f,\\
&h(w_i) = h_i,\ h'(w_i) = h_i',\ \text{for } i \in I_h.
\end{aligned}$$

Finite-dimensional maximization problem. Using Lemma 3 and the new notation above, reformulate ($\mathcal{O}^{\text{inner}}$) as

$$\mathcal{R}(M) = \left(\begin{array}{l}
\text{maximize} \quad f_N + h_N - f_{-1} - h_{-1}\\
\text{subject to}\\
f_i \ge f_j + \langle f_j', w_i - w_j\rangle + \frac{1}{2L}\|f_i' - f_j'\|^2, \quad i,j \in I_f, i \ne j,\\
h_i \ge h_j + \langle h_j', w_i - w_j\rangle, \quad i,j \in I_h, i \ne j,\\
w_{N+i+1} = w_0 - \sum_{k=0}^{i} \frac{\phi_{i+1,k}}{L} f_k' - \sum_{k=0}^{i} \frac{\psi_{i+1,k}}{L} h_{N+k+1}', \quad i \in [0:N-1],\\
w_{i+1} = w_0 - \sum_{k=0}^{i} \frac{\alpha_{i+1,k}}{L} f_k' - \sum_{k=0}^{i} \frac{\beta_{i+1,k}}{L} h_{N+k+1}', \quad i \in [0:N-1],\\
f_{-1}' + h_{-1}' = 0,\\
w_{-1} = 0,\\
\|w_0 - w_{-1}\|^2 \le R^2,
\end{array}\right)$$

where the decision variables are $\{w_i, f_i', f_i\}_{i\in I_f} \subseteq \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ and $\{w_i, h_i', h_i\}_{i\in I_h} \subseteq \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$. The optimization problem is now finite-dimensional, although nonconvex.

Grammian formulation. Next, we formulate ($\mathcal{O}^{\text{inner}}$) into a (finite-dimensional convex) SDP. Let

$$\begin{aligned}
H &= [\ w_0 \mid f_{-1}' \mid f_0' \mid \cdots \mid f_N' \mid h_N' \mid h_{N+1}' \mid \cdots \mid h_{2N}'\ ] \in \mathbb{R}^{d\times(2N+4)},\\
G &= H^\intercal H \in \mathbb{S}_+^{2N+4},\\
F &= [\ f_{-1} \mid f_0 \mid \cdots \mid f_N \mid h_{-1} \mid h_N \mid h_{N+1} \mid \cdots \mid h_{2N}\ ] \in \mathbb{R}^{1\times(2N+4)}.
\end{aligned}$$

Note that $\mathbf{rank}\, G \le d$. Define the following notation for selecting columns and elements of H and F:

$$\begin{aligned}
&\mathbf{w}_{-1} = \mathbf{0} \in \mathbb{R}^{2N+4},\ \mathbf{w}_0 = e_0 \in \mathbb{R}^{2N+4},\\
&\mathbf{f}_i' = e_{i+2} \in \mathbb{R}^{2N+4} \text{ for } i \in [-1:N],\\
&\mathbf{h}_{-1}' = -e_1 \in \mathbb{R}^{2N+4},\ \mathbf{h}_{N+i}' = e_{N+i+3} \in \mathbb{R}^{2N+4} \text{ for } i \in [0:N],\\
&\mathbf{w}_{N+i+1} = \mathbf{w}_0 - \sum_{k=0}^{i} \frac{\phi_{i+1,k}}{L} \mathbf{f}_k' - \sum_{k=0}^{i} \frac{\psi_{i+1,k}}{L} \mathbf{h}_{N+k+1}' \in \mathbb{R}^{2N+4} \text{ for } i \in [0:N-1],\\
&\mathbf{w}_{i+1} = \mathbf{w}_0 - \sum_{k=0}^{i} \frac{\alpha_{i+1,k}}{L} \mathbf{f}_k' - \sum_{k=0}^{i} \frac{\beta_{i+1,k}}{L} \mathbf{h}_{N+k+1}' \in \mathbb{R}^{2N+4} \in \text{ for } i \in [0:N-1],\\
&\mathbf{f}_i = e_{i+1} \in \mathbb{R}^{2N+4} \text{ for } i \in [-1,N],\\
&\mathbf{h}_{-1} = e_{N+2} \in \mathbb{R}^{2N+4},\ \mathbf{h}_{N+i} = e_{N+i+3} \in \mathbb{R}^{2N+4} \text{ for } i \in [0:N].
\end{aligned}$$

Note that $\mathbf{w}_i$ depends linearly on $\{\phi_{i,j}\}$, $\{\psi_{i,j}\}$, $\{\alpha_{i,j}\}$, and $\{\beta_{i,j}\}$. This notation is defined so that for all $i \in [-1:2N]$ we have

$$\begin{aligned}
w_i &= H\mathbf{w}_i,\\
f_i' &= H\mathbf{f}_i',\ h_i' = H\mathbf{h}_i',\\
f_i &= F\mathbf{f}_i,\ h_i = F\mathbf{h}_i.
\end{aligned}$$

Also, our choice ensures that $f'_{-1} + h'_{-1} = H\left(\mathbf{f}'_{-1} + \mathbf{h}'_{-1}\right) = H\left(e_1 - e_1\right) = 0$. Furthermore, for i, j that belong to either I_f or I_h, define:

$$\begin{aligned}
&A^f_{i,j}(\phi,\psi,\alpha,\beta) = \mathbf{f}'_j \odot (\mathbf{w}_i - \mathbf{w}_j),\\
&A^h_{i,j}(\phi,\psi,\alpha,\beta) = \mathbf{h}'_j \odot (\mathbf{w}_i - \mathbf{w}_j),\\
&B_{i,j}(\phi,\psi,\alpha,\beta) = (\mathbf{w}_i - \mathbf{w}_j) \odot (\mathbf{w}_i - \mathbf{w}_j),\\
&C^f_{i,j} = (\mathbf{f}'_i - \mathbf{f}'_j) \odot (\mathbf{f}'_i - \mathbf{f}'_j),\\
&a^f_{i,j} = \mathbf{f}_j - \mathbf{f}_i,\ a^h_{i,j} = \mathbf{h}_j - \mathbf{h}_i.
\end{aligned}$$

Note that $A^f_{i,j}(\phi,\psi,\alpha,\beta)$ and $A^h_{i,j}(\phi,\psi,\alpha,\beta)$ are affine and $B_{i,j}(\phi,\psi,\alpha,\beta)$ is quadratic in the entries of $\{\phi_{i,j}\}$, $\{\psi_{i,j}\}$, $\{\alpha_{i,j}\}$, and $\{\beta_{i,j}\}$. This notation is defined so that

$$\begin{aligned}
&\left\langle f'_j,\, w_i - w_j\right\rangle = \mathbf{tr}\, G A^f_{i,j}(\phi,\psi,\alpha,\beta),\ \left\langle h'_j,\, w_i - w_j\right\rangle = \mathbf{tr}\, G A^h_{i,j}(\phi,\psi,\alpha,\beta),\\
&\|w_i - w_j\|^2 = \mathbf{tr}\, G B_{i,j}(\phi,\psi,\alpha,\beta),\\
&\|f'_i - f'_j\|^2 = \mathbf{tr}\, G C^f_{i,j},\\
&f_j - f_i = F a^f_{i,j},\ h_j - h_i = F a^h_{i,j}.
\end{aligned}$$

for all i, j that belong to either I_f or I_h. Using this notation, formulate $(\mathcal{O}^{\text{inner}})$ as

$$\mathcal{R}(M) = \begin{pmatrix}
\text{maximize} \quad F a^f_{-1,N} + F a^h_{-1,N}\\
\text{subject to}\\
F a^f_{i,j} + \mathbf{tr}\, A^f_{i,j}(\phi,\psi,\alpha,\beta) G + \frac{1}{2L}\mathbf{tr}\, C^f_{i,j} G \le 0,\ i,j \in I_f, i \ne j,\\
F a^h_{i,j} + \mathbf{tr}\, A^h_{i,j}(\phi,\psi,\alpha,\beta) G \le 0,\ i,j \in I_g, i \ne j,\\
G \succeq 0,\ \mathbf{rank}(G) \le d,\\
\mathbf{tr}\, B_{0,-1} G \le R^2,
\end{pmatrix}$$

where $G \in \mathbb{R}^{(2N+4)\times(2N+4)}$ and $F \in \mathbb{R}^{1\times(2N+4)}$ are the decision variables. The equivalence relies on the fact that given a $G \in \mathbb{S}^{2N+4}_+$ satisfying $\mathbf{rank}(G) \le d$, there exists a $H \in \mathbb{R}^{d\times(2N+4)}$ such that $G = H^\intercal H$. The argument is further detailed in [69, §3.2]. This formulation is not yet a convex SDP due to the rank constraint $\mathbf{rank}(G) \le d$.

SDP representation. Next, we make the following large-scale assumption.

Assumption 1 *We have $d \ge 2N + 4$.*

Under this assumption, the constraint $\mathbf{rank}\, G \le d$ becomes vacuous, since $G \in \mathbb{S}^{2N+4}_+$. We drop the rank constraint and formulate $(\mathcal{O}^{\text{inner}})$ as a convex SDP

$$\mathcal{R}(M) = \begin{pmatrix}
\text{maximize} \quad F a^f_{-1,N} + F a^h_{-1,N}\\
\text{subject to}\\
F a^f_{i,j} + \mathbf{tr}\, A^f_{i,j}(\phi,\psi,\alpha,\beta) G + \frac{1}{2L}\mathbf{tr}\, C^f_{i,j} G \le 0,\ i,j \in I_f, i \ne j, \quad \triangleright \text{ dual var. } \lambda_{i,j} \ge 0\\
F a^h_{i,j} + \mathbf{tr}\, A^h_{i,j}(\phi,\psi,\alpha,\beta) G \le 0,\ i,j \in I_g, i \ne j, \quad \triangleright \text{ dual var. } \tau_{i,j} \ge 0\\
-G \preceq 0, \quad \triangleright \text{ dual var. } Z \succeq 0\\
\mathbf{tr}\, B_{0,-1} G - R^2 \le 0, \quad \triangleright \text{ dual var. } \nu \ge 0
\end{pmatrix} \tag{1}$$

where $F \in \mathbb{R}^{1\times(2N+4)}$ and $G \in \mathbb{R}^{(2N+4)\times(2N+4)}$ are the decision variables. We represent the dual variables corresponding to the right-hand side constraints using the notation $\triangleright$ dual var., which will be employed in later sections. It is important to note that omitting the rank constraint does not constitute a relaxation; the optimization problem (1) and its solution remain independent of dimension d, as long as the large-scale assumption $d \ge 2N + 4$ is satisfied. Further discussion on this topic can be found in [69, §3.3].

Dualization. We now utilize convex duality to recast ($\mathcal{O}^{\text{inner}}$), initially a maximization problem, as a minimization problem. Taking the dual of (1) yields

$$\overline{\mathcal{R}}(M) = \left(\begin{array}{l} \text{minimize} \quad \nu R^2 \\ \text{subject to} \\ -a^f_{-1,N} - a^h_{-1,N} + \sum_{i,j\in I_f, i\neq j} \lambda_{i,j} a^f_{i,j} + \sum_{i,j\in I_g, i\neq j} \tau_{i,j} a^h_{i,j} = 0, \\ \nu B_{0,-1} + \sum_{i,j\in I_f, i\neq j} \lambda_{i,j} \left(A^f_{i,j}(\phi,\psi,\alpha,\beta) + \frac{1}{2L} C^f_{i,j} \right) \\ \quad + \sum_{i,j\in I_g, i\neq j} \tau_{i,j} A^h_{i,j}(\phi,\psi,\alpha,\beta) = Z, \\ Z \succeq 0, \\ \nu \geq 0, \\ \lambda_{i,j} \geq 0, \quad i,j \in I_f, i \neq j, \\ \tau_{i,j} \geq 0, \quad i,j \in I_h, i \neq j, \end{array}\right) \tag{2}$$

where $\nu \in \mathbb{R}$, $\lambda = \{\lambda_{i,j}\}_{i,j\in I_f : i\neq j}$, $\tau = \{\tau_{i,j}\}_{i,j\in I_h, i\neq j}$ with $\lambda_{i,j} \in \mathbb{R}$, $\tau_{i,j} \in \mathbb{R}$, and $Z \in \mathbb{S}^{2N+4}_+$ are the decision variables. (Note, we write $\overline{\mathcal{R}}$ rather than $\mathcal{R}$ here.) We call ν, λ, τ and Z the *inner-dual variables*. By weak duality of convex SDPs, we have

$$\mathcal{R}(M) \leq \overline{\mathcal{R}}(M).$$

In convex SDPs, strong duality holds often but not always. For the sake of simplicity, we assume strong duality holds. Note that because we establish a matching lower bound, we know, after the fact, that strong duality holds.

Assumption 2 *Strong duality holds between* (1) *and* (2)*, i.e.,*

$$\mathcal{R}(M) = \overline{\mathcal{R}}(M).$$

Again, we do not need this assumption in our specific setup as we establish a matching lower bound. However, in a generic variant of our setup, Assumption 2 can be eliminated by employing the reasoning outlined in [60, Claim 4]; however, the approach is somewhat complicated.

At this juncture, our formulation diverges from the previous steps of [68].

3.3.2 Formulating the outer problem ($\mathcal{O}^{\text{outer}}$) as a QCQP

With the inner problem ($\mathcal{O}^{\text{inner}}$) formulated as a minimization problem, the outer optimization problem ($\mathcal{O}^{\text{outer}}$) now involves joint minimization over both the inner dual variables and the stepsizes. Although the inner problem is convex, the outer minimization problem is not convex in all variables. The nonconvex outer optimization problem requires minimizing over the stepsizes ϕ, ψ, α, and β and the inner dual variables of (2). We directly address this nonconvexity by formulating ($\mathcal{O}^{\text{outer}}$) as a (nonconvex) QCQP and solving it exactly using spatial branch-and-bound algorithms. To achieve this, we substitute the semidefinite constraint with a quadratic constraint using the Cholesky factorization.

Lemma 4 ([37, Corollary 7.2.9]) *A matrix $Z \in \mathbb{S}^n$ is positive semidefinite if and only if it has a Cholesky factorization $PP^\intercal = Z$, where $P \in \mathbb{R}^{n\times n}$ is lower triangular with nonnegative diagonals.*

Using Lemma 4, we have

$$\begin{aligned} (Z \succeq 0) \quad &\Leftrightarrow \quad \left(\begin{array}{l} P \text{ is lower triangular with nonnegative diagonals,} \\ PP^\intercal = Z. \end{array}\right) \\ &\Leftrightarrow \quad \left(\begin{array}{l} P_{j,j} \geq 0, \quad j \in [1:2N+4], \\ P_{i,j} = 0, \quad 1 \leq i < j \leq 2N+4, \\ \sum_{k=1}^{j} P_{i,k} P_{j,k} = Z_{i,j}, \quad 1 \leq j \leq i \leq 2N+4. \end{array}\right) \end{aligned}$$

We now formulate ($\mathcal{O}^{\text{outer}}$), the problem to find an optimal (N-DF-FSFOM), as the following (nonconvex) QCQP:

$$\mathcal{R}^\star(\mathcal{M}_N) = \left(\begin{array}{l} \text{minimize} \quad \nu R^2 \\ \text{subject to} \\ -a^f_{-1,N} - a^h_{-1,N} + \sum_{i,j\in I_f, i\neq j} \lambda_{i,j} a^f_{i,j} + \sum_{i,j\in I_g, i\neq j} \tau_{i,j} a^h_{i,j} = 0, \\ \nu B_{0,-1} + \sum_{i,j\in I_f, i\neq j} \lambda_{i,j} \left(A^f_{i,j}(\phi,\psi,\alpha,\beta) + \frac{1}{2L} C^f_{i,j} \right) \\ \quad + \sum_{i,j\in I_g, i\neq j} \tau_{i,j} A^h_{i,j}(\phi,\psi,\alpha,\beta) = Z, \\ P \text{ is lower triangular with nonnegative diagonals}, \\ PP^\intercal = Z, \\ \nu \geq 0, \\ \lambda_{i,j} \geq 0, \quad i,j \in I_f, i \neq j, \\ \tau_{i,j} \geq 0, \quad i,j \in I_h, i \neq j, \end{array}\right)$$

where the inner dual variables ν, λ, τ, Z, and the stepsizes $\phi, \psi, \alpha, \beta$ are the decision variables. We now have a nonconvex QCQP that can be solved to global optimality using spatial branch-and-bound algorithms [1,48,38,16]. The optimal stepsizes found will correspond to OptISTA.

3.4 Finding analytical solution of QCQP $\Leftrightarrow$ Finding OptISTA and its convergence proof

We use a spatial branch-and-bound algorithm to find global solutions to the nonconvex QCQP equivalent to ($\mathcal{O}^{\text{outer}}$) for $N = 1, \dots 5$. The numerical solutions allow us to guess the analytical forms of the optimal stepsizes and inner-dual variables as a function of N. Surprisingly, if we additionally assume $\alpha = \beta$ and $\phi = \psi$ in ($\mathcal{O}^{\text{outer}}$), many of the values were similar to the stepsize of OGM without hurting the optimal performance. This observation, along with some trial-and-error, enables to recover the analytical form. We then solve the nonconvex QCQP for larger values of N (e.g., $N = 6, \dots, 25$) and confirm that the numerical values of optimal stepsizes and inner-dual variables agree with our conjectured guess. This gives us confidence that our guess correctly represents the optimal stepsizes and inner-dual variables, although this is not yet a formal proof. The formal proof of Theorem 1 is obtained by identifying the inner-dual variables of ($\mathcal{O}^{\text{outer}}$).

We list optimal stepsizes $\phi, \psi, \alpha, \beta$ and dual variables ν, λ, τ, Z as follows. For primal variables,

$$\alpha_{i+1,j} = \begin{cases} \alpha_{j+1,j} + \sum_{k=j+1}^{i} \left(\frac{2\theta_j}{\theta_{k+1}} - \frac{1}{\theta_{k+1}} \alpha_{k,j} \right) & \text{if } j \in [0 : i-1] \\ 1 + \frac{2\theta_i - 1}{\theta_{i+1}} & \text{if } j = i \end{cases} \qquad \text{and} \quad \beta = \alpha.$$

$$\phi_{i+1,j} = \alpha_{N,j} \text{ for } 1 \leq i \leq N, 0 \leq j \leq i \quad \text{and} \quad \psi = \phi.$$

For inner-dual variables, $\nu = \frac{1}{2(\theta_N^2 - 1)}$ and

$$\lambda = \begin{cases} \lambda_{\star,i} = \frac{2\theta_i}{\theta_N^2} \text{ for } i \in [0 : N-1] \\ \lambda_{\star,N} = \frac{1}{\theta_N} \\ \lambda_{i,i+1} = \frac{2\theta_i^2}{\theta_N^2} \text{ for } \in [0 : N-1] \\ \lambda_{i,j} = 0 \text{ otherwise}, \end{cases} \qquad \tau = \begin{cases} \tau_{-1,N+i} = \frac{2\tilde{\theta}_{i-1}}{\theta_N^2 - 1} \text{ for } i \in [1 : N] \\ \tau_{N+i,N+j} = \frac{2\tilde{\theta}_{j-1}}{\theta_N^2 - 2\theta_i^2 + \theta_i} - \frac{2\tilde{\theta}_{j-1}}{\theta_N^2 - 2\theta_{i-1}^2 + \theta_{i-1}} \text{ for } 1 \leq i < j \leq N \\ \tau_{N+i+1,N+i} = \frac{\theta_i - 1}{\theta_N^2 - 2\theta_i^2 + \theta_i} \text{ for } i \in [1 : N-1] \\ \tau_{i,j} = 0 \text{ otherwise}, \end{cases}$$

and

$$Z = \nu B_{0,-1} + \sum_{i,j\in I_f, i\neq j} \lambda_{i,j} \left(A^f_{i,j}(\phi,\psi,\alpha,\beta) + \frac{1}{2L} C^f_{i,j} \right) \quad + \sum_{i,j\in I_g, i\neq j} \tau_{i,j} A^h_{i,j}(\phi,\psi,\alpha,\beta).$$

Of course, the most direct way to prove Theorem 1 is to show that the dual variables of ($\mathcal{O}^{\text{outer}}$) are indeed a feasible point. However, this approach is inevitably very complex, as it requires, for example, demonstrating that Z is positive semidefinite even if the analytical form is known. Therefore,

we constructed the proof of Theorem 1 by utilizing inner dual variables and stepsizes to construct nonincreasing Lyapunov sequences.

In fact, the dual variables are related to the $\{\mathcal{U}_k\}_{k\in[-1:N]}$ sequence in such a way that the sequence is a linear combination of interpolation inequalities weighted by a dual variable. For example, in a single L-smooth convex function minimization, we define $\mathcal{U}_k$ to be

$$\begin{aligned}\mathcal{U}_k = \nu\|x_0 - x_\star\|^2 + \sum_{i=0}^{k-1} \lambda_{i,i+1}\left(f(x_{i+1}) - f(x_i) + \langle \nabla f(x_i), x_{i+1} - x_i\rangle\right)\\ + \sum_{i=0}^{k} \lambda_{\star,i}\left(f(x_\star) - f(x_i) + \langle \nabla f(x_i), x_\star - x_i\rangle\right),\end{aligned}$$

where ν, $\lambda_{i,i+1}$, and $\lambda_{\star,i}$ are nonnegative dual variables of corresponding stepsize-optimization PEP [40]. If one start the analysis from here, then the dissipative property of $\mathcal{U}_k$ is straightforward from the cococervity inequality of f. However, the main source of difficulty is moved to showing that $\mathcal{U}_N \geq f(x_N) - f(x_\star)$. Unfortunately, this is fundamentally equivalent to showing that the dual variable Z is positive semidefinite.

In short, our options were whether to start the proof from assuming $\mathcal{U}_N \geq f(x_N) - f(x_\star)$ and showing the nonincreasing property, or to choose the opposite approach which is essentially equal to showing that Z is positive semidefinite. We chose the former approach.

Potential non-uniqueness of the optimal method. In solving ($\mathcal{O}^{\text{outer}}$) numerically, we additionally assumed $\alpha = \beta$ and $\phi = \psi$ since we observed that adding those constraints did not worsen the worst-case performance. Under this additional assumption, our numerics lead us to believe that the following values are unique:

$$\begin{aligned}\tau_{-1,N+i} &= \frac{2\tilde{\theta}_{i-1}}{\theta_N^2 - 1} && \text{for } i \in [1:N],\\ \phi_{i,0} &= \frac{2(\theta_N^2 - 1)}{\theta_N^2} && \text{for } i \in [1:N],\\ \phi_{N,i} &= \alpha_{N,i} && \text{for } i \in [0:N-1].\end{aligned}$$

However, the remaining values of ϕ, ψ and τ seem to be not unique, even with the additional constraints $\alpha = \beta$ and $\phi = \psi$. The presented stepsizes and dual variables of OptISTA were chosen as they were the most analytically tractable that the authors could find. It may be that OptISTA is the only analytically tractable choice, or it may be that there is a simpler exact optimal algorithm. Further exploring the set of exact optimal algorithms beyond OptISTA is an interesting direction of future work.

Global optimality through lower bound. The resulting proof of Theorem 1 is likely globally optimal over methods of the form (N-DF-FSFOM) as it agrees with the global numerical solutions of the spatial branch-and-bound solvers, but Theorem 1, by itself, does not guarantee global optimality. In other words, the proof of Theorem 1, by itself, shows that it is a valid proof, but it does not show whether or not the proof can be improved. Rather, global optimality is established through the matching lower bound of Section 4. In Section 4, we prove that all deterministic first-order methods (precisely defined later) respect a lower bound that exactly matches the upper bound of Theorem 1. Since (N-DF-FSFOM) are instances of deterministic first-order methods, global optimality is proved.

4 Exact matching lower bounds

In this section, we present complexity lower bounds that establish exact optimality of OptISTA and OPPA. The explicit construction uses double-function semi-interpolated zero-chain construction, which extends the prior construction of [25] to the double-function setup.

4.1 Composite optimization lower bound

Let $N > 0$ be the total iteration count and $x_0 = z_0 \in \mathbb{R}^d$ be a starting point. We say a method satisfies the *double-function span condition* if it produces an output x_N satisfying:

$$\begin{aligned}
&\delta_i \in \{0, 1\} && \text{for } i = 0, \dots, 2N-1,\\
&\sum_{i=0}^{2N-1} \delta_i = \sum_{i=0}^{2N-1} (1-\delta_i) = N, && \triangleright \text{ exactly } N \text{ evaluations of } \nabla f \text{ and } \mathbf{prox}_{\gamma_i h} \text{ in any order}\\
&d_i = \begin{cases} \nabla f(z_i) & \text{if } \delta_i = 0,\\ z_i - \mathbf{prox}_{\gamma_i h}(z_i) \text{ for some } \gamma_i > 0, & \text{if } \delta_i = 1, \end{cases} && \text{for } i = 0, \dots, 2N-1,\\
&z_i \in x_0 + \operatorname{span}\{d_0, \dots, d_{i-1}\} && \text{for } i = 1, \dots, 2N-1,\\
&x_N \in x_0 + \operatorname{span}\{d_0, \dots, d_{2N-1}\}. && \triangleright \text{ call output of method } x_N \text{ for consistency with Section 3}
\end{aligned}$$

To clarify, we make no assumptions about the order in which the gradient and proximal oracles are used. We only assume that the gradient and proximal oracles are each called exactly N times.

Theorem 2 *Let $L > 0$, $R > 0$, $N > 0$, and $d \geq N + 1$. Let $x_0 \in \mathbb{R}^d$ be any starting point and x_N be generated by a method satisfying the double-function span condition. Then, there is an $f \colon \mathbb{R}^d \to \mathbb{R}$ that is L-smooth and convex and an $h \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ that is an indicator function of a nonempty closed convex set such that there is an $x_\star \in \operatorname{argmin}(f + h)$ satisfying $\|x_0 - x_\star\| = R$ and*

$$f(x_N) + h(x_N) - f(x_\star) - h(x_\star) \geq \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2 - 1)},$$

where θ_N is as defined for OptISTA*.*

Prior lower bounds for the single-function setup of minimizing an L-smooth convex function f (without h) are applicable to OptISTA. Such bounds include the classical $3L\|x_0 - x_\star\|^2/(32(N+1)^2)$ bound by [58, Theorem 2.1.7] and the more modern $L\|x_0 - x_\star\|^2/(2\theta_N^2)$ bound by [23]. However, the composite optimization setup is a harder problem setup, and our lower bound, based on a double-function construction, establishes a stronger (larger) lower bound and exactly matches the rate of OptISTA.

The fact that our construction uses an h that is an indicator function (so that $\mathbf{prox}_{\gamma_i h}$ is a projection onto a convex set) implies that OptISTA is also an optimal projected-gradient type method. Loosely speaking, if we write $\mathcal{PC}$ to denote the problem complexity, then

$$\mathcal{PC}\Big(\underset{x}{\text{minimize}}\ f(x)\Big) < \mathcal{PC}\Big(\underset{x \in C}{\text{minimize}}\ f(x)\Big) = \mathcal{PC}\Big(\underset{x}{\text{minimize}}\ f(x) + h(x)\Big)$$

where C denotes nonempty closed convex sets, if the methods utilize ∇f, $\mathit{\Pi}_C$, and $\mathbf{prox}_{\gamma_i h}$ as oracles, where $\mathit{\Pi}_C$ denotes the projection onto C.

4.2 Double-function semi-interpolated zero-chain construction

Next, we describe the double-function semi-interpolated zero-chain construction used to establish the lower bound of Theorem 2. Let $I = [0 : N] \cup \{\star\}$, let $x_i, g_i \in \mathbb{R}^{N+1}$, $f_i \in \mathbb{R}$ for $i \in I$, and let $\mathit{\Delta}_I = \{\alpha \in \mathbb{R}^{|I|} \mid \sum_{i \in I} \alpha_i = 1,\ \alpha_i \geq 0 \text{ for } i \in I\}$. We define $f \colon \mathbb{R}^{N+1} \to \mathbb{R}$ as

$$f(x) = \max_{\alpha \in \mathit{\Delta}_I} \tfrac{L}{2}\|x\|^2 - \tfrac{L}{2}\Big\|x - \tfrac{1}{L}\sum_{i \in I} \alpha_i g_i\Big\|^2 + \sum_{i \in I} \alpha_i \left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2\right),$$

and $h \colon \mathbb{R}^{N+1} \to \mathbb{R} \cup \{\infty\}$ as $h(x) = \delta_C(x)$, where δ_C is the indicator function of the convex set $C = x_\star + \operatorname{cone}\{g_0, g_1, \dots, g_N\}$ (here cone denotes the set of conic combinations), satisfying $\delta_C(x) = 0$ if

$x \in C$ and $\delta_C(x) = \infty$ if $x \notin C$. The construction of f is the semi-interpolated zero-chain construction and it is L-smooth and convex [25, Theorem 1].

In the single-function semi-interpolated zero-chain construction of [25], the authors carefully choose x_i, g_i and f_i so that the corresponding parameterized function f satisfies so-called *first-order zero-chain property*:

$$x \in \operatorname{span}\{e_0, \dots, e_{i-1}\} \Rightarrow \left[\, \nabla f(x) \in \operatorname{span}\{e_0, \dots, e_i\} \,\right]$$

for all $i \in [0 : N]$ and $\gamma > 0$ (we use the convention $\operatorname{span}\{\} = \{0\}$) and standard unit vectors $e_0, \dots, e_N \in \mathbb{R}^{N+1}$. This property makes each iteration of a first-order method contained in a subspace spanned by unit vectors. However, similar properties have not been known for the proximal operator of a possibly nonsmooth CCP function. Our contribution extends the single-function semi-interpolated zero-chain construction of [26] to identify conditions under which the proximal operators also exhibit zero-chain-like properties on δ_C, and the following lemma demonstrates this.

Lemma 5 *Let $e_0, \dots, e_N \in \mathbb{R}^{N+1}$ be the standard unit vectors. If*

$$\begin{aligned}
&g_i = La_i e_i,\ a_i > 0, \quad \text{for } i \in [0:N],\\
&x_0 = 0, \qquad x_i \in -\operatorname{cone}\{e_0, \dots, e_{i-1}\} \quad \text{for } i \in [1:N], \qquad x_\star \in -\operatorname{cone}\{e_0, e_1, \dots, e_N\},\\
&f_j - \tfrac{1}{L}\langle g_i, g_j\rangle + \tfrac{1}{2L}\|g_j - Lx_j\|^2 - \tfrac{L}{2}\|x_j\|^2 \geq f_k - \tfrac{1}{L}\langle g_i, g_k\rangle + \tfrac{1}{2L}\|g_k - Lx_k\|^2 - \tfrac{L}{2}\|x_k\|^2\\
&\qquad\qquad \text{for } \ i \in [0:N],\ j \in [0:N-1],\ k \in [j+1:N],\\
&\sigma_i \geq 0,\ i \in [0:N], \qquad \sum_{i=0}^N \sigma_i = 1, \qquad g_\star = \sum_{i=0}^N \sigma_i g_i,\\
&\sum_{i=0}^N \sigma_i \left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2\right) = f_\star + \tfrac{1}{2L}\|g_\star - Lx_\star\|^2 - \tfrac{L}{2}\|x_\star\|^2,
\end{aligned}$$

then, for all $i \in [0:N]$ and $\gamma > 0$ (we use the convention $\operatorname{span}\{\} = \{0\}$),

$$x \in \operatorname{span}\{e_0, \dots, e_{i-1}\} \Rightarrow \left[\, \nabla f(x) \in \operatorname{span}\{e_0, \dots, e_i\} \,\right]$$

$$x \in \operatorname{span}\{e_0, \dots, e_{i-1}\} \Rightarrow \left[\, \mathbf{prox}_{\gamma h}(x) = \mathbf{prox}_h(x) \in \operatorname{span}\{e_0, \dots, e_{i-1}\} \,\right]$$

and

$$\inf_{x \in \operatorname{span}\{e_0, \dots, e_{N-1}\}} f(x) \geq f_N.$$

To clarify, the $x_0, \dots, x_N$ in Lemma 5 is unrelated to the x-, y-, and z-iterates in OptISTA. We devote next section for the proof.

4.3 Proof of Lemma 5 and Theorem 2

Let $I = [0:N] \cup \{\star\}$, and let $x_i, g_i \in \mathbb{R}^{N+1}$ and $f_i \in \mathbb{R}$ for $i \in I$. As in the previous section, define the L-smooth convex function $f\colon \mathbb{R}^{N+1} \to \mathbb{R}$ as

$$v(x, \alpha) = \tfrac{L}{2}\|x\|^2 - \tfrac{L}{2}\|x - \tfrac{1}{L}\sum_{i \in I} \alpha_i g_i\|^2 + \sum_{i \in I} \alpha_i \left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2\right),$$

$$f(x) = \max_{\alpha \in \Delta_I} v(x, \alpha) = \max_{\alpha \in \Delta_I} \tfrac{L}{2}\|x\|^2 - \tfrac{L}{2}\|x - \tfrac{1}{L}\sum_{i \in I} \alpha_i g_i\|^2 + \sum_{i \in I} \alpha_i \left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2\right),$$

and $h\colon \mathbb{R}^{N+1} \to \mathbb{R} \cup \{\infty\}$ as $h(x) = \delta_C(x)$, where $C = x_\star + \operatorname{cone}\{g_0, g_1, \dots, g_N\}$. We start with some preliminary lemmas.

Lemma 6 *Let* $\Delta_{I\setminus\{\star\}} = \{(\tilde{\alpha}_0, \dots, \tilde{\alpha}_N) \mid \sum_{i=0}^N \tilde{\alpha}_i = 1,\ \tilde{\alpha}_i \geq 0 \text{ for } i \in [0:N]\}$. *Suppose,*

$$\tilde{f}(x) = \max_{\tilde{\alpha}\in\Delta_{I\setminus\{\star\}}} v(x,\tilde{\alpha}) = \max_{\tilde{\alpha}\in\Delta_{I\setminus\{\star\}}} \tfrac{L}{2}\|x\|^2 - \frac{L}{2}\|x - \tfrac{1}{L}\sum_{i\in I\setminus\{\star\}} \tilde{\alpha}_i g_i\|^2 + \sum_{i\in I\setminus\{\star\}} \tilde{\alpha}_i \left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2\right).$$

and

$$\begin{aligned}
&x_0 = 0,\\
&\langle g_i, g_j\rangle = \langle g_i, g_k\rangle \text{ for } 0 \leq i < j \leq N-1,\ k \in K_j,\\
&f_j - \frac{1}{L}\langle g_i, g_j\rangle + \frac{1}{2L}\|g_j - Lx_j\|^2 - \frac{L}{2}\|x_j\|^2 \geq f_k - \frac{1}{L}\langle g_i, g_k\rangle + \frac{1}{2L}\|g_k - Lx_k\|^2 - \frac{L}{2}\|x_k\|^2\\
&\text{for } i \in I,\ j \in [0:N-1],\ k \in K_j,
\end{aligned}$$

and g_j *is linearly separable from* $\{g_k\}_{k\in K_j}$, *where*

$$K_j = \{k \in I \setminus \{\star\} : g_k \text{ is linearly independent of} \{g_0, \dots, g_j\}\}.$$

Then if $x \in \operatorname{span}\{g_0, \dots, g_{i-1}\}$, *then* $\nabla f(x) \in \operatorname{span}\{g_0, \dots, g_i\}$ *for* $i \in [0:N-1]$.

Proof This is an instance of [25, Theorem 3]. □

Lemma 7 *The following inequality holds for any* $x \in \mathbb{R}^d$.

$$f(x) \geq f_i + \langle g_i, x - x_i\rangle, \quad \text{for } i \in I.$$

Proof We refer the reader to [25, Theorem 1]. □

Lemma 8 *Let* $\Delta_{I\setminus\{\star\}} = \{(\alpha_0, \dots, \alpha_N) \mid \sum_{i=0}^N \alpha_i = 1\ \alpha_i \geq 0 \text{ for } i \in [0:N]\}$ *and let*

$$\tilde{f}(x) = \max_{\tilde{\alpha}\in\Delta_{I\setminus\{\star\}}} v(x,\tilde{\alpha}) = \max_{\tilde{\alpha}\in\Delta_{I\setminus\{\star\}}} \tfrac{L}{2}\|x\|^2 - \frac{L}{2}\|x - \tfrac{1}{L}\sum_{i\in I\setminus\{\star\}} \tilde{\alpha}_i g_i\|^2 + \sum_{i\in I\setminus\{\star\}} \tilde{\alpha}_i \left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2\right).$$

If

$$\begin{aligned}
&g_\star = \sum_{i=0}^N \sigma_i g_i \text{ for some } \sigma_i \geq 0,\ \sum_{i=0}^N \sigma_i = 1,\\
&\sum_{i=0}^N \sigma_i \left(f_i + \frac{1}{2L}\|g_i - Lx_i\|^2 - \frac{L}{2}\|x_i\|^2\right) = f_\star + \frac{1}{2L}\|g_\star - Lx_\star\|^2 - \frac{L}{2}\|x_\star\|^2,
\end{aligned}$$

Then $f \equiv \tilde{f}$.

Proof It is clear that $\tilde{f} \leq f$ pointwise. Now fix x and let $\alpha^\star = (\alpha_0^\star, \dots, \alpha_N^\star, \alpha_\star^\star) \in \Delta_I$ be optimal for x in f. Then define

$$\tilde{\alpha}^\star = (\alpha_0^\star + \sigma_0\alpha_\star^\star, \dots, \alpha_N^\star + \sigma_N\alpha_\star^\star) \in \Delta_{I\setminus\{\star\}}.$$

Under the assumptions of the lemma,

$$v(x,\tilde{\alpha}) \leq \tilde{f}(x) \leq f(x) = v(x,\alpha^\star) = v(x,\tilde{\alpha}^\star) \quad \forall \tilde{\alpha} \in \Delta_{I\setminus\{\star\}}.$$

Therefore $\tilde{\alpha}^\star$ is optimal for x in $\tilde{f}$ and $f(x) = \tilde{f}(x)$. □

Lemma 9 *Suppose* $\{g_i\}_{i\in[0:N]}$ *are orthogonal and*

$$x_\star = -\sum_{i=0}^N s_i g_i \text{ for some } s_i \geq 0.$$

Then, for any $\gamma > 0$ *and* $J \subseteq [0:N]$, *if* $x \in \operatorname{span}\{g_i\}_{i\in J}$, *then* $\mathbf{prox}_{\gamma h}(x) \in \operatorname{span}\{g_i\}_{i\in J}$.

Proof Let $h_i(x) = \delta_{\{x \mid x \geq -s_i\}}(x)$ for $i \in [0:N]$. If we represent x and $x_\star$ into coordinates with respect to orthogonal basis $\{g_i\}_{i\in[0:N]}$, i.e.,

$$x = (t_0, t_1, \ldots, t_N) \quad x_\star = (-s_0, -s_1, \ldots, -s_N).$$

Then $x \in C$ if and only if $t_i \geq -s_i$ for $i \in [0:N]$. Also, It is well known that

$$\mathbf{prox}_{\gamma h}(x) = \mathit{\Pi}_C(x) = \operatorname*{argmin}_{y\in C} \|y - x\|^2$$

and $\mathbf{prox}_{\gamma h}(\cdot)$ splits coordinate-wise:

$$\mathbf{prox}_{\gamma h}(x) = (\mathbf{prox}_{\gamma h_0}(t_0), \ldots, \mathbf{prox}_{\gamma h_N}(t_N)),$$

$$\mathbf{prox}_{\gamma h_i}(t_i) = \operatorname*{argmin}_{y_i \geq -s_i} \|y_i - x_i\|^2.$$

Then if $i \notin J$,

$$\mathbf{prox}_{\gamma h_i}(t_i) = \mathbf{prox}_{\gamma h_i}(0) = 0.$$

Therefore $\mathbf{prox}_{\gamma h}(x) \in \{g_i\}_{i\in J}$. □

Now we can prove Lemma 5.

Proof of Lemma 5. Recall the conditions of Lemma 5:

$$g_i = La_ie_i,\ a_i > 0, \quad \text{for } i \in [0:N], \tag{3}$$

$$x_0 = 0, \qquad x_i \in -\text{cone}\{e_0, \ldots, e_{i-1}\} \ \text{ for } i \in [1:N], \qquad x_\star \in -\text{cone}\{e_0, e_1, \ldots, e_N\}, \tag{4}$$

$$f_j - \tfrac{1}{L}\langle g_i, g_j\rangle + \tfrac{1}{2L}\|g_j - Lx_j\|^2 - \tfrac{L}{2}\|x_j\|^2 \geq f_k - \tfrac{1}{L}\langle g_i, g_k\rangle + \tfrac{1}{2L}\|g_k - Lx_k\|^2 - \tfrac{L}{2}\|x_k\|^2$$
$$\text{for } \ i \in [0:N],\ j \in [0:N-1],\ k \in [j+1:N], \tag{5}$$

$$\sigma_i \geq 0,\ i \in [0:N], \qquad \sum_{i=0}^N \sigma_i = 1, \qquad g_\star = \sum_{i=0}^N \sigma_i g_i, \tag{6}$$

$$\sum_{i=0}^N \sigma_i \left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2\right) = f_\star + \tfrac{1}{2L}\|g_\star - Lx_\star\|^2 - \tfrac{L}{2}\|x_\star\|^2. \tag{7}$$

Under the orthogonality assumption (3),

$$K_j = \{k \in I \setminus \{\star\} : g_k \text{ is linearly independent of} \{g_0, \ldots, g_j\}\} = [j+1:N],$$

$$\langle g_i, g_j\rangle = \langle g_i, g_k\rangle = 0 \text{ for } 0 \leq i < j \leq N-1,\ k \in K_j,$$

and g_j is linearly separable from $\{g_k\}_{k\in K_j}$. Together with (5), we can apply Lemma 6. Therefore, we get:

$$x \in \text{span}\{e_0, \ldots, e_{i-1}\} \Rightarrow \nabla\tilde{f}(x) \in \text{span}\{e_0, \ldots, e_i\} \text{ for } i \in [0:N-1].$$

Additionally, we also have (6) and (7), so can apply Lemma 8 to conclude that

$$x \in \text{span}\{e_0, \ldots, e_{i-1}\} \Rightarrow \nabla f(x) = \nabla\tilde{f}(x) \in \text{span}\{e_0, \ldots, e_i\} \text{ for } i \in [0:N-1].$$

For $i = N$, it is immediate from the fact that $\nabla f(x) = \sum_{i\in I} \alpha_i^\star g_i$ where $\alpha^\star$ is optimal for x. Now using the orthogonality condition (3) and condition on $x_\star$ of (4), we can apply Lemma 9 to get

$$x \in \text{span}\{e_0, \ldots, e_i\} \Rightarrow \mathbf{prox}_{\gamma h}(x) = \mathbf{prox}_h(x) \in \text{span}\{e_0, \ldots, e_i\} \text{ for } i \in [0:N],$$

for any $\gamma > 0$. Now we are left to show

$$\inf_{x\in\text{span}\{e_0,\ldots,e_{N-1}\}} f(x) \geq f_N.$$

By Lemma 7, we have

$$f(x) \geq f_N + \langle g_N, x - x_N \rangle.$$

We take $\inf_{x \in \mathrm{span}\{e_0,\ldots,e_{N-1}\}}$ on both sides to conclude

$$\inf_{x \in \mathrm{span}\{e_0,\ldots,e_{N-1}\}} f(x) \geq f_N + \inf_{x \in \mathrm{span}\{e_0,\ldots,e_{N-1}\}} \langle g_N, x \rangle - \langle g_N, x_N \rangle = f_N - \langle g_N, x_N \rangle = f_N.$$

where the first equality is from orthogonality of (3) and the second equality is from condition on x_N in (4). □

It remains to find a choice of $\sigma_0, \ldots, \sigma_N$, $a_0, \ldots, a_N$, $x_0, \ldots, x_N, x_\star$, $g_\star$, $f_0, \ldots, f_N$, and $f_\star$ while satisfying the constraints of Lemma 5. Since the value of f_N serves as a lower bound for OptISTA, we find the choice that maximizes f_N. In Appendix D, we show that the choice below leads to the lower bound of Theorem 2:

$$\begin{aligned}
&\sigma_i = \tfrac{2\theta_i}{\theta_N^2},\ i \in [0:N-1], \qquad \sigma_N = \tfrac{1}{\theta_N},\\
&\zeta_{N+1} = \tfrac{(\theta_N - 1)R^2}{\theta_N^2(2\theta_N - 1)}, \quad \zeta_N = \tfrac{\theta_N}{\theta_N - 1}\zeta_{N+1}, \quad \zeta_i = \tfrac{2\theta_i}{2\theta_i - 1}\zeta_{i+1}\ , i \in [0:N-1],\\
&a_i = \tfrac{1}{\theta_N^2 - 1} \cdot \tfrac{\zeta_i}{\sigma_i\sqrt{\zeta_i - \zeta_{i+1}}},\ i \in [0:N], \qquad x_i = -(\theta_N^2 - 1)\sum_{k=0}^{i-1} \sigma_k a_k e_k,\\
&f_i = \tfrac{L}{2}a_i^2(4\theta_i - 1) - \tfrac{LR^2}{2(\theta_N^2 - 1)^2}\ , i \in [0:N-1], \qquad f_N = \tfrac{LR^2}{2(\theta_N^2 - 1)},\\
&x_\star = -(\theta_N^2 - 1)\sum_{k=0}^{N} \sigma_k a_k e_k, \qquad g_\star = -\tfrac{L}{\theta_N^2 - 1}x_\star, \qquad f_\star = 0.
\end{aligned}$$

4.4 Proximal minimization lower bound

Consider the problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ h(x),$$

where $h\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed, convex, and proper (possibly nonsmooth). Assume a minimizer $x_\star$ exists (not necessarily unique).

Let $N > 0$ be the total iteration count. Let the proximal stepsizes $\gamma_0, \gamma_1, \ldots, \gamma_{N-1}$ be pre-specified positive numbers. The **O**ptimized **P**roximal **P**oint **A**lgorithm (OPPA), presented by Monteiro-Svaiter [52] and Barré-Taylor-Bach [2], has the form

$$\begin{aligned}
y_{i+1} &= \mathbf{prox}_{\gamma_i h}(x_i)\\
x_{i+1} &= y_{i+1} + \frac{\rho_{i+1}(\eta_i - \rho_i)}{\rho_i \eta_{i+1}}(y_{i+1} - y_i) + \frac{\rho_{i+1}\eta_i}{\rho_i \eta_{i+1}}(y_{i+1} - x_i)
\end{aligned} \tag{OPPA}$$

for $i = 0, \ldots, N-1$, where $x_0 = y_0 \in \mathbb{R}^d$ is a starting point and

$$\rho_i = \frac{\gamma_i}{\gamma_0} \text{ for } i = 0, \ldots, N-1, \qquad \eta_i = \begin{cases} 1 & \text{if } i = 0,\\ \frac{\rho_i + \sqrt{\rho_i^2 + \frac{4\rho_i}{\rho_{i-1}}\eta_{i-1}^2}}{2} & \text{if } 1 \leq i \leq N-1. \end{cases}$$

To clarify, this version of OPPA is equivalent to A-HPE of [52] and ORI-PPA [2] in the specific cases detailed in Section 1.1, but is expressed slightly differently. The prior results [2, Theorem 1] establishes the rate:

$$h(y_N) - h(x_\star) \leq \frac{\gamma_{N-1}\|x_0 - x_\star\|^2}{4\gamma_0^2\eta_{N-1}^2}.$$

We note that (OptISTA) does not reduce to (OPPA) when $f = 0$. More precisely, given $f = 0$, $N \geq 2$, and any $L > 0$, there is no choice of $\gamma_0, \ldots, \gamma_{N-1}$ for (OPPA) that makes (OPPA) equivalent to (OptISTA) with the $\gamma_0, \ldots, \gamma_{N-1}$ as specified by the definition of (OptISTA). (Recall, (OPPA) allows the user to choose $\gamma_0, \ldots, \gamma_{N-1}$ while (OptISTA) specifies the values for $\gamma_0, \ldots, \gamma_{N-1}$.) This contrasts with how (OptISTA) reduces to OGM when $g = 0$.

We say a method satisfies the *proximal span condition* if it produces an output x_N satisfying:

$$x_i \in x_0 + \operatorname{span}\{x_0 - \mathbf{prox}_{\gamma_0 h}(x_0), \ldots, x_{i-1} - \mathbf{prox}_{\gamma_{i-1} h}(x_{i-1})\} \qquad \text{for } i = 1, \ldots, N.$$

Theorem 3 *Let $R > 0$, $N > 0$, and $d \geq N + 1$. Let $\gamma_0, \gamma_1, \ldots, \gamma_{N-1}$ be positive numbers. Let $x_0 \in \mathbb{R}^d$ be any starting point and x_N be generated by an method satisfying the proximal span condition. Then there is a closed convex proper function $h \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ such that there is an $x_\star \in \operatorname{argmin} h$ satisfying $\|x_0 - x_\star\| = R$ and*

$$h(x_N) - h(x_\star) \geq \frac{\gamma_{N-1} \|x_0 - x_\star\|^2}{4\gamma_0^2 \eta_{N-1}^2},$$

where η_{N-1} is as defined for OPPA.

We devote next section for the proof. As a short remark, η is a function of "ratio of γ_i", so its direct dependence on γ_i is somewhat complicated. However, if $\frac{\gamma_i}{\gamma_0} \approx 1$ for all i, then $\eta_N \approx \frac{N}{2}$ and thus we have $\mathcal{O}(1/N^2)$ lower bound on the proximal setup. This concludes that $\mathcal{O}(1/N^2)$ upper bound of Güler's second method [35] with uniform proximal stepsister is indeed optimal.

To the best of our knowledge, there are no prior lower bounds for the proximal minimization setup. There are lower bounds for the proximal point method applied to the more general monotone inclusion problems [20,61], but these results lower bound the rate of convergence of fixed-point residuals rather than function values.

4.5 Proof of Theorem 3

In this section, let $I = [0 : N] \cup \{\star\}$, let $x_i \in \mathbb{R}^{N+1}$ for $i \in [0 : N-1] \cup \{\star\}$, $a_i \in \mathbb{R}$ for $i \in [0 : N-1]$, and let $h_i \in \mathbb{R}$ for $i \in I$. Denote $e_0, \ldots, e_N$ for standard unit vectors in $\mathbb{R}^{N+1}$. We assume proximal stepsizes $\{\gamma_i\}_{i \in [0:N-1]}$ are given and fixed. We define the following notations for closed, convex, and proper functions on $\mathbb{R}^{N+1}$:

$$\begin{aligned}
H_i(x) &\triangleq h_i + \langle a_i e_i, x - x_i \rangle \text{ for } i \in [0 : N-1], \\
H_N(x) &\triangleq h_N + \mathbf{1}_{\{x \,|\, \langle e_N, x \rangle \leq 0\}}(x), \quad H_\star(x) = h_\star, \\
H_{[0:i]}(x) &\triangleq \max_{j \in [0:i]} \{h_j + \langle a_j e_j, x - x_j \rangle\} = \max_{j \in [0:i]} H_j(x) \text{ for } i \in [0 : N-1], \\
H_I(x) &\triangleq \max_{j \in I} H_j(x).
\end{aligned}$$

To clarify, $x_0, \ldots, x_N$ is unrelated to the x- and y- iterates in OPPA. We start the proof with preliminary lemmas.

Lemma 10 *Suppose $x \in \operatorname{span}\{e_0, \ldots, e_{N-1}\}$ and $i \in [0 : N-1]$ are fixed. If,*

$$H_i(\mathbf{prox}_{\gamma_i H_{[0:i]}}(x)) \geq H_j(\mathbf{prox}_{\gamma_i H_{[0:i]}}(x)),$$

for all $i < j \leq N, j = \star$, then

$$\mathbf{prox}_{\gamma_i H_{[0:i]}}(x) = \mathbf{prox}_{\gamma_i H_I}(x).$$

Proof For notational convenience, let $z = \mathbf{prox}_{\gamma_i H_{[0:i]}}(x)$. Then $z \in \mathrm{span}\{e_0, \dots, e_{N-1}\}$ since the function value of $H_{[0:i]}(x)$ does not depend on $\langle e_N, x\rangle$, the last coordinate. Under the given condition $H_i(z) \geq H_j(z)$ for all $i < j \leq N, j = \star$, the following holds for every $\tilde{z}$:

$$\begin{aligned} H_I(z) + \frac{\|z-x\|^2}{2\gamma_i} = \max_{\ell \in I} H_\ell(z) + \frac{\|z-x\|^2}{2\gamma_i} &= \max_{0 \leq \ell \leq i} H_\ell(z) + \frac{\|z-x\|^2}{2\gamma_i} \\ &\leq \max_{0 \leq \ell \leq i} H_\ell(\tilde{z}) + \frac{\|\tilde{z}-x\|^2}{2\gamma_i} \\ &\leq \max_{\ell \in I} H_\ell(\tilde{z}) + \frac{\|\tilde{z}-x\|^2}{2\gamma_i} \\ &= H_I(\tilde{z}) + \frac{\|\tilde{z}-x\|^2}{2\gamma_i}. \end{aligned}$$

The second inequality follows from $[0:i] \subseteq I = \{0, 1, \dots, N, \star\}$. So z minimizes $H_I(\cdot) + \frac{\|\cdot - x\|^2}{2\gamma_i}$ and therefore $z = \mathbf{prox}_{\gamma_i H_{[0:i]}}(x) = \mathbf{prox}_{\gamma_i H_I}(x)$. □

The following lemma is a well-known fact of pointwise maximum function.

Lemma 11 *The following holds.*

$$\partial H_{[0:i]}(x) = \left\{ \sum_{\ell=0}^{i} \sigma_\ell a_\ell e_\ell \mid \sigma_\ell \geq 0,\ \sum_\ell \sigma_\ell = 1 \right\}.$$

Proof We refer the readers to [14, Proposition 2.3.12]. □

Now we have the following lemma, which is analogous to Lemma 5.

Lemma 12 *If*

$$x_i \in \mathrm{span}\{e_0, \dots, e_{i-1}\} \quad \textit{for } i \in [0:N-1], \tag{8}$$

$$h_i - \gamma_i a_i^2 \geq h_j \textit{ for } i \in [0:N-1],\ j \in \{i+1, \dots, N, \star\}, \tag{9}$$

then the following holds for $i \in [0:N-1]$. We use the convention $\mathrm{span}\{\} = \{0\}$.

$$x \in \mathrm{span}\{e_0, \dots, e_{i-1}\} \Rightarrow \mathbf{prox}_{\gamma_i H_I}(x) \in \mathrm{span}\{e_0, \dots, e_i\}.$$

Proof Fix $i \in [0:N-1]$ and assume $x \in \mathrm{span}\{e_0, \dots, e_{i-1}\}$ and consider $z = \mathbf{prox}_{\gamma_i H_{[0:i]}}(x)$. Then,

$$0 \in \partial H_{[0:i]}(z) + \frac{1}{\gamma_i}(z - x)$$

and hence

$$z \in x - \gamma_i \partial H_{[0:i]}(z).$$

Therefore by lemma 11,

$$z = \mathbf{prox}_{\gamma_i H_{[0:i]}}(x) = x - \gamma_i \sum_{\ell=0}^{i} \sigma_\ell a_\ell e_\ell,$$

where $\sigma_\ell \geq 0$ and $\sum_{\ell=0}^{i} \sigma_\ell = 1$. Now we show that z satisfies the condition of Lemma 10. By (8), the following holds,

$$\begin{aligned} H_i(\mathbf{prox}_{\gamma_i H_{[0:i]}}(x)) = H_i\Big(x - \gamma_i \sum_{\ell=0}^{i} \sigma_\ell a_\ell e_\ell\Big) &= h_i + \left\langle a_i e_i, x - \gamma_i \sum_{\ell=0}^{i} \sigma_\ell a_\ell e_\ell - x_i \right\rangle \\ &= h_i + \langle a_i e_i, x\rangle - \gamma_i \left\langle a_i e_i, \sum_{\ell=0}^{i} \sigma_\ell a_\ell e_\ell \right\rangle - \langle a_i e_i, x_i\rangle \\ &= h_i - \gamma_i \sigma_i a_i^2 - \langle a_i e_i, x_i\rangle \qquad \triangleright \langle e_i, x\rangle = 0 \\ &\geq h_i - \gamma_i a_i^2. \qquad \triangleright \langle e_i, x_i\rangle = 0 \end{aligned}$$

For $j \in \{i+1, \dots, N-1\}$,

$$\begin{aligned} H_j(\mathbf{prox}_{\gamma_i H_{[0:i]}}(x)) &= H_j\Big(x - \gamma_j \sum_{\ell=0}^{i} \sigma_\ell a_\ell e_\ell\Big) \\ &= h_j + \left\langle a_j e_j, x - \gamma_i \sum_{\ell=0}^{i} \sigma_\ell a_\ell e_\ell - x_j \right\rangle \\ &= h_j + \langle a_j e_j, x\rangle - \gamma_i \left\langle a_j e_j, \sum_{\ell=0}^{i} \sigma_\ell a_\ell e_\ell \right\rangle - \langle g_j, x_j\rangle \qquad \triangleright \langle a_j e_j, x\rangle = 0 \\ &= h_j. \qquad \triangleright \langle e_j, x_j\rangle = 0 \end{aligned}$$

For $j = N$ and $j = \star$,

$$H_N(\mathbf{prox}_{\gamma_i H_{[0:i]}}(x)) = h_N \text{ and } H_\star(\mathbf{prox}_{\gamma_i H_{[0:i]}}(x)) = h_\star$$

are by definition. Therefore by (9),

$$H_i(\mathbf{prox}_{\gamma_i H_{[0:i]}}(x)) \geq H_j(\mathbf{prox}_{\gamma_i H_{[0:i]}}(x))$$

for all $j \in \{i+1, \dots, N, \star\}$. Then by Lemma 10,

$$\mathbf{prox}_{\gamma_i H_{[0:i]}}(x) = \mathbf{prox}_{\gamma_i H_I}(x) \in \operatorname{span}\{e_0, \dots, e_i\}.$$

□

We now introduce the sufficient conditions for Lemma 12.

Lemma 13 *Let $a_i > 0, b_i > 0$, and $\zeta > 0$ for $i \in [0 : N-1]$. If*

$$a_i b_i - a_{i+1} b_{i+1} - \gamma_i a_i^2 \geq 0, \ i \in [0 : N-2], \tag{10}$$

and

$$a_{N-1} b_{N-1} - \gamma_{N-1} a_{N-1}^2 \geq \zeta, \tag{11}$$

then

$$\begin{aligned} x_i &= -\sum_{k=0}^{i-1} b_k e_k, \ i \in [0 : N-1] \\ h_i &= a_i b_i \ i \in [0 : N-1], \quad h_N = \zeta, \quad h_\star = 0. \end{aligned} \tag{12}$$

satisfies (8) and (9), which are the conditions of Lemma 12.

Proof We already have (8) from (12). For (9), fix $i \in [0 : N-2]$ and first assume $j \in \{i+1, \dots, N-1\}$. Then,

$$h_i = a_i b_i$$

and

$$h_j + \gamma_i a_i^2 = a_j b_j + \gamma_i a_i^2.$$

Therefore

$$\begin{aligned} h_i - \gamma_i a_i^2 - h_j &= a_i b_i - \gamma_i a_i^2 - a_j b_j \\ &\geq a_{i+1} b_{i+1} - a_j b_j \\ &\geq a_{i+1} b_{i+1} - a_j b_j - \gamma_{i+1} a_{i+1}^2 \\ &\geq a_{i+2} b_{i+2} - a_j b_j \\ &\geq \cdots \geq a_{j-1} b_{j-1} - a_j b_j \geq \gamma_{j-1} a_{j-1}^2 \geq 0. \end{aligned}$$

If $j = N$, then

$$h_i - h_N - \gamma_i a_i^2 = a_i b_i - \zeta - \gamma_i a_i^2 \geq a_{i+1} b_{i+1} - \zeta \geq \cdots \geq a_{N-1} b_{N-1} - \zeta \geq \gamma_{N-1} a_{N-1}^2 \geq 0.$$

If $j = \star$, then

$$h_i - h_\star - \gamma_i a_i^2 = a_i b_i - \gamma_i a_i^2 \geq a_{i+1} b_{i+1} \geq 0.$$

If $i = N-1$, $h_{N-1} - \gamma_{N-1} a_{N-1}^2 \geq h_N > h_\star = 0$ is immediate from the definition. So, we have (9) for $j \in \{i+1, \ldots, N, \star\}$. □

Now we present our choice of $a_0, \ldots, a_{N-1}$, $b_0, \ldots, b_{N-1}$ and ζ for (12), which satisfies (10) and (11), leads to the lower bound of Theorem 3 for given $R > 0$ and $N > 0$. First, define the followings:

$$\zeta_N = \tfrac{\gamma_{N-1} R^2}{4\gamma_0^2 \eta_{N-1}^2}, \quad \zeta_i = \tfrac{\frac{2\eta_i}{\rho_i}}{\frac{2\eta_i}{\rho_i} - 1} \zeta_{i+1} \ , i \in [0 : N-1], \tag{13}$$

where $\{\rho_i\}_{i \in [0:N-1]}$ and $\{\eta_i\}_{i \in [0:N-1}$ are defined as

$$\rho_i = \frac{\gamma_i}{\gamma_0} \text{ for } i = 0, \ldots, N-1, \qquad \eta_i = \begin{cases} 1 & \text{if } i = 0, \\ \frac{\rho_i + \sqrt{\rho_i^2 + \frac{4\rho_i}{\rho_{i-1}} \eta_{i-1}^2}}{2} & \text{if } 1 \leq i \leq N-1. \end{cases}$$

Then, define $a_0, \ldots, a_{N-1}$, $b_0, \ldots, b_{N-1}$ and ζ as:

$$a_i = \tfrac{\sqrt{\zeta_i - \zeta_{i+1}}}{\sqrt{\gamma_i}}, \ i \in [0 : N-1], \qquad b_i = \tfrac{\sqrt{\gamma_i} \zeta_i}{\sqrt{\zeta_i - \zeta_{i+1}}}, \ i \in [0 : N-1], \qquad \zeta = \zeta_N \tag{14}$$

Note that ζ is strictly decreasing by definition and thus a_i and b_i are well-defined. In next lemma, we show additional properties of (14), which include (10) and (11) of Lemma 13.

Lemma 14 *The following holds for a_i and b_i in* (14)*.*

$$\begin{aligned} &a_i b_i - a_{i+1} b_{i+1} - \gamma_i a_i^2 \geq 0, \ i \in [0 : N-2], \\ &a_{N-1} b_{N-1} - \gamma_{N-1} a_{N-1}^2 \geq \zeta, \\ &\sum_{i=0}^{N-1} b_i^2 = R^2. \end{aligned}$$

Proof By the definition, for $i \in [0 : N-2]$,

$$a_{i+1} b_{i+1} = \zeta_{i+1},$$

and

$$a_i b_i - \gamma_i a_i^2 = \zeta_i - \gamma_i \frac{\zeta_i - \zeta_{i+1}}{\gamma_i} = \zeta_{i+1}.$$

So, $a_i b_i - a_{i+1} b_{i+1} - \gamma_i a_i^2 = 0$. For $i = N-1$,

$$a_{N-1} b_{N-1} = \zeta_{N-1},$$

and

$$a_{N-1} b_{N-1} - \gamma_{N-1} a_{N-1}^2 = \zeta_{N-1} - \gamma_{N-1} \frac{\zeta_{N-1} - \zeta_N}{\gamma_{N-1}} = \zeta_N.$$

So, $a_{N-1}b_{N-1} - \gamma_{N-1}a_{N-1}^2 = \zeta$. Now we prove the second equation. We start with the following:

$$\sum_{i=0}^{N-1} b_i^2 = \sum_{i=0}^{N-1} \gamma_i \frac{\zeta_i^2}{\zeta_i - \zeta_{i+1}} = \sum_{i=0}^{N-1} \gamma_i \frac{\zeta_i^2}{\zeta_i(1 - \frac{\frac{2\eta_i}{\rho_i}-1}{\frac{2\eta_i}{\rho_i}})}$$
$$= \sum_{i=0}^{N-1} \gamma_i \frac{\zeta_i}{\frac{1}{\frac{2\eta_i}{\rho_i}}} = \sum_{i=0}^{N-1} 2\gamma_i\eta_i\zeta_i \frac{1}{\rho_i} = \gamma_0 \sum_{i=0}^{N-1} 2\eta_i\zeta_i.$$

The term $\sum_{i=0}^{N-1} 2\eta_i\zeta_i$ can be rearranged as:

$$\begin{aligned}
\sum_{i=0}^{N-1} 2\eta_i\zeta_i &= \sum_{i=0}^{N-1} 2\eta_i\zeta_N \prod_{j=i}^{N-1} \frac{\frac{2\eta_j}{\rho_j}}{\frac{2\eta_j}{\rho_j} - 1} \\
&= 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}} - 1} + 2\zeta_N \sum_{i=0}^{N-2} \frac{\frac{2\eta_i^2}{\rho_i}}{\frac{2\eta_i}{\rho_i} - 1} \prod_{j=i+1}^{N-1} \frac{\frac{2\eta_j}{\rho_j}}{\frac{2\eta_j}{\rho_j} - 1} \\
&= 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}} - 1} + 2\zeta_N \sum_{i=0}^{N-2} \frac{\frac{2\eta_{i+1}^2}{\rho_{i+1}} - 2\eta_{i+1}}{\frac{2\eta_i}{\rho_i} - 1} \prod_{j=i+1}^{N-1} \frac{\frac{2\eta_j}{\rho_j}}{\frac{2\eta_j}{\rho_j} - 1} \quad \triangleright \text{using } \frac{2\eta_i^2}{\rho_i} = \frac{2\eta_{i+1}^2}{\rho_{i+1}} - 2\eta_{i+1} \\
&= 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}} - 1} + 2\zeta_N \sum_{i=0}^{N-2} \frac{\frac{2\eta_{i+1}}{\rho_{i+1}} - 2}{\frac{2\eta_i}{\rho_i} - 1}\eta_{i+1} \prod_{j=i+1}^{N-1} \frac{\frac{2\eta_j}{\rho_j}}{\frac{2\eta_j}{\rho_j} - 1} \\
&= 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}} - 1} + 2\zeta_N \sum_{i=0}^{N-2} \frac{\frac{2\eta_{i+1}}{\rho_{i+1}} - 2}{\frac{2\eta_i}{\rho_i} - 1} \frac{\frac{2\eta_{i+1}^2}{\rho_{i+1}}}{\frac{2\eta_{i+1}}{\rho_{i+1}} - 1} \prod_{j=i+2}^{N-1} \frac{\frac{2\eta_j}{\rho_j}}{\frac{2\eta_j}{\rho_j} - 1} \\
&= \cdots \\
&= 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}} - 1} + 2\zeta_N \sum_{i=0}^{N-2} \prod_{j=i}^{N-2} \frac{\frac{2\eta_{j+1}}{\rho_{j+1}} - 2}{\frac{2\eta_j}{\rho_j} - 1}\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}} - 1} \\
&= 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}} - 1} + 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}} - 1} \sum_{i=0}^{N-2} \prod_{j=i}^{N-2} \frac{\frac{2\eta_{j+1}}{\rho_{j+1}} - 2}{\frac{2\eta_j}{\rho_j} - 1}.
\end{aligned} \tag{15}$$

Meanwhile,

$$\sum_{i=0}^{N-2} \prod_{j=i}^{N-2} \frac{\frac{2\eta_{j+1}}{\rho_{j+1}} - 2}{\frac{2\eta_j}{\rho_j} - 1} = \frac{\frac{2\eta_{N-1}}{\rho_{N-1}} - 2}{\frac{2\eta_{N-2}}{\rho_{N-2}} - 1}\Big(\frac{\frac{2\eta_{N-2}}{\rho_{N-2}} - 2}{\frac{2\eta_{N-3}}{\rho_{N-3}} - 1}\Big(\cdots\Big(\frac{\frac{2\eta_2}{\rho_2} - 2}{\frac{2\eta_1}{\rho_1} - 1}\Big(\frac{\frac{2\eta_1}{\rho_1} - 2}{\frac{2\eta_0}{\rho_0} - 1} + 1\Big) + 1\Big)\cdots\Big) + 1\Big).$$

Since $\frac{2\eta_0}{\rho_0} - 1 = 1$, by the telescopic product, it reduces to

$$\sum_{i=0}^{N-2} \prod_{j=i}^{N-2} \frac{\frac{2\eta_{j+1}}{\rho_{j+1}} - 2}{\frac{2\eta_j}{\rho_j} - 1} = \frac{2\eta_{N-1}}{\rho_{N-1}} - 2.$$

Then (15) reduces to

$$\begin{aligned}\sum_{i=0}^{N-1} 2\eta_i\zeta_i &= 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}}-1} + 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}}-1}\sum_{i=0}^{N-2}\prod_{j=i}^{N-2}\frac{\frac{2\eta_{i+1}}{\rho_{i+1}}-2}{\frac{2\eta_i}{\rho_i}-1}\\ &= 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}}-1} + 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}}-1}\Big(\frac{2\eta_{N-1}}{\rho_{N-1}}-2\Big)\\ &= 2\zeta_N\eta_{N-1}\frac{\frac{2\eta_{N-1}}{\rho_{N-1}}}{\frac{2\eta_{N-1}}{\rho_{N-1}}-1}\left(1+\frac{2\eta_{N-1}}{\rho_{N-1}}-2\right)\\ &= 2\zeta_N\eta_{N-1}\frac{2\eta_{N-1}}{\rho_{N-1}} = 4\zeta_N\frac{\eta_{N-1}^2}{\rho_{N-1}}.\end{aligned}$$

Hence,

$$\sum_{i=0}^{N-1} b_i^2 = \gamma_0\sum_{i=1}^{N-1} 2\eta_i\zeta_i = 4\gamma_0\zeta_N\frac{\eta_{N-1}^2}{\rho_{N-1}} = 4\gamma_0\frac{\gamma_{N-1}}{4\gamma_0^2\eta_{N-1}^2}R^2\cdot\frac{\eta_{N-1}^2}{\rho_{N-1}} = R^2.$$

□

The following lemma is an restriction of Theorem 3 in the sense that it has fixed starting point $x_0 = z_0 = 0$ and fixed dimension $d = N+1$.

Lemma 15 *Let $R > 0$, $N > 0$ and let $x_0 = z_0 = 0 \in \mathbb{R}^{N+1}$. Let closed, convex, and proper function $H_I\colon \mathbb{R}^{N+1} \to \mathbb{R}\cup\{\infty\}$ as*

$$\begin{aligned} H_i(x) &\triangleq h_i + \langle a_i e_i, x - x_i\rangle \text{ for } i \in [0:N-1],\\ H_N(x) &\triangleq h_N + \mathbf{1}_{\{x\,|\,\langle e_N, x\rangle \le 0\}}(x), \quad H_\star(x) = h_\star,\\ H_{[0:i]}(x) &\triangleq \max_{j\in[0:i]}\{h_j + \langle a_j e_j, x - x_j\rangle\} = \max_{j\in[0:i]} H_j(x) \text{ for } i \in [0:N-1],\\ H_I(x) &\triangleq \max_{j\in I} H_j(x).\end{aligned}$$

with the choice of

$$\begin{aligned} x_i &= -\sum_{k=0}^{i-1} b_k e_k,\ i\in[0:N-1], \quad x_\star = -\sum_{k=0}^{N-1} b_k e_{k+1}\\ h_i &= a_i b_i\ i\in[0:N-1], \quad h_N = \zeta, \quad h_\star = 0.\end{aligned}$$

where

$$a_i = \frac{\sqrt{\zeta_i - \zeta_{i+1}}}{\sqrt{\gamma_i}},\ i\in[0:N-1], \qquad b_i = \frac{\sqrt{\gamma_i}\zeta_i}{\sqrt{\zeta_i-\zeta_{i+1}}},\ i\in[0:N-1], \qquad \zeta = \zeta_N.$$

Then $x_\star \in \operatorname{argmin} H_I$ and satisfies $\|x_0 - x_\star\| = R$ and

$$H_I(x_N) - H_I(x_\star) \ge \frac{\gamma_{N-1}\|x_0 - x_\star\|^2}{4\gamma_0^2\eta_{N-1}^2}$$

for any $\{x_i\}_{i\in[0:N]}$ satisfying the following proximal span condition:

$$x_i \in x_0 + \operatorname{span}\{x_0 - \mathbf{prox}_{\gamma_0 h}(x_0), \ldots, x_{i-1} - \mathbf{prox}_{\gamma_{i-1}h}(x_{i-1})\} \qquad \text{for } i = 1,\ldots,N. \tag{16}$$

To clarify, $\{x_i\}_{i\in[0:N]}$ in (16) is unrelated to all previously defined $\{x_i\}_{i\in[0:N-1]}$.

Proof Note that for $i \in [0 : N-1]$,

$$H_i(x_\star) = h_i + \langle a_i e_i, x_\star - x_i \rangle = a_i b_i + \left\langle a_i e_i, \sum_{k=0}^{N-1} b_k e_k \right\rangle = a_i b_i - a_i b_i = 0.$$

so $H_I(x_\star) = 0$. Also note that $H_I(x) \geq H_\star(x) = h_\star$. Thus $x_\star \in \operatorname{argmin} h$ and $\|x_0 - x_\star\| = R$ by Lemma 14. By Lemma 14 and 13, we have the result of Lemma 12. So, each proximal evaluation will give at most one new next coordinate. Then after N proximal evaluations, the output x_N will be in the span of $\{e_0, \ldots, e_{N-1}\}$ under the condition (16) of the lemma. Now we apply the second result of Lemma 12 to conclude that

$$\begin{aligned} H_I(x_N) - H_I(x_\star) &\geq h_N - 0 \\ &\geq \zeta = \frac{\gamma_{N-1}\|x_0 - x_\star\|^2}{4\gamma_0^2 \eta_{N-1}^2}. \end{aligned}$$

where the first inequality is from the deinition of H_I. □

Then we expand the condition of Lemma 15 to arrive at Theorem 3.

Proof of Theorem 3 Assume $d \geq N + 1$. Take $H_I \in \mathcal{F}_{0,\infty}$ be function defined in Lemma 15, which is embedded in $\mathbb{R}^d$. Call $\tilde{x}_\star$ to be the element of $\operatorname{argmin}_{x \in \mathbb{R}^d} H_I$ in Lemma 15. Now for arbitrary x_0, let $h \colon \mathbb{R}^d \to \mathbb{R}$ be translation of H_I by x_0:

$$h(x) = H_I(x - x_0).$$

Then h is closed, convex, and proper, and $x_\star \triangleq \tilde{x}_\star + x_0 \in \operatorname{argmin}_{x \in \mathbb{R}^d} h$. Now assume $\{x_i\}_{i \in [0:N]}$ is produced from an method that satisfies the proximal span condition. That is,

$$x_i \in x_0 + \operatorname{span}\{x_0 - \mathbf{prox}_{\gamma_0 h}(x_0), \ldots, x_{i-1} - \mathbf{prox}_{\gamma_{i-1} h}(x_{i-1})\} \qquad \text{for } i = 1, \ldots, N.$$

Then for $\tilde{x}_i \triangleq x_i - x_0$, $\{\tilde{x}_i\}_{i \in [0:N]}$ satisfies:

$$\tilde{x}_i \in \tilde{x}_0 + \operatorname{span}\left\{\tilde{x}_0 - \mathbf{prox}_{\gamma_0 H_I}(\tilde{x}_0), \ldots, \tilde{x}_{i-1} - \mathbf{prox}_{\gamma_{i-1} H_I}(\tilde{x}_{i-1})\right\}.$$

This is because $\tilde{x}_0 = 0$ and

$$\mathbf{prox}_{\gamma_i h}(x_i) - x_0 = \mathbf{prox}_{\gamma_i H_I}(x_i - x_0) = \mathbf{prox}_{\gamma_i H_I}(\tilde{x}_i)$$

So,

$$\tilde{x_i} - \mathbf{prox}_{\gamma_i H_I}(\tilde{x}_i) = x_i - x_0 - \mathbf{prox}_{\gamma_i h}(x_i) + x_0 = x_i - \mathbf{prox}_{\gamma_i h}(x_i).$$

Hence, we can apply Lemma 15 on $\{\tilde{x}_i\}_{i \in [0:N]}$ to get

$$H_I(\tilde{x}_N) - H_I(\tilde{x}_\star) \geq \frac{\gamma_{N-1}\|0 - \tilde{x}_\star\|^2}{4\gamma_0^2 \eta_{N-1}^2} = \frac{\gamma_{N-1}\|x_0 - x_\star\|^2}{4\gamma_0^2 \eta_{N-1}^2} = \frac{\gamma_{N-1} R^2}{4\gamma_0^2 \eta_{N-1}^2}.$$

Then we finally get

$$\begin{aligned} h(x_N) - h(\tilde{x}_\star + x_0) &= H_I(\tilde{x}_N) - H_I(\tilde{x}_\star) \\ &\geq \frac{\gamma_{N-1}\|0 - \tilde{x}_\star\|^2}{4\gamma_0^2 \eta_{N-1}^2} = \frac{\gamma_{N-1}\|x_0 - x_\star\|^2}{4\gamma_0^2 \eta_{N-1}^2} = \frac{\gamma_{N-1} R^2}{4\gamma_0^2 \eta_{N-1}^2}, \end{aligned}$$

and h is our desired function. □

4.6 Generalization to deterministic first-order methods via resisting oracle technique

Using Nemirovsky's resisting oracle technique [54,12], we can extend the lower bound of Theorems 2 and 3 to a broader class of deterministic methods that do not necessarily satisfy the span condition. The methods we consider are deterministic methods whose behavior only depends on the output of the oracle evaluations. In Appendix E, we precisely define the classes as "N-step deterministic double-oracle method" and "N-step deterministic proximal-oracle method" in Appendix F . For the composite minimization setup, we make no assumptions about the order in which the gradient and proximal oracles are used. For these classes of deterministic methods, we retain the same lower bound.

Theorem 4 *Let $L > 0$, $R > 0$, $N > 0$, and $d \geq 3N + 1$. Then for any starting point x_0 and N-step deterministic double-oracle method, which output we call x_N, there is an $f\colon \mathbb{R}^d \to \mathbb{R}$ that is L-smooth and convex and an $h\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ that is an indicator function of nonempty closed convex set such that there is a $x_\star \in \operatorname{argmin}(f + h)$ satisfying $\|x_0 - x_\star\| = R$ and*

$$f(x_N) + h(x_N) - f(x_\star) - h(x_\star) \geq \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2 - 1)},$$

where θ_N is as defined for OptISTA*.*

Theorem 5 *Let $R > 0$, $N > 0$, and $d \geq 2N + 1$. Let $\gamma_0, \gamma_1, \ldots, \gamma_{N-1}$ be positive numbers. Then, for any starting point x_0, N-step deterministic proximal-oracle method, which output we call x_N, there is an $h\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ that is closed convex proper function such that there is a $x_\star \in \operatorname{argmin} h$ satisfying $\|x_0 - x_\star\| = R$ and*

$$h(x_N) - h(x_\star) \geq \frac{\gamma_{N-1}\|x_0 - x_\star\|^2}{4\gamma_0^2\eta_{N-1}^2},$$

where η_{N-1} is as defined for OPPA.

5 Conclusion

In this work, we present two methodologies, the double-function PEP and the double-function semi-interpolated zero-chain construction, and use them to find the exact optimal accelerated composite optimization method OptISTA and establish an exact matching lower bound. We also follow an analogous approach to establish a lower bound in the proximal setup that matches the prior upper bound of OPPA. By establishing exact optimality, this work concludes the search for the fastest methods, with respect to the performance measure of *worst-case function value suboptimality*, for the proximal, projected-gradient, and proximal-gradient setups in composite minimization problem involving a smooth convex function and a closed proper convex function. The presented methodology has broad potential applications beyond the setups considered in this paper. The analysis and design of optimization methods with the splitting-method-structure of Douglas–Rachford [21,47,65], ADMM [33,31,29,7], and Frank–Wolfe [30,9] are interesting directions of future work.

A Method reference

We quickly list relevant first-order methods and their convergence rates.

Nesterov's FGM [57]:

$$y_{i+1} = x_i - \frac{1}{L}\nabla f(x_i)$$
$$x_{i+1} = y_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}(y_{i+1} - y_i)$$

with starting point $x_0 = y_0$, with $f \in \mathcal{F}_{0,L}$, and with the sequence θ_i satisfying $\theta_0 = 1$ and $\theta_i = \left(1+\sqrt{1+4\theta_{i-1}^2}\right)/2$ for $i \in [1 : N-1]$. Its convergence rate is

$$f(y_N) - f(x_\star) \leq \frac{L\|x_0 - x_\star\|^2}{2\theta_{N-1}^2} \leq \frac{2L\|x_0 - x_\star\|^2}{(N+1)^2}.$$

OGM [40]:

$$y_{i+1} = x_i - \frac{1}{L}\nabla f(x_i)$$
$$x_{i+1} = y_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}(y_{i+1} - y_i) + \frac{\theta_i}{\theta_{i+1}}(y_{i+1} - x_i)$$

with starting point $x_0 = y_0$, with $f \in \mathcal{F}_{0,L}$, and with the sequence θ_i satisfying $\theta_0 = 1$, $\theta_i = \left(1+\sqrt{1+4\theta_{i-1}^2}\right)/2$ for $i \in [1 : N-1]$, and $\theta_N = \left(1+\sqrt{1+8\theta_{N-1}^2}\right)/2$. Its convergence rate is

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

FISTA [6]:

$$y_{i+1} = \mathbf{prox}_{\frac{1}{L}h}\left(x_i - \frac{1}{L}\nabla f(x_i)\right)$$
$$x_{i+1} = y_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}(y_{i+1} - y_i)$$

with starting point $x_0 = y_0$, with $f \in \mathcal{F}_{0,L}$ and $h \in \mathcal{F}_{0,\infty}$, and with the sequence θ_i satisfying $\theta_0 = 1$, $\theta_i = \left(1+\sqrt{1+4\theta_{i-1}^2}\right)/2$ for $i \in [1 : N-1]$. Its convergence rate is

$$f(y_N) + h(y_N) - f(x_\star) - h(x_\star) \leq \frac{L\|x_0 - x_\star\|^2}{2\theta_{N-1}^2} \leq \frac{2L\|x_0 - x_\star\|^2}{(N+1)^2}.$$

Güler's second method [35]:

$$y_{i+1} = \mathbf{prox}_{\gamma_i h}(x_i)$$
$$x_{i+1} = y_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}(y_{i+1} - y_i) + \frac{\theta_i}{\theta_{i+1}}(y_{i+1} - x_i)$$

with starting point $x_0 = y_0$, with $h \in \mathcal{F}_{0,\infty}$, with $\gamma_{i+1} \geq \gamma_i > 0$ for $i \in [0 : N-1]$, and with the sequence θ_i satisfying $\theta_0 = 1$ and $\theta_i = \left(1+\sqrt{1+4\theta_{i-1}^2}\right)/2$ for $i \in [1 : N-1]$. Its convergence rate is

$$h(y_N) - h(x_\star) \leq \frac{\|x_0 - x_\star\|^2}{4\gamma_0\theta_{N-1}^2}.$$

OPPA[52, 2]:

$$y_{i+1} = \mathbf{prox}_{\gamma_i h}(x_i)$$
$$x_{i+1} = y_{i+1} + \frac{\rho_{i+1}(\eta_i - \rho_i)}{\rho_i\eta_{i+1}}(y_{i+1} - y_i) + \frac{\rho_{i+1}\eta_i}{\rho_i\eta_{i+1}}(y_{i+1} - x_i)$$

with starting point $x_0 = y_0$, with $h \in \mathcal{F}_{0,\infty}$, with the sequence ρ_i satisfying γ_i/γ_0 for $i \in [0 : N-1]$ and η_i satisfying $\eta_0 = 1$ and $\eta_i = \left(\rho_i+\sqrt{\rho_i^2+4\rho_i\eta_{i-1}^2/\rho_{i-1}}\right)/2$ for $i \in [1 : N-1]$. Its convergence rate is

$$h(y_N) - h(x_\star) \leq \frac{\gamma_{N-1}\|x_0 - x_\star\|^2}{4\gamma_0^2\eta_{N-1}^2}.$$

OptISTA:

$$\begin{aligned}
y_{i+1} &= \mathbf{prox}_{\frac{\gamma_i}{L}h}\left(y_i - \frac{\gamma_i}{L}\nabla f(x_i)\right)\\
z_{i+1} &= x_i + \frac{1}{\gamma_i}\left(y_{i+1} - y_i\right)\\
x_{i+1} &= z_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}\left(z_{i+1} - z_i\right) + \frac{\theta_i}{\theta_{i+1}}\left(z_{i+1} - x_i\right)
\end{aligned}$$

with starting point $x_0 = y_0 = z_0$, with $f \in \mathcal{F}_{0,L}$ and $h \in \mathcal{F}_{0,\infty}$, with the sequence θ_i satisfying $\theta_0 = 1$, $\theta_i = \left(1+\sqrt{1+4\theta_{i-1}^2}\right)/2$ for $i \in [1:N-1]$, and $\theta_N = \left(1+\sqrt{1+8\theta_{N-1}^2}\right)/2$, and with $\gamma_i = {2\theta_i\left(\theta_N^2 - 2\theta_i + \theta_i\right)}/{\theta_N^2} > 0$ for $i \in [0:N-1]$. Its convergence rate is

$$f(y_N) + h(y_N) - f(x_\star) - h(x_\star) \leq \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2 - 1)} \leq \frac{L\|x_0 - x_\star\|^2}{(N+1)^2}.$$

B Omitted proof of Lemma 2

We begin the section by introducing some lemmas.

Lemma 16 *Let* $\{\alpha_{i,j}\}_{i\in[1:N],j\in[0:N-1]}$ *and* $\{h_{i,j}\}_{i\in[1:N],j\in[0:N-1]}$ *be*

$$\alpha_{i+1,j} = \begin{cases} \alpha_{j+1,j} + \sum_{k=j+1}^{i}\left(\frac{2\theta_j}{\theta_{k+1}} - \frac{1}{\theta_{k+1}}\alpha_{k,j}\right) & \text{if } j \in [0:i-1],\\ 1 + \frac{2\theta_i - 1}{\theta_{i+1}} & \text{if } j = i, \end{cases}$$

$$h_{i+1,j} = \begin{cases} \alpha_{i+1,j} - \alpha_{i,j} & \text{if } j \in [0:i-1],\\ \alpha_{i+1,j} & \text{if } j = i. \end{cases}$$

Then,

$$\alpha_{i+1,j} = \sum_{k=j+1}^{i+1} h_{k,j} \text{ for } j \in [0:i] \text{ and } h_{i+1,j} = \begin{cases} \frac{\theta_i - 1}{\theta_{i+1}} h_{i,j} & \text{if } j \in [0:i-2],\\ \frac{\theta_i - 1}{\theta_{i+1}}\left(h_{i,i-1} - 1\right) & \text{if } j = i-1,\\ 1 + \frac{2\theta_i - 1}{\theta_{i+1}} & \text{if } j = i. \end{cases}$$

Proof Note that

$$\alpha_{i+1,j} = \begin{cases} \alpha_{i,j} + h_{i+1,j}, & j \in [0:i-1],\\ h_{i+1,j}, & j = i. \end{cases}$$

So, for $j \in [0:i-1]$, we get the following equation:

$$\alpha_{i+1,j} = \alpha_{i,j} + h_{i+1,j} = \alpha_{i-1,j} + h_{i,j} + h_{i+1,j} = \cdots = \alpha_{j+1,j} + \cdots + h_{i+1,j} = h_{j+1,j} + \cdots + h_{i+1,j} = \sum_{k=j+1}^{i+1} h_{k,j}.$$

Since $\alpha_{i+1,j} = h_{i+1,j}$, we have

$$\alpha_{i+1,j} = \sum_{k=j+1}^{i+1} h_{k,j} \text{ for } j \in [0:i].$$

For the second equality, we refer the reader to [40, Proposition 3]. □

Now we convert x and y-iterates of OptISTA into (N-DF-FSFOM) form.

Lemma 17 *Let*

$$\gamma_i = \frac{2\theta_i}{\theta_N^2}\left(\theta_N^2 - 2\theta_i^2 + \theta_i\right) \text{ for } i \in [0:N-1],$$

$$\alpha_{i+1,j} = \begin{cases} \alpha_{j+1,j} + \sum_{k=j+1}^{i}\left(\frac{2\theta_j}{\theta_{k+1}} - \frac{1}{\theta_{k+1}}\alpha_{k,j}\right) & \text{if } j \in [0:i-1],\\ 1 + \frac{2\theta_i - 1}{\theta_{i+1}} & \text{if } j = i. \end{cases}$$

Then, $\{x_1, \ldots, x_N\}$ *and* $\{y_1, \ldots, y_N\}$ *of the following* (N-DF-FSFOM) *form:*

$$\begin{aligned}
y_{i+1} &= x_0 - \sum_{j\in[0:i]} \frac{\gamma_j}{L}\nabla f(x_j) - \sum_{j\in[0:i]} \frac{\gamma_j}{L} h'(y_{j+1}) \quad \text{for } i \in [0:N-1]\\
x_{i+1} &= x_0 - \sum_{j\in[0:i]} \frac{\alpha_{i+1,j}}{L}\nabla f(x_j) - \sum_{j\in[0:i]} \frac{\alpha_{i+1,j}}{L} h'(y_{j+1}) \quad \text{for } i \in [0:N-1]
\end{aligned} \tag{17}$$

is equal to x-iterates and y-iterates generated by OptISTA *respectively.*

Proof Denote $\{\hat{x}_1, \ldots, \hat{x}_N\}$ and $\{\hat{y}_1, \ldots, \hat{y}_N\}$ to be sequence generated by OptISTA. We use mathematical induction. Since we have the same starting point $x_0 = y_0$, $\hat{y}_1 = y_1$ is immediate from the uniqueness of proximal operator, and

$$\hat{y}_1 = \mathbf{prox}_{\frac{\gamma_0}{L}h}\Big(y_0 - \frac{\gamma_0}{L}\nabla f(x_0)\Big) = x_0 - \frac{\gamma_0}{L}\nabla f(x_0) - \frac{\gamma_0}{L}h'(\hat{y}_1).$$

For x-iterate, we have,

$$\begin{aligned}
\hat{x}_1 &= z_1 + \frac{\theta_0 - 1}{\theta_1}(z_1 - z_0) + \frac{\theta_0}{\theta_1}(z_1 - x_0) = x_0 + \frac{1}{\gamma_0}(\hat{y}_1 - y_0) + \frac{1}{\theta_1}\left(x_0 + \frac{1}{\gamma_0}(\hat{y}_1 - y_0) - x_0\right) \\
&= x_0 - \frac{1}{\gamma_0}\cdot\frac{\gamma_0}{L}\left(\nabla f(x_0) + h'(\hat{y}_1)\right) - \frac{1}{\theta_1}\cdot\frac{1}{\gamma_0}\cdot\frac{\gamma_0}{L}\left(\nabla f(x_0) + h'(\hat{y}_1)\right) \\
&= x_0 - \frac{1}{L}\left(1 + \frac{1}{\theta_1}\right)\nabla f(x_0) - \frac{1}{L}\left(1 + \frac{1}{\theta_1}\right)h'(\hat{y}_1) \\
&= x_0 - \frac{\alpha_{1,0}}{L}\nabla f(x_0) - \frac{\alpha_{1,0}}{L}h'(\hat{y}_1) = x_1.
\end{aligned}$$

Now suppose $x_k = \hat{x}_k$ and $y_k = \hat{y}_k$ for $1 \le k \le i$ and $1 \le i \le N-1$. Then, we get $\hat{y}_{i+1} = y_{i+1}$ from the uniqueness of the proximal operator, and

$$\begin{aligned}
\hat{y}_{i+1} &= \mathbf{prox}_{\frac{\gamma_i}{L}h}\Big(\hat{y}_i - \frac{\gamma_i}{L}\nabla f(x_i)\Big) = y_i - \frac{\gamma_i}{L}\nabla f(x_i) - \frac{\gamma_i}{L}h'(\hat{y}_{i+1}) \\
&= x_0 - \sum_{j\in[0:i-1]}\frac{\gamma_j}{L}\nabla f(x_j) - \sum_{j\in[0:i-1]}\frac{\gamma_j}{L}h'(y_{j+1}) - \frac{\gamma_i}{L}\nabla f(x_i) - \frac{\gamma_i}{L}h'(\hat{y}_{i+1}) \\
&= x_0 - \sum_{j\in[0:i]}\frac{\gamma_j}{L}\nabla f(x_j) - \sum_{j\in[0:i]}\frac{\gamma_j}{L}h'(y_{j+1}).
\end{aligned}$$

For x-iterate, we have,

$$\begin{aligned}
z_{i+1} &= x_0 - \sum_{j\in[0:i-1]}\frac{\alpha_{i,j}}{L}\nabla f(x_j) - \sum_{j\in[0:i-1]}\frac{\alpha_{i,j}}{L}h'(y_{j+1}) - \frac{1}{\gamma_i}\left(\frac{\gamma_i}{L}\nabla f(x_i) + \frac{\gamma_i}{L}h'(y_{i+1})\right) \\
&= x_0 - \sum_{j\in[0:i-1]}\frac{\alpha_{i,j}}{L}\nabla f(x_j) - \sum_{j\in[0:i-1]}\frac{\alpha_{i,j}}{L}h'(y_{j+1}) - \frac{1}{L}\nabla f(x_i) - \frac{1}{L}h'(y_{i+1}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\hat{x}_{i+1} &= z_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}(z_{i+1} - z_i) + \frac{\theta_i}{\theta_{i+1}}(z_{i+1} - \hat{x}_i) \\
&= x_0 - \sum_{j\in[0:i-1]}\frac{\alpha_{i,j}}{L}\nabla f(x_j) - \sum_{j\in[0:i-1]}\frac{\alpha_{i,j}}{L}h'(y_{j+1}) - \frac{1}{L}\nabla f(x_i) - \frac{1}{L}h'(y_{i+1}) \\
&- \frac{\theta_i - 1}{\theta_{i+1}}\left(\sum_{j\in[0:i-2]}\frac{h_{i,j}}{L}\nabla f(x_j) + \sum_{j\in[0:i-2]}\frac{h_{i,j}}{L}h'(y_{j+1})\right) \qquad \triangleright \text{ definition of } h_{i,j} \\
&- \frac{\theta_i - 1}{\theta_{i+1}}\left(\frac{\alpha_{i,i-1} - 1}{L}\nabla f(x_{i-1}) + \frac{\alpha_{i,i-1} - 1}{L}h'(y_i)\right) \\
&- \frac{\theta_i - 1}{\theta_{i+1}}\left(\frac{1}{L}\nabla f(x_i) + \frac{1}{L}h'(y_{i+1})\right) - \frac{\theta_i}{\theta_{i+1}}\left(\frac{1}{L}\nabla f(x_i) + \frac{1}{L}h'(y_{i+1})\right).
\end{aligned} \tag{18}$$

By the second equation of Lemma 16, We have

$$\frac{\theta_i - 1}{\theta_{i+1}}h_{i,j} = h_{i+1,j} \text{ for } j \in [0 : i-2],$$

$$\frac{\theta_i - 1}{\theta_{i+1}}\left(\alpha_{i,i-1} - 1\right) = h_{i+1,i-1}.$$

So (18) reduces to

$$\begin{aligned}
\hat{x}_{i+1} &= x_0 - \sum_{j\in[0:i-1]} \frac{\alpha_{i,j}}{L}\nabla f(x_j) - \sum_{j\in[0:i-1]} \frac{\alpha_{i,j}}{L} h'(y_{j+1}) - \frac{1}{L}\nabla f(x_i) - \frac{1}{L}h'(y_{i+1}) \\
&\quad - \sum_{j\in[0:i-2]} \frac{h_{i+1,j}}{L}\nabla f(x_j) - \sum_{j\in[0:i-2]} \frac{h_{i+1,j}}{L} h'(y_{j+1}) - \frac{h_{i+1,i-1}}{L}\nabla f(x_{i-1}) - \frac{h_{i+1,i-1}}{L}h'(y_i) \\
&\quad - \frac{\theta_i - 1}{\theta_{i+1}}\left(\frac{1}{L}\nabla f(x_i) + \frac{1}{L}h'(y_{i+1})\right) - \frac{\theta_i}{\theta_{i+1}}\left(\frac{1}{L}\nabla f(x_i) + \frac{1}{L}h'(y_{i+1})\right) \\
&= x_0 - \sum_{j\in[0:i-1]} \frac{\alpha_{i+1,j}}{L}\nabla f(x_j) - \sum_{j\in[0:i-1]} \frac{\alpha_{i+1,j}}{L} h'(y_{j+1}) - \frac{1}{L}\left(1 + \frac{2\theta_i - 1}{\theta_{i+1}}\right)\nabla f(x_i) - \frac{1}{L}\left(1 + \frac{2\theta_i - 1}{\theta_{i+1}}\right)h'(y_{i+1}) \\
&= x_0 - \sum_{j\in[0:i]} \frac{\alpha_{i+1,j}}{L}\nabla f(x_j) - \sum_{j\in[0:i]} \frac{\alpha_{i+1,j}}{L} h'(y_{j+1}) = x_{i+1}.
\end{aligned}$$

□

We show $x_N = y_N$ by using the following lemma.

Lemma 18 *$\alpha_{N,j}$ is completely characterized by the following recurrent relationship:*

$$\frac{\theta_{j+1}}{2\theta_j - 1}\left(\alpha_{N,j} - 1\right) - 1 = \frac{\theta_{j+1} - 1}{2\theta_{j+1} - 1}\left(\alpha_{N,j+1} - 1\right).$$

Proof By Lemma 16, we have the following for $j \in [0 : N-1]$:

$$\begin{aligned}
\alpha_{N,j} &= \sum_{k=j+1}^{N} h_{k,j} = h_{j+1,j} + \frac{\theta_{j+1} - 1}{\theta_{j+2}}(h_{j+1,j} - 1) + \frac{\theta_{j+2} - 1}{\theta_{j+3}}\frac{\theta_{j+1} - 1}{\theta_{j+2}}(h_{j+1,j} - 1) + \\
&\cdots + \prod_{\ell=j+1}^{N-1}\frac{\theta_\ell - 1}{\theta_{\ell+1}}\left(h_{j+1,j} - 1\right) = h_{j+1,j} + \sum_{k=j+1}^{N-1}\prod_{\ell=j+1}^{k}\frac{\theta_\ell - 1}{\theta_{\ell+1}}\left(h_{j+1,j} - 1\right).
\end{aligned}$$

So we have

$$\alpha_{N,j} - 1 = \left(1 + \sum_{k=j+1}^{N-1}\prod_{\ell=j+1}^{k}\frac{\theta_\ell - 1}{\theta_{\ell+1}}\right)\left(h_{j+1,j} - 1\right) = \left(1 + \sum_{k=j+1}^{N-1}\prod_{\ell=j+1}^{k}\frac{\theta_\ell - 1}{\theta_{\ell+1}}\right)\cdot\frac{2\theta_j - 1}{\theta_{j+1}}.$$

Multiply both sides by $\frac{\theta_{j+1}}{2\theta_j - 1}$ and subtract 1 to get

$$\frac{\theta_{j+1}}{2\theta_j - 1}\left(\alpha_{N,j} - 1\right) - 1 = \sum_{k=j+1}^{N-1}\prod_{\ell=j+1}^{k}\frac{\theta_\ell - 1}{\theta_{\ell+1}}. \tag{19}$$

Similarly for $j+1$, we have the following:

$$\begin{aligned}
\frac{\theta_{j+1} - 1}{2\theta_{j+1} - 1}\left(\alpha_{N,j+1} - 1\right) &= \left(1 + \sum_{k=j+2}^{N-1}\prod_{\ell=j+2}^{k}\frac{\theta_\ell - 1}{\theta_{\ell+1}}\right)\cdot\frac{2\theta_{j+1} - 1}{\theta_{j+2}}\cdot\frac{\theta_{j+1} - 1}{2\theta_{j+1} - 1} \\
&= \left(1 + \sum_{k=j+2}^{N-1}\prod_{\ell=j+2}^{k}\frac{\theta_\ell - 1}{\theta_{\ell+1}}\right)\cdot\frac{\theta_{j+1} - 1}{\theta_{j+2}} = \sum_{k=j+1}^{N-1}\prod_{\ell=j+1}^{k}\frac{\theta_\ell - 1}{\theta_{\ell+1}}.
\end{aligned} \tag{20}$$

Combining (19) and (20), we get the desired result. □

Now we are ready to prove Lemma 2.

Proof of Lemma 2. By Lemma 17, it suffices to show the following holds for $j \in [0 : N-1]$.

$$\gamma_j = \frac{2\theta_j}{\theta_N^2}\left(\theta_N^2 - 2\theta_j^2 + \theta_j\right) = \alpha_{N,j}.$$

First we show that $\gamma_{N-1} = \alpha_{N,N-1}$:

$$\gamma_{N-1} = \frac{2\theta_{N-1}}{\theta_N^2}\left(\theta_N^2 - 2\theta_{N-1}^2 + \theta_{N-1}\right) = \frac{2\theta_{N-1}}{\theta_N^2}\left(\theta_N + \theta_{N-1}\right) = \frac{2\theta_N\theta_{N-1} + 2\theta_{N-1}^2}{\theta_N^2}$$
$$= \frac{2\theta_N\theta_{N-1} + \theta_N^2 - \theta_N}{\theta_N^2} = \frac{2\theta_{N-1} + \theta_N - 1}{\theta_N} = 1 + \frac{2\theta_{N-1} - 1}{\theta_N} = \alpha_{N,N-1}.$$

Now we complete the proof by showing that γ_j satisfies the same defining recurrent relationship satisfied by $\alpha_{N,j}$ in Lemma 18. For $j \in [0 : N-2]$,

$$\begin{aligned}
\frac{\theta_{j+1}}{2\theta_j - 1}\left(\gamma_j - 1\right) - 1 &= \frac{\theta_{j+1}}{2\theta_j - 1} \cdot \left(\frac{2\theta_j}{\theta_N^2}\left(\theta_N^2 - 2\theta_j^2 + \theta_j\right) - 1\right) - 1 \\
&= \frac{2\theta_{j+1}}{\theta_N^2} \cdot \frac{\theta_j}{2\theta_j - 1}\left(\theta_N^2 - 2\theta_j^2 + \theta_j\right) - \frac{\theta_{j+1}}{2\theta_j - 1} - 1 \\
&= \frac{2\theta_{j+1}\theta_N^2}{\theta_N^2} \cdot \frac{\theta_j}{2\theta_j - 1} - \frac{2\theta_{j+1}\theta_j^2}{\theta_N^2} - \frac{\theta_{j+1}}{2\theta_j - 1} - 1 \\
&= \frac{2\theta_{j+1}\theta_j}{2\theta_j - 1} - \frac{2\theta_{j+1}\theta_j^2}{\theta_N^2} - \frac{\theta_{j+1}}{2\theta_j - 1} - 1 \\
&= \frac{\theta_{j+1}\left(2\theta_j - 1\right)}{2\theta_j - 1} - \frac{2\theta_{j+1}\theta_j^2}{\theta_N^2} - 1 \\
&= \theta_{j+1} - \frac{2\theta_{j+1}(\theta_{j+1}^2 - \theta_{j+1})}{\theta_N^2} - 1 = \theta_{j+1} - \frac{2\theta_{j+1}^2(\theta_{j+1} - 1)}{\theta_N^2} - 1
\end{aligned}$$

and

$$\begin{aligned}
\frac{\theta_{j+1} - 1}{2\theta_{j+1} - 1}\left(\gamma_{j+1} - 1\right) &= \frac{\theta_{j+1} - 1}{2\theta_{j+1} - 1} \cdot \left(\frac{2\theta_{j+1}}{\theta_N^2}\left(\theta_N^2 - 2\theta_{j+1}^2 + \theta_{j+1}\right) - 1\right) \\
&= \frac{\theta_{j+1} - 1}{2\theta_{j+1} - 1} \cdot \left(2\theta_{j+1} - 1 - \frac{2\theta_{j+1}^2\left(2\theta_{j+1} - 1\right)}{\theta_N^2}\right) \\
&= \theta_{j+1} - 1 - \frac{2\theta_{j+1}^2\left(\theta_{j+1} - 1\right)}{\theta_N^2}.
\end{aligned}$$

□

C Omitted proof of Theorem 1

First, we prove preliminary lemmas that will be used in the main proof.

Lemma 19 *Let* $\{\theta_i\}_{i=0,\ldots,N}$ *defined as*

$$\theta_i = \begin{cases} 1, & \text{if } i = 0 \\ \frac{1+\sqrt{1+4\theta_{i-1}^2}}{2}, & \text{if } 1 \le i \le N-1, \\ \frac{1+\sqrt{1+8\theta_{N-1}^2}}{2}, & \text{if } i = N. \end{cases}$$

Then $\{\theta_i\}_{i=0,\ldots,N}$ *satisfies*

$$\theta_{i+1}^2 - \theta_{i+1} - \theta_i^2 = 0 \text{ for } i \in [0 : N-2], \tag{21}$$
$$\theta_N^2 - \theta_N - 2\theta_{N-1}^2 = 0, \tag{22}$$
$$\sum_{j=0}^{i} \theta_j = \theta_i^2 \text{ for } i \in [0 : N-1]. \tag{23}$$

Proof Since θ_{i+1} is a root of $x^2 - x - \theta_i^2 = 0$ for $i \in [0 : N-2]$, we have (21). Similarly, θ_N is a root of $x^2 - x - 2\theta_{N-1}^2 = 0$, we have (22). For (23),

$$\sum_{j=0}^{i} \theta_j = \theta_0 + \sum_{j=1}^{i}\left(\theta_j^2 - \theta_{j-1}^2\right) = \theta_0 + \theta_i^2 - \theta_0^2 = \theta_i^2,$$

where $i \in [0 : N-1]$ and the second equality is from the telescopic sum. □

Lemma 20 *Let* $\{\tilde{\theta}_i\}_{i=0,\dots,N-1}$ *defined as*

$$\tilde{\theta}_i = \begin{cases} \theta_i & \text{if } i \in [0:N-2], \\ \frac{2\theta_{N-1}+\theta_N-1}{2} & \text{if } i = N-1. \end{cases}$$

Then,

$$\sum_{j=0}^{N-1} \tilde{\theta}_j = \frac{\theta_N^2 - 1}{2}. \tag{24}$$

Proof

$$\sum_{i=0}^{N-1} \tilde{\theta}_i = \sum_{i=0}^{N-2} \theta_i + \frac{2\theta_{N-1}+\theta_N-1}{2} = \frac{2\theta_{N-2}^2 + 2\theta_{N-1} + \theta_N - 1}{2} = \frac{2\theta_{N-1}^2 + \theta_N - 1}{2} = \frac{\theta_N^2 - 1}{2}.$$

where the second equality uses (23), the third uses (21), and the fourth uses (22) of Lemma 19. □

Now we are ready to start the main proof. We provide the explicit form of Lyapunov sequence $\mathcal{U}_k$ for $k \in [-1:N]$. We first define $\{\tau_{i,j}\}$ where $i,j \in \{\star, 0, 1, \dots, N\}$ as

$$\tau = \begin{cases} \tau_{\star,i} = \frac{2\tilde{\theta}_{i-1}}{\theta_N^2-1} \text{ if } i \in [1:N] \\ \tau_{i,j} = \frac{2\tilde{\theta}_{j-1}}{\theta_N^2 - 2\theta_i^2 + \theta_i} - \frac{2\tilde{\theta}_{j-1}}{\theta_N^2 - 2\theta_{i-1}^2 + \theta_{i-1}} \text{ if } 1 \le i < j \le N, \\ \tau_{i+1,i} = \frac{\theta_i - 1}{\theta_N^2 - 2\theta_i^2 + \theta_i} \text{ if } i \in [1:N-1], \\ \tau_{i,j} = 0 \text{ otherwise.} \end{cases}$$

Let $\theta_{-1} = 0$ and define $\{\mathcal{F}_k\}_{k\in[-1:N]}$ to be:

- $k = N$

$$\mathcal{F}_N = f(x_N) - f(x_\star) + \frac{L}{2\theta_N^2} \left\| w_N - x_\star + \frac{1}{L}\nabla f(x_\star) + \frac{2\theta_{N-1}}{L} h'(y_N) - \frac{\theta_N}{L}\nabla f(x_N) - \frac{2\tilde{\theta}_{N-1}}{L} h'(y_N) \right\|^2,$$

- $k \in [-1:N-1]$

$$\mathcal{F}_k = \frac{2\theta_k^2}{\theta_N^2}\left(f(x_k) - f(x_\star)\right) + \frac{L}{2\theta_N^2}\left\| w_{k+1} - x_\star + \frac{1}{L}\nabla f(x_\star) + \frac{2\theta_k}{L} h'(y_{k+1}) \right\|^2 - \left(\frac{1}{2L} - \frac{\theta_k^2}{L\theta_N^2}\right)\|\nabla f(x_\star)\|^2 - \frac{\theta_k^2}{L\theta_N^2}\|\nabla f(x_k)\|^2,$$

and $\{\mathcal{H}_k\}_{k\in[-1:N]}$ to be

- $k = N$

$$\begin{aligned} \mathcal{H}_N &= h(y_N) - h(x_\star) \\ &+ \frac{L}{2\theta_N^2(\theta_N^2-1)} \left\| x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star) - \sum_{i=0}^{N-1} \frac{2\tilde{\theta}_i}{L} h'(y_{i+1}) \right\|^2 \\ &+ \sum_{i\neq j, i,j\in[1:N]} \frac{\tilde{\theta}_{i-1}\tilde{\theta}_{j-1}}{L\theta_N^2(\theta_N^2-1)} \|h'(y_i) - h'(y_j)\|^2 + \sum_{i=1}^{N-1} \frac{\tilde{\theta}_{i-1}^2}{L\theta_N^2} \|h'(y_i) - h'(y_{i+1})\|^2, \end{aligned}$$

- $k \in [1:N-1]$

$$\begin{aligned} \mathcal{H}_k &= \sum_{i,j\in\{\star,1,\dots,k\}} \tau_{i,j}\left(h(y_j) - h(y_i)\right) \\ &+ \frac{L}{2\theta_N^2(\theta_N^2-1)} \left\| x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star) - \sum_{i=0}^{k-1} \frac{2\theta_i}{L} h'(y_{i+1}) \right\|^2 \\ &+ \sum_{i\neq j, i,j\in[1:k]} \frac{\theta_{i-1}\theta_{j-1}}{L\theta_N^2(\theta_N^2-1)} \|h'(y_i) - h'(y_j)\|^2 + \sum_{i=1}^{k-1} \frac{\theta_{i-1}^2}{L\theta_N^2} \|h'(y_i) - h'(y_{i+1})\|^2 \\ &+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle \nabla f(x_k), h'(y_k)\rangle + \sum_{i=1}^{k}\sum_{\ell=k}^{N-1} \frac{2\tilde{\theta}_\ell\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)} \|h'(y_i)\|^2 + \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)\|^2, \end{aligned}$$

- $k = 0, -1$

$$\mathcal{H}_0 = \mathcal{H}_{-1} = \frac{L}{2\theta_N^2(\theta_N^2 - 1)} \left\| x_0 - x_\star - \frac{\theta_N^2 - 1}{L} \nabla f(x_\star) \right\|^2,$$

Then we let

$$\mathcal{U}_k = \mathcal{F}_k + \mathcal{H}_k \qquad k \in [-1 : N].$$

In order to show that $\{\mathcal{U}_k\}_{k\in[-1:N]}$ is nonincreasing, we start with calculating $\mathcal{F}_k - \mathcal{F}_{k+1}$ for $k \in [-1 : N-1]$. We define $\theta_{-1} = 0$.

Lemma 21 *The following holds.*

$$\mathcal{F}_{N-1} - \mathcal{F}_N \geq \frac{2\tilde{\theta}_{N-1}}{\theta_N^2} \left\langle w_N - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_N) \right\rangle + \frac{\tilde{\theta}_{N-1}(4\theta_{N-1} - 2\tilde{\theta}_{N-1})}{L\theta_N^2} \|h'(y_N)\|^2$$

$$\mathcal{F}_k - \mathcal{F}_{k+1} \geq \frac{2\theta_k}{\theta_N^2} \left\langle w_{k+1} - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_{k+1}) \right\rangle + \frac{2\theta_k^2}{L\theta_N^2} \|h'(y_{k+1})\|^2 + \frac{2\theta_k^2}{L\theta_N^2} \langle h'(y_{k+1}), \nabla f(x_{k+1}) \rangle. \qquad k \in [-1 : N-2].$$

Proof We expand the squares of $\mathcal{F}_{N-1}$ and $\mathcal{F}_N$ to get:

$$
\begin{aligned}
\mathcal{F}_{N-1} - \mathcal{F}_N &= \frac{2\theta_{N-1}^2}{\theta_N^2}\left(f(x_{N-1}) - f(x_\star)\right) - \left(f(x_N) - f(x_\star)\right) - \frac{1}{2L\theta_N}\|\nabla f(x_\star)\|^2 - \frac{\theta_{N-1}^2}{L\theta_N^2}\|\nabla f(x_{N-1})\|^2 \\
&+ \frac{L}{2\theta_N^2}\left\|w_N - x_\star + \frac{1}{L}\nabla f(x_\star) + \frac{2\theta_{N-1}}{L}h'(y_N)\right\|^2 - \frac{L}{2\theta_N^2}\left\|w_N - x_\star + \frac{1}{L}\nabla f(x_\star) + \frac{2\theta_{N-1}}{L}h'(y_N)\right\|^2 \\
&+ \frac{L}{\theta_N^2}\left\langle \frac{\theta_N}{L}\nabla f(x_N) + \frac{2\tilde{\theta}_{N-1}}{L}h'(y_N), w_N - x_\star + \frac{1}{L}\nabla f(x_\star) + \frac{2\theta_{N-1}}{L}h'(y_N)\right\rangle - \frac{L}{2\theta_N^2}\left\|\frac{\theta_N}{L}\nabla f(x_N) + \frac{2\tilde{\theta}_{N-1}}{L}h'(y_N)\right\|^2 \\
&= \frac{2\theta_{N-1}^2}{\theta_N^2}\left(f(x_{N-1}) - f(x_\star) - \frac{1}{2L}\|\nabla f(x_{N-1})\|^2\right) - \left(f(x_N) - f(x_\star)\right) - \frac{1}{2L\theta_N}\|\nabla f(x_\star)\|^2 \\
&+ \frac{L}{\theta_N^2}\left\langle \frac{\theta_N}{L}\nabla f(x_N) + \frac{2\tilde{\theta}_{N-1}}{L}h'(y_N), w_N - x_\star + \frac{1}{L}\nabla f(x_\star) + \frac{2\theta_{N-1}}{L}h'(y_N)\right\rangle \\
&- \frac{1}{2L}\|\nabla f(x_N)\|^2 - \frac{2\tilde{\theta}_{N-1}}{L\theta_N}\left\langle \nabla f(x_N), h'(y_N)\right\rangle - \frac{2\tilde{\theta}_{N-1}^2}{L\theta_N^2}\|h'(y_N)\|^2 \\
&= \frac{2\theta_{N-1}^2}{\theta_N^2}\left(f(x_{N-1}) - f(x_\star) - \frac{1}{2L}\|\nabla f(x_{N-1})\|^2\right) - \left(f(x_N) - f(x_\star)\right) - \frac{1}{2L\theta_N}\|\nabla f(x_\star)\|^2 \\
&+ \frac{1}{\theta_N}\left\langle w_N - x_\star + \frac{1}{L}\nabla f(x_\star), \nabla f(x_N)\right\rangle + \frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_N)\right\rangle \\
&+ \frac{2\theta_{N-1}}{L\theta_N}\left\langle \nabla f(x_N), h'(y_N)\right\rangle + \frac{4\theta_{N-1}\tilde{\theta}_{N-1}}{L\theta_N^2}\|h'(y_N)\|^2 - \frac{1}{2L}\|\nabla f(x_N)\|^2 - \frac{2\tilde{\theta}_{N-1}}{L\theta_N}\left\langle \nabla f(x_N), h'(y_N)\right\rangle - \frac{2\tilde{\theta}_{N-1}^2}{L\theta_N^2}\|h'(y_N)\|^2 \\
&= \frac{2\theta_{N-1}^2}{\theta_N^2}\left(f(x_{N-1}) - f(x_\star) - \frac{1}{2L}\|\nabla f(x_{N-1})\|^2\right) - \frac{\theta_N^2}{\theta_N^2}\left(f(x_N) - f(x_\star) + \frac{1}{2L}\|\nabla f(x_N)\|^2\right) \\
&- \frac{1}{2L\theta_N}\|\nabla f(x_\star)\|^2 + \frac{1}{\theta_N}\left\langle w_N - x_\star + \frac{1}{L}\nabla f(x_\star), \nabla f(x_N)\right\rangle + \frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_N)\right\rangle \\
&- \frac{\theta_N - 1}{L\theta_N}\left\langle \nabla f(x_N), h'(y_N)\right\rangle + \frac{4\theta_{N-1}\tilde{\theta}_{N-1} - 2\tilde{\theta}_{N-1}^2}{L\theta_N^2}\|h'(y_N)\|^2 \\
&= \frac{\theta_N^2 - \theta_N}{\theta_N^2}\left(f(x_{N-1}) - f(x_\star) - \frac{1}{2L}\|\nabla f(x_{N-1})\|^2\right) - \frac{\theta_N^2}{\theta_N^2}\left(f(x_N) - f(x_\star) + \frac{1}{2L}\|\nabla f(x_N)\|^2\right) \qquad \triangleright\, \theta_{N-1}^2 = \theta_N^2 - \theta_N \\
&- \frac{1}{2L\theta_N}\|\nabla f(x_\star)\|^2 + \frac{1}{\theta_N}\left\langle w_N - x_\star, \nabla f(x_N)\right\rangle + \frac{1}{L\theta_N}\left\langle \nabla f(x_\star), \nabla f(x_N)\right\rangle + \frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_N)\right\rangle \\
&- \frac{\theta_N - 1}{L\theta_N}\left\langle \nabla f(x_N), h'(y_N)\right\rangle + \frac{\tilde{\theta}_{N-1}(4\theta_{N-1} - 2\tilde{\theta}_{N-1})}{L\theta_N^2}\|h'(y_N)\|^2 \\
&= \frac{\theta_N^2 - \theta_N}{\theta_N^2}\left(f(x_{N-1}) - f(x_\star) - \frac{1}{2L}\|\nabla f(x_{N-1})\|^2 - f(x_N) + f(x_\star) - \frac{1}{2L}\|\nabla f(x_N)\|^2\right) \\
&+ \frac{\theta_N}{\theta_N^2}\left(f(x_\star) - f(x_N) - \frac{1}{2L}\|\nabla f(x_N)\|^2 - \frac{1}{2L}\|\nabla f(x_\star)\|^2\right) \\
&+ \frac{1}{\theta_N}\left\langle w_N - x_N, \nabla f(x_N)\right\rangle + \frac{1}{\theta_N}\left\langle x_N - x_\star, \nabla f(x_N)\right\rangle + \frac{1}{L\theta_N}\left\langle \nabla f(x_\star), \nabla f(x_N)\right\rangle + \frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_N)\right\rangle \\
&- \frac{\theta_N - 1}{L\theta_N}\left\langle \nabla f(x_N), h'(y_N)\right\rangle + \frac{\tilde{\theta}_{N-1}(4\theta_{N-1} - 2\tilde{\theta}_{N-1})}{L\theta_N^2}\|h'(y_N)\|^2 \\
&= \frac{\theta_N^2 - \theta_N}{\theta_N^2}\left(f(x_{N-1}) - \frac{1}{2L}\|\nabla f(x_{N-1})\|^2 - f(x_N) - \frac{1}{2L}\|\nabla f(x_N)\|^2\right) \\
&+ \frac{\theta_N}{\theta_N^2}\left(f(x_\star) - f(x_N) - \frac{1}{2L}\|\nabla f(x_N)\|^2 - \frac{1}{2L}\|\nabla f(x_\star)\|^2 + \frac{1}{L}\left\langle \nabla f(x_\star), \nabla f(x_N)\right\rangle + \left\langle x_N - x_\star, \nabla f(x_N)\right\rangle\right) \\
&+ \frac{1}{\theta_N}\left\langle w_N - x_N, \nabla f(x_N)\right\rangle + \frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_N)\right\rangle \\
&- \frac{\theta_N - 1}{L\theta_N}\left\langle \nabla f(x_N), h'(y_N)\right\rangle + \frac{\tilde{\theta}_{N-1}(4\theta_{N-1} - 2\tilde{\theta}_{N-1})}{L\theta_N^2}\|h'(y_N)\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\geq \frac{\theta_N^2-\theta_N}{\theta_N^2}\left(f(x_{N-1})-\frac{1}{2L}\|\nabla f(x_{N-1})\|^2-f(x_N)-\frac{1}{2L}\|\nabla f(x_N)\|^2\right)+\frac{1}{\theta_N}\langle w_N-x_N,\nabla f(x_N)\rangle \qquad \triangleright \text{cocoercivity of } f\\
&+\frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N-x_\star+\frac{1}{L}\nabla f(x_\star),h'(y_N)\right\rangle\\
&-\frac{\theta_N-1}{L\theta_N}\langle\nabla f(x_N),h'(y_N)\rangle+\frac{\tilde{\theta}_{N-1}(4\theta_{N-1}-2\tilde{\theta}_{N-1})}{L\theta_N^2}\|h'(y_N)\|^2\\
&=\frac{\theta_N^2-\theta_N}{\theta_N^2}\left(f(x_{N-1})-\frac{1}{2L}\|\nabla f(x_{N-1})\|^2-f(x_N)-\frac{1}{2L}\|\nabla f(x_N)\|^2\right)\\
&+\frac{\theta_N-1}{\theta_N}\left\langle x_N-x_{N-1}+\frac{1}{L}\nabla f(x_{N-1})+\frac{1}{L}h'(y_N),\nabla f(x_N)\right\rangle \qquad \triangleright w_N=\theta_N x_N-(\theta_N-1)z_N, z_N=x_{N-1}-\frac{1}{L}\nabla f(x_{N-1})-\frac{1}{L}h'(y_N)\\
&+\frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N-x_\star+\frac{1}{L}\nabla f(x_\star),h'(y_N)\right\rangle-\frac{\theta_N-1}{L\theta_N}\langle\nabla f(x_N),h'(y_N)\rangle+\frac{\tilde{\theta}_{N-1}(4\theta_{N-1}-2\tilde{\theta}_{N-1})}{L\theta_N^2}\|h'(y_N)\|^2\\
&=\frac{\theta_N^2-\theta_N}{\theta_N^2}\left(f(x_{N-1})-f(x_N)-\frac{1}{2L}\|\nabla f(x_{N-1})\|^2-\frac{1}{2L}\|\nabla f(x_N)\|^2\right)\\
&+\frac{\theta_N-1}{\theta_N}\left\langle x_N-x_{N-1}+\frac{1}{L}\nabla f(x_{N-1}),\nabla f(x_N)\right\rangle\\
&+\frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N-x_\star+\frac{1}{L}\nabla f(x_\star),h'(y_N)\right\rangle+\frac{\tilde{\theta}_{N-1}(4\theta_{N-1}-2\tilde{\theta}_{N-1})}{L\theta_N^2}\|h'(y_N)\|^2\\
&=\frac{\theta_N^2-\theta_N}{\theta_N^2}\left(f(x_{N-1})-\frac{1}{2L}\|\nabla f(x_{N-1})\|^2-f(x_N)-\frac{1}{2L}\|\nabla f(x_N)\|^2+\langle x_N-x_{N-1},\nabla f(x_N)\rangle+\frac{1}{L}\langle\nabla f(x_{N-1}),\nabla f(x_N)\rangle\right)\\
&+\frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N-x_\star+\frac{1}{L}\nabla f(x_\star),h'(y_N)\right\rangle+\frac{\tilde{\theta}_{N-1}(4\theta_{N-1}-2\tilde{\theta}_{N-1})}{L\theta_N^2}\|h'(y_N)\|^2\\
&\geq\frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N-x_\star+\frac{1}{L}\nabla f(x_\star),h'(y_N)\right\rangle+\frac{\tilde{\theta}_{N-1}(4\theta_{N-1}-2\tilde{\theta}_{N-1})}{L\theta_N^2}\|h'(y_N)\|^2 \qquad \triangleright \text{cocoercivity of } f.
\end{aligned}
$$

For $k\in[-1:N-2]$,

$$
\begin{aligned}
&\mathcal{F}_k-\mathcal{F}_{k+1}=\frac{2\theta_k^2}{\theta_N^2}(f(x_k)-f(x_\star))+\frac{L}{2\theta_N^2}\left\|w_{k+1}-x_\star+\frac{1}{L}\nabla f(x_\star)+\frac{2\theta_k}{L}h'(y_{k+1})\right\|^2-\left(\frac{1}{2L}-\frac{\theta_k^2}{L\theta_N^2}\right)\|\nabla f(x_\star)\|^2-\frac{\theta_k^2}{L\theta_N^2}\|\nabla f(x_k)\|^2\\
&-\frac{2\theta_{k+1}^2}{\theta_N^2}(f(x_{k+1})-f(x_\star))-\frac{L}{2\theta_N^2}\left\|w_{k+2}-x_\star+\frac{1}{L}\nabla f(x_\star)+\frac{2\theta_{k+1}}{L}h'(y_{k+2})\right\|^2+\left(\frac{1}{2L}-\frac{\theta_{k+1}^2}{L\theta_N^2}\right)\|\nabla f(x_\star)\|^2+\frac{\theta_{k+1}^2}{L\theta_N^2}\|\nabla f(x_{k+1})\|^2\\
&=\frac{2\theta_k^2}{\theta_N^2}(f(x_k)-f(x_\star))-\frac{2\theta_{k+1}^2}{\theta_N^2}(f(x_{k+1})-f(x_\star))+\frac{L}{2\theta_N^2}\left\|w_{k+1}-x_\star+\frac{1}{L}\nabla f(x_\star)+\frac{2\theta_k}{L}h'(y_{k+1})\right\|^2\\
&+\left(\frac{\theta_k^2-\theta_{k+1}^2}{L\theta_N^2}\right)\|\nabla f(x_\star)\|^2-\frac{\theta_k^2}{L\theta_N^2}\|\nabla f(x_k)\|^2-\frac{L}{2\theta_N^2}\left\|w_{k+1}-x_\star+\frac{1}{L}\nabla f(x_\star)-\frac{2\theta_{k+1}}{L}\nabla f(x_{k+1})\right\|^2+\frac{\theta_{k+1}^2}{L\theta_N^2}\|\nabla f(x_{k+1})\|^2\\
&=\frac{2\theta_k^2}{\theta_N^2}\left(f(x_k)-f(x_\star)-\frac{1}{2L}\|\nabla f(x_k)\|^2\right)-\frac{2\theta_{k+1}^2}{\theta_N^2}\left(f(x_{k+1})-f(x_\star)-\frac{1}{2L}\|\nabla f(x_{k+1})\|^2\right)\\
&+\frac{L}{\theta_N^2}\left\langle w_{k+1}-x_\star+\frac{1}{L}\nabla f(x_\star),\frac{2\theta_k}{L}h'(y_{k+1})+\frac{2\theta_{k+1}}{L}\nabla f(x_{k+1})\right\rangle-\frac{2\theta_{k+1}^2}{L\theta_N^2}\|\nabla f(x_{k+1})\|^2\\
&+\frac{2\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2-\frac{\theta_{k+1}}{L\theta_N^2}\|\nabla f(x_\star)\|^2\\
&=\frac{2\theta_{k+1}^2-2\theta_{k+1}}{\theta_N^2}\left(f(x_k)-f(x_{k+1})-\frac{1}{2L}\|\nabla f(x_k)\|^2-\frac{1}{2L}\|\nabla f(x_{k+1})\|^2\right)+\frac{2\theta_{k+1}}{\theta_N^2}\left(f(x_\star)-f(x_{k+1})-\frac{1}{2L}\|\nabla f(x_{k+1})\|^2\right)\\
&+\frac{2\theta_{k+1}}{\theta_N^2}\langle w_{k+1}-x_\star,\nabla f(x_{k+1})\rangle+\frac{2\theta_{k+1}}{L\theta_N^2}\langle\nabla f(x_\star),\nabla f(x_{k+1})\rangle+\frac{2\theta_k}{\theta_N^2}\left\langle w_{k+1}-x_\star+\frac{1}{L}\nabla f(x_\star),h'(y_{k+1})\right\rangle\\
&+\frac{2\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2-\frac{\theta_{k+1}}{L\theta_N^2}\|\nabla f(x_\star)\|^2 \qquad \triangleright \theta_k^2=\theta_{k+1}^2-\theta_{k+1}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{2\theta_{k+1}^2 - 2\theta_{k+1}}{\theta_N^2}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(x_k)\|^2 - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2\right)\\
&+ \frac{2\theta_{k+1}}{\theta_N^2}\left(f(x_\star) - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2 - \frac{1}{2L}\|\nabla f(x_\star)\|^2 + \frac{1}{L}\langle \nabla f(x_\star), \nabla f(x_{k+1})\rangle + \langle x_{k+1} - x_\star, \nabla f(x_{k+1})\rangle\right)\\
&+ \frac{2\theta_{k+1}}{\theta_N^2}\langle w_{k+1} - x_{k+1}, \nabla f(x_{k+1})\rangle + \frac{2\theta_k}{\theta_N^2}\left\langle w_{k+1} - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle + \frac{2\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2\\
&\geq \frac{2\theta_{k+1}^2 - 2\theta_{k+1}}{\theta_N^2}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(x_k)\|^2 - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2\right) + \frac{2\theta_{k+1}}{\theta_N^2}\langle w_{k+1} - x_{k+1}, \nabla f(x_{k+1})\rangle\\
&+ \frac{2\theta_k}{\theta_N^2}\left\langle w_{k+1} - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle + \frac{2\theta_k^2}{L\theta_N^2}\|h'(y_1)\|^2 \qquad \triangleright \text{cocoercivity of } f\\
&= \frac{2\theta_{k+1}^2 - 2\theta_{k+1}}{\theta_N^2}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(x_k)\|^2 - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2\right)\\
&+ \frac{2\theta_{k+1}(\theta_{k+1}-1)}{\theta_N^2}\left\langle x_{k+1} - x_k + \frac{1}{L}\nabla f(x_k) + \frac{1}{L}h'(y_{k+1}), \nabla f(x_{k+1})\right\rangle \qquad \triangleright\ w_{k+1} = \theta_{k+1}x_{k+1} - (\theta_{k+1}-1)z_{k+1}\\
&+ \frac{2\theta_k}{\theta_N^2}\left\langle w_{k+1} - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle + \frac{2\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2\\
&= \frac{2\theta_{k+1}^2 - 2\theta_{k+1}}{\theta_N^2}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(x_k)\|^2 - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2 + \langle x_{k+1} - x_k, \nabla f(x_{k+1})\rangle + \frac{1}{L}\langle \nabla f(x_k), \nabla f(x_{k+1})\rangle\right)\\
&+ \frac{2\theta_k}{\theta_N^2}\left\langle w_{k+1} - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle + \frac{2\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 + \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle\\
&\geq \frac{2\theta_k}{\theta_N^2}\left\langle w_{k+1} - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle + \frac{2\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 + \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle. \qquad \triangleright \text{cocoercivity of } f
\end{aligned}
$$

□

Now we begin calculating $\mathcal{H}_k - \mathcal{H}_{k+1}$ for $k \in [-1 : N-1]$. We first prove the following two lemmas.

Lemma 22 *The following equality holds.*

$$\sum_{i,j\in\{\star,1,\dots,N\}} \tau_{i,j}\left(h(y_j) - h(y_i)\right) = \sum_{i=1}^N \tau_{\star,i}\left(h(y_i) - h(x_\star)\right) + \sum_{i<j\in[1:N]} \tau_{i,j}\left(h(y_j) - h(y_i)\right) + \sum_{i=1}^{N-1} \tau_{i+1,i}\left(h(y_i) - h(y_{i+1})\right) = h(y_N) - h(x_\star)$$

Proof We prove the lemma by showing (25), (26), and (27).

$$\sum_{i=1}^N \tau_{\star,i} = 1, \tag{25}$$

$$\tau_{\star,N} + \sum_{i=1}^{N-1} \tau_{i,N} - \tau_{N,N-1} = 1, \quad \tau_{\star,1} + \tau_{2,1} - \sum_{i=2}^N \tau_{1,i} = 0, \tag{26}$$

$$\tau_{\star,i} + \sum_{j=1}^{i-1} \tau_{j,i} + \tau_{i+1,i} - \sum_{j=i+1}^N \tau_{i,j} - \tau_{i,i-1} = 0 \text{ for } i \in [2 : N-1]. \tag{27}$$

For (25),

$$\sum_{i=1}^N \tau_{\star,i} = \frac{2}{\theta_N^2 - 1}\sum_{i=1}^N \tilde{\theta}_{i-1} = \frac{2}{\theta_N^2 - 1}\cdot\frac{\theta_N^2 - 1}{2} = 1. \qquad \triangleright \text{using (24) of Lemma 20}$$

For (26),

$$
\begin{aligned}
\tau_{\star,N} + \sum_{i=1}^{N-1} \tau_{i,N} - \tau_{N,N-1} &= \frac{2\tilde{\theta}_{N-1}}{\theta_N^2 - 1} + 2\tilde{\theta}_{N-1}\sum_{i=1}^{N-1}\left(\frac{1}{\theta_N^2 - 2\theta_i^2 + \theta_i} - \frac{1}{\theta_N^2 - 2\theta_{i-1}^2 + \theta_{i-1}}\right) - \frac{\theta_{N-1} - 1}{\theta_N^2 - 2\theta_{N-1}^2 + \theta_{N-1}}\\
&= \frac{2\tilde{\theta}_{N-1}}{\theta_N^2 - 1} + 2\tilde{\theta}_{N-1}\left(\frac{1}{\theta_N^2 - 2\theta_{N-1}^2 + \theta_{N-1}} - \frac{1}{\theta_N^2 - 1}\right) - \frac{\theta_{N-1} - 1}{\theta_N^2 - 2\theta_{N-1}^2 + \theta_{N-1}} \qquad \triangleright \text{telescoping sum}\\
&= \frac{2\tilde{\theta}_{N-1} - \theta_{N-1} + 1}{\theta_N^2 - 2\theta_{N-1}^2 + \theta_{N-1}} = \frac{2\theta_{N-1} + \theta_N - 1 - \theta_{N-1} + 1}{\theta_N + \theta_{N-1}} = 1, \qquad \triangleright \text{substituting } \tilde{\theta}_{N-1} = \frac{2\theta_{N-1} + \theta_N - 1}{2}
\end{aligned}
$$

and

$$\begin{aligned}
&\tau_{\star,1}+\tau_{2,1}-\sum_{i=2}^{N}\tau_{1,i}=\frac{2}{\theta_N^2-1}+\frac{\theta_1-1}{\theta_N^2-2\theta_1^2+\theta_1}-\left(\frac{2}{\theta_N^2-2\theta_1^2+\theta_1}-\frac{2}{\theta_N^2-1}\right)\sum_{i=2}^{N}\tilde{\theta}_{i-1}\\
&=\frac{2}{\theta_N^2-1}+\frac{\theta_1-1}{\theta_N^2-2\theta_1^2+\theta_1}-\left(\frac{2}{\theta_N^2-2\theta_1^2+\theta_1}-\frac{2}{\theta_N^2-1}\right)\left(\frac{\theta_N^2-1}{2}-1\right) \qquad \triangleright \text{using (24) of Lemma 20}\\
&=\frac{1}{\theta_N^2-1}\left(2+(\theta_N^2-1)-2\right)+\frac{1}{\theta_N^2-2\theta_1^2+\theta_1}\left(\theta_1-1-(\theta_N^2-1)+2\right)\\
&=1+\frac{1}{\theta_N^2-\theta_1-2}\left(-\theta_N^2+\theta_1+2\right)=0. \qquad \triangleright \text{using (21) of Lemma 19}
\end{aligned}$$

Lastly, for (27) when $i\in[2:N-1]$,

$$\begin{aligned}
&\tau_{\star,i}+\sum_{j=1}^{i-1}\tau_{j,i}+\tau_{i+1,i}-\sum_{j=i+1}^{N}\tau_{i,j}-\tau_{i,i-1}\\
&=\frac{2\tilde{\theta}_{i-1}}{\theta_N^2-1}+2\tilde{\theta}_{i-1}\sum_{j=1}^{i-1}\left(\frac{1}{\theta_N^2-2\theta_j^2+\theta_j}-\frac{1}{\theta_N^2-2\theta_{j-1}^2+\theta_{j-1}}\right)+\frac{\theta_i-1}{\theta_N^2-2\theta_i^2+\theta_i}\\
&-\left(\frac{2}{\theta_N^2-2\theta_i^2+\theta_i}-\frac{2}{\theta_N^2-2\theta_{i-1}^2+\theta_{i-1}}\right)\sum_{j=i+1}^{N}\tilde{\theta}_{j-1}-\frac{\theta_{i-1}-1}{\theta_N^2-2\theta_{i-1}^2+\theta_{i-1}}\\
&=\frac{2\tilde{\theta}_{i-1}}{\theta_N^2-1}+2\tilde{\theta}_{i-1}\left(\frac{1}{\theta_N^2-2\theta_{i-1}^2+\theta_{i-1}}-\frac{1}{\theta_N^2-1}\right)+\frac{\theta_i-1}{\theta_N^2-2\theta_i^2+\theta_i} \qquad \triangleright \text{telescoping sum}\\
&-\left(\frac{2}{\theta_N^2-2\theta_i^2+\theta_i}-\frac{2}{\theta_N^2-2\theta_{i-1}^2+\theta_{i-1}}\right)\left(\frac{\theta_N^2-1}{2}-\tilde{\theta}_{i-1}^2\right) \qquad \triangleright \text{using (23) of Lemma 19, (24) of Lemma 19}\\
&-\frac{\theta_{i-1}-1}{\theta_N^2-2\theta_{i-1}^2+\theta_{i-1}}\\
&=\frac{1}{\theta_N^2-2\theta_i^2+\theta_i}\left(\theta_i-1-(\theta_N^2-1)+2\tilde{\theta}_{i-1}^2\right)+\frac{1}{\theta_N^2-2\theta_{i-1}^2+\theta_{i-1}}\left(2\tilde{\theta}_{i-1}+(\theta_N^2-1)-2\tilde{\theta}_{i-1}^2-\theta_{i-1}+1\right)\\
&=\frac{1}{\theta_N^2-2\theta_i^2+\theta_i}\left(-\theta_N^2+\theta_i+2\theta_i^2-2\theta_i\right) \qquad \triangleright \text{using (21) of Lemma 19 and } \tilde{\theta}_{i-1}=\theta_{i-1}\\
&+\frac{1}{\theta_N^2-2\theta_{i-1}^2+\theta_{i-1}}\left(\theta_N^2-2\theta_{i-1}^2+\theta_{i-1}\right) \qquad \triangleright \text{using } \tilde{\theta}_{i-1}=\theta_{i-1}\\
&=-1+1=0.
\end{aligned}$$

□

Lemma 23 *The following holds.*

$$\begin{aligned}
\mathcal{H}_{N-1}-\mathcal{H}_N=&\sum_{i=1}^{N-1}\tau_{i,N}\left(h(y_i)-h(y_N)\right)+\tau_{\star,N}\left(h(x_\star)-h(y_N)\right)+\tau_{N,N-1}\left(h(y_N)-h(y_{N-1})\right)\\
&+\frac{2\tilde{\theta}_{N-1}}{\theta_N^2(\theta_N^2-1)}\left\langle x_0-x_\star-\frac{\theta_N^2-1}{L}\nabla f(x_\star),h'(y_N)\right\rangle-\frac{\tilde{\theta}_{N-1}+\theta_{N-2}^2}{L\theta_N^2}\|h'(y_N)\|^2+\frac{2\theta_{N-2}^2}{L\theta_N^2}\langle h'(y_{N-1}),h'(y_N)+\nabla f(x_{N-1})\rangle.
\end{aligned}$$

For $k\in[0:N-2]$,

$$\begin{aligned}
\mathcal{H}_k-\mathcal{H}_{k+1}=&\sum_{i=1}^{k}\tau_{i,k+1}\left(h(y_i)-h(y_{k+1})\right)+\tau_{\star,k+1}\left(h(x_\star)-h(y_{k+1})\right)+\tau_{k+1,k}\left(h(y_{k+1})-h(y_k)\right)\\
&+\frac{2\theta_k}{\theta_N^2(\theta_N^2-1)}\left\langle x_0-x_\star-\frac{\theta_N^2-1}{L}\nabla f(x_\star),h'(y_{k+1})\right\rangle-\frac{2\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2+\frac{2\theta_{k-1}^2}{L\theta_N^2}\langle h'(y_k),h'(y_{k+1})+\nabla f(x_k)\rangle\\
&-\frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}),\nabla f(x_{k+1})\rangle.
\end{aligned}$$

Proof For $k = N-1$,

$$
\begin{aligned}
\mathcal{H}_{N-1} - \mathcal{H}_N &= \sum_{i,j\in\{\star,1,\dots,N-1\}} \tau_{i,j}(h(y_j)-h(y_i)) - h(y_N) + h(x_\star)\\
&+ \frac{L}{2\theta_N^2(\theta_N^2-1)}\left[\left\|x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star) - \sum_{i=0}^{N-2}\frac{2\theta_i}{L}h'(y_{i+1})\right\|^2 - \left\|x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star) - \sum_{i=0}^{N-1}\frac{2\tilde{\theta}_i}{L}h'(y_{i+1})\right\|^2\right]\\
&+ \sum_{i\neq j, i,j\in[1:N-1]} \frac{\theta_{i-1}\theta_{j-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)-h'(y_j)\|^2 + \sum_{i=1}^{N-2}\frac{\theta_{i-1}^2}{L\theta_N^2}\|h'(y_i)-h'(y_{i+1})\|^2\\
&- \sum_{i\neq j, i,j\in[1:N]} \frac{\tilde{\theta}_{i-1}\tilde{\theta}_{j-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)-h'(y_j)\|^2 - \sum_{i=1}^{N-1}\frac{\tilde{\theta}_{i-1}^2}{L\theta_N^2}\|h'(y_i)-h'(y_{i+1})\|^2\\
&+ \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle\nabla f(x_{N-1}), h'(y_{N-1})\rangle + \sum_{i=1}^{N-1}\frac{2\tilde{\theta}_{N-1}\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 + \frac{\theta_{N-2}^2}{L\theta_N^2}\|h'(y_{N-1})\|^2\\
&= \sum_{i,j\in\{\star,1,\dots,N-1\}} \tau_{i,j}(h(y_j)-h(y_i)) - \sum_{i,j\in\{\star,1,\dots,N\}} \tau_{i,j}(h(y_j)-h(y_i))\\
&+ \frac{L}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star) - \sum_{i=0}^{N-2}\frac{2\theta_i}{L}h'(h_{i+1}), \frac{2\tilde{\theta}_{N-1}}{L}h'(y_N)\right\rangle - \frac{2\tilde{\theta}_{N-1}^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_N)\|^2\\
&- \sum_{i=1}^{N-1}\frac{2\tilde{\theta}_{N-1}\tilde{\theta}_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_N)-h'(y_i)\|^2 - \frac{\tilde{\theta}_{N-2}^2}{L\theta_N^2}\|h'(y_N)-h'(y_{N-1})\|^2\\
&+ \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle\nabla f(x_{N-1}), h'(y_{N-1})\rangle + \sum_{i=1}^{N-1}\frac{2\tilde{\theta}_{N-1}\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 + \frac{\theta_{N-2}^2}{L\theta_N^2}\|h'(y_{N-1})\|^2\\
&= \sum_{i=1}^{N-1}\tau_{i,N}\left(h(y_i)-h(y_N)\right) + \tau_{\star,N}\left(h(x_\star)-h(y_N)\right) + \tau_{N,N-1}\left(h(y_N)-h(y_{N-1})\right)\\
&+ \frac{L}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), \frac{2\tilde{\theta}_{N-1}}{L}h'(y_N)\right\rangle - \sum_{i=1}^{N-1}\frac{4\theta_{i-1}\tilde{\theta}_{N-1}}{L\theta_N^2(\theta_N^2-1)}\langle h'(y_i), h'(y_N)\rangle - \frac{2\tilde{\theta}_{N-1}^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_N)\|^2\\
&- \sum_{i=1}^{N-1}\frac{2\tilde{\theta}_{N-1}\tilde{\theta}_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_N)\|^2 + \sum_{i=1}^{N-1}\frac{4\theta_{i-1}\tilde{\theta}_{N-1}}{L\theta_N^2(\theta_N^2-1)}\langle h'(y_i), h'(y_N)\rangle - \sum_{i=1}^{N-1}\frac{2\tilde{\theta}_{N-1}\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2\\
&- \frac{\tilde{\theta}_{N-2}^2}{L\theta_N^2}\|h'(y_N)-h'(y_{N-1})\|^2 + \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle\nabla f(x_{N-1}), h'(y_{N-1})\rangle + \sum_{i=1}^{N-1}\frac{2\tilde{\theta}_{N-1}\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 + \frac{\theta_{N-2}^2}{L\theta_N^2}\|h'(y_{N-1})\|^2\\
&= \sum_{i=1}^{N-1}\tau_{i,N}\left(h(y_i)-h(y_N)\right) + \tau_{\star,N}\left(h(x_\star)-h(y_N)\right) + \tau_{N,N-1}\left(h(y_N)-h(y_{N-1})\right)\\
&+ \frac{L}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), \frac{2\tilde{\theta}_{N-1}}{L}h'(y_N)\right\rangle - \frac{2\tilde{\theta}_{N-1}^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_N)\|^2\\
&- \sum_{i=1}^{N-1}\frac{2\tilde{\theta}_{N-1}\tilde{\theta}_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_N)\|^2 - \frac{\tilde{\theta}_{N-2}^2}{L\theta_N^2}\|h'(y_N)-h'(y_{N-1})\|^2 + \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle\nabla f(x_{N-1}), h'(y_{N-1})\rangle + \frac{\theta_{N-2}^2}{L\theta_N^2}\|h'(y_{N-1})\|^2\\
&= \sum_{i=1}^{N-1}\tau_{i,N}\left(h(y_i)-h(y_N)\right) + \tau_{\star,N}\left(h(x_\star)-h(y_N)\right) + \tau_{N,N-1}\left(h(y_N)-h(y_{N-1})\right)\\
&+ \frac{2\tilde{\theta}_{N-1}}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_N)\right\rangle - \frac{2\tilde{\theta}_{N-1}^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_N)\|^2\\
&- \frac{\tilde{\theta}_{N-1}}{L\theta_N^2(\theta_N^2-1)}\left(\theta_N^2 - 1 - 2\tilde{\theta}_{N-1}\right)\|h'(y_N)\|^2 - \frac{\tilde{\theta}_{N-2}^2}{L\theta_N^2}\|h'(y_N)\|^2 + \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle h'(y_N), h'(y_{N-1})\rangle - \frac{\theta_{N-2}^2}{L\theta_N^2}\|h'(y_{N-1})\|^2\\
&+ \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle\nabla f(x_{N-1}), h'(y_{N-1})\rangle + \frac{\theta_{N-2}^2}{L\theta_N^2}\|h'(y_{N-1})\|^2 \qquad \triangleright\ \sum_{i=1}^{N-1}\tilde{\theta}_{i-1} = \sum_{i=0}^{N-1}\tilde{\theta}_i - \tilde{\theta}_{N-1} = \frac{\theta_N^2-1}{2} - \tilde{\theta}_{N-1}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{N-1} \tau_{i,N} \left(h(y_i) - h(y_N)\right) + \tau_{\star,N} \left(h(x_\star) - h(y_N)\right) + \tau_{N,N-1} \left(h(y_N) - h(y_{N-1})\right) \\
&+ \frac{2\tilde{\theta}_{N-1}}{\theta_N^2(\theta_N^2-1)} \left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_N) \right\rangle - \frac{\tilde{\theta}_{N-1}}{L\theta_N^2}\|h'(y_N)\|^2 - \frac{\tilde{\theta}_{N-2}^2}{L\theta_N^2}\|h'(y_N)\|^2 \\
&+ \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle h'(y_N), h'(y_{N-1})\rangle + \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle \nabla f(x_{N-1}), h'(y_{N-1})\rangle \\
&= \sum_{i=1}^{N-1} \tau_{i,N} \left(h(y_i) - h(y_N)\right) + \tau_{\star,N} \left(h(x_\star) - h(y_N)\right) + \tau_{N,N-1} \left(h(y_N) - h(y_{N-1})\right) \\
&+ \frac{2\tilde{\theta}_{N-1}}{\theta_N^2(\theta_N^2-1)} \left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_N) \right\rangle - \frac{\tilde{\theta}_{N-1}+\theta_{N-2}^2}{L\theta_N^2}\|h'(y_N)\|^2 + \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle h'(y_{N-1}), h'(y_N) + \nabla f(x_{N-1})\rangle
\end{aligned}
$$

For $k \in [-1 : N-2]$,

$$
\begin{aligned}
\mathcal{H}_k - \mathcal{H}_{k+1} &= \sum_{i,j\in\{\star,1,\ldots,k\}} \tau_{i,j} \left(h(y_j) - h(y_i)\right) - \sum_{i,j\in\{\star,1,\ldots,k+1\}} \tau_{i,j} \left(h(y_j) - h(y_i)\right) \\
&+ \frac{L}{2\theta_N^2(\theta_N^2-1)} \left[\left\| x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star) - \sum_{i=0}^{k-1} \frac{2\theta_i}{L} h'(y_{i+1}) \right\|^2 - \left\| x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star) - \sum_{i=0}^{k} \frac{2\theta_i}{L} h'(y_{i+1}) \right\|^2 \right] \\
&+ \sum_{i\neq j, i,j\in[1:k]} \frac{\theta_{i-1}\theta_{j-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i) - h'(y_j)\|^2 + \sum_{i=1}^{k-1} \frac{\theta_{i-1}^2}{L\theta_N^2}\|h'(y_i) - h'(y_{i+1})\|^2 \\
&+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle \nabla f(x_k), h'(y_k)\rangle + \sum_{i=1}^{k}\sum_{\ell=k}^{N-1} \frac{2\tilde{\theta}_\ell\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 + \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)\|^2 \\
&- \sum_{i\neq j, i,j\in[1:k+1]} \frac{\theta_{i-1}\theta_{j-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i) - h'(y_j)\|^2 - \sum_{i=1}^{k} \frac{\theta_{i-1}^2}{L\theta_N^2}\|h'(y_i) - h'(y_{i+1})\|^2 \\
&- \frac{2\theta_k^2}{L\theta_N^2}\langle \nabla f(x_{k+1}), h'(y_{k+1})\rangle - \sum_{i=1}^{k+1}\sum_{\ell=k+1}^{N-1} \frac{2\tilde{\theta}_\ell\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 - \frac{\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 \\
&= \sum_{i=1}^{k} \tau_{i,k+1} \left(h(y_i) - h(y_{k+1})\right) + \tau_{\star,k+1} \left(h(x_\star) - h(y_{k+1})\right) + \tau_{k+1,k} \left(h(y_{k+1}) - h(y_k)\right) \\
&+ \frac{L}{2\theta_N^2(\theta_N^2-1)} \left[\left\| x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star) - \sum_{i=0}^{k} \frac{2\theta_i}{L} h'(y_{i+1}) \right\|^2 - \left\| x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star) - \sum_{i=0}^{k-1} \frac{2\theta_i}{L} h'(y_{i+1}) \right\|^2 \right] \\
&- \sum_{i=1}^{k} \frac{2\theta_k\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1}) - h'(y_i)\|^2 - \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k) - h'(y_{k+1})\|^2 \\
&+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle \nabla f(x_k), h'(y_k)\rangle - \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle + \sum_{i=1}^{k}\sum_{\ell=k}^{N-1} \frac{2\tilde{\theta}_\ell\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 - \sum_{i=1}^{k+1}\sum_{\ell=k+1}^{N-1} \frac{2\tilde{\theta}_\ell\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 \\
&+ \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)\|^2 - \frac{\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 \\
&= \sum_{i=1}^{k} \tau_{i,k+1} \left(h(y_i) - h(y_{k+1})\right) + \tau_{\star,k+1} \left(h(x_\star) - h(y_{k+1})\right) + \tau_{k+1,k} \left(h(y_{k+1}) - h(y_k)\right) \\
&+ \frac{2\theta_k}{\theta_N^2(\theta_N^2-1)} \left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star) - \sum_{i=0}^{k-1} \frac{2\theta_i}{L} h'(y_{i+1}), h'(y_{k+1}) \right\rangle - \frac{2\theta_k^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2
\end{aligned}
$$

$$
\begin{aligned}
&- \sum_{i=1}^{k} \frac{2\theta_k\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})-h'(y_i)\|^2 - \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)-h'(y_{k+1})\|^2 \\
&+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle \nabla f(x_k), h'(y_k)\rangle - \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle + \sum_{i=1}^{k}\sum_{\ell=k}^{N-1} \frac{2\tilde{\theta}_\ell\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 - \sum_{i=1}^{k+1}\sum_{\ell=k+1}^{N-1} \frac{2\tilde{\theta}_\ell\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 \\
&+ \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)\|^2 - \frac{\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 \\
&= \sum_{i=1}^{k} \tau_{i,k+1}\left(h(y_i)-h(y_{k+1})\right) + \tau_{\star,k+1}\left(h(x_\star)-h(y_{k+1})\right) + \tau_{k+1,k}\left(h(y_{k+1})-h(y_k)\right) \\
&+ \frac{2\theta_k}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle - \sum_{i=0}^{k-1}\frac{4\theta_k\theta_i}{L\theta_N^2(\theta_N^2-1)}\langle h'(y_{i+1}), h'(y_{k+1})\rangle - \frac{2\theta_k^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 \\
&- \sum_{i=1}^{k}\frac{2\theta_k\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 - \sum_{i=1}^{k}\frac{2\theta_k\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 + \sum_{i=1}^{k}\frac{4\theta_k\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\langle h'(y_i), h'(y_{k+1})\rangle - \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)-h'(y_{k+1})\|^2 \\
&+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle \nabla f(x_k), h'(y_k)\rangle - \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle + \sum_{i=1}^{k}\sum_{\ell=k}^{N-1} \frac{2\tilde{\theta}_\ell\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 - \sum_{i=1}^{k}\sum_{\ell=k+1}^{N-1} \frac{2\tilde{\theta}_\ell\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 \\
&- \sum_{\ell=k+1}^{N-1}\frac{2\tilde{\theta}_\ell\theta_k}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 + \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)\|^2 - \frac{\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 \\
&= \sum_{i=1}^{k} \tau_{i,k+1}\left(h(y_i)-h(y_{k+1})\right) + \tau_{\star,k+1}\left(h(x_\star)-h(y_{k+1})\right) + \tau_{k+1,k}\left(h(y_{k+1})-h(y_k)\right) \\
&+ \frac{2\theta_k}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle - \frac{2\theta_k^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 \\
&- \sum_{i=0}^{k-1}\frac{2\theta_k\tilde{\theta}_i}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 - \sum_{i=1}^{k}\frac{2\theta_k\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 - \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)-h'(y_{k+1})\|^2 \\
&+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle \nabla f(x_k), h'(y_k)\rangle - \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle + \sum_{i=1}^{k}\frac{2\theta_k\theta_{i-1}}{L\theta_N^2(\theta_N^2-1)}\|h'(y_i)\|^2 - \sum_{\ell=k+1}^{N-1}\frac{2\tilde{\theta}_\ell\theta_k}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 \\
&+ \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)\|^2 - \frac{\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 \\
&= \sum_{i=1}^{k} \tau_{i,k+1}\left(h(y_i)-h(y_{k+1})\right) + \tau_{\star,k+1}\left(h(x_\star)-h(y_{k+1})\right) + \tau_{k+1,k}\left(h(y_{k+1})-h(y_k)\right) \\
&+ \frac{2\theta_k}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle - \frac{2\theta_k^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 \\
&- \frac{2\theta_k\theta_{k-1}^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 - \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)-h'(y_{k+1})\|^2 \\
&+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle \nabla f(x_k), h'(y_k)\rangle - \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle - \frac{(\theta_N^2-1-2\theta_k^2)\theta_k}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 \qquad \triangleright \sum_{\ell=k+1}^{N-1}\tilde{\theta}_\ell = \sum_{\ell=0}^{N-1}\tilde{\theta}_\ell - \sum_{\ell=0}^{k}\tilde{\theta}_\ell = \frac{\theta_N^2-1}{2} - \theta_k^2 \\
&+ \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)\|^2 - \frac{\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 \\
&= \sum_{i=1}^{k} \tau_{i,k+1}\left(h(y_i)-h(y_{k+1})\right) + \tau_{\star,k+1}\left(h(x_\star)-h(y_{k+1})\right) + \tau_{k+1,k}\left(h(y_{k+1})-h(y_k)\right) \\
&+ \frac{2\theta_k}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle - \frac{2\theta_k^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 \\
&- \frac{2\theta_k\theta_{k-1}^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 - \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)\|^2 - \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 + \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle h'(y_k), h'(y_{k+1})\rangle \\
&+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle \nabla f(x_k), h'(y_k)\rangle - \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle - \frac{\theta_k}{L\theta_N^2}\|h'(y_{k+1})\|^2 + \frac{2\theta_k^3}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 \\
&+ \frac{\theta_{k-1}^2}{L\theta_N^2}\|h'(y_k)\|^2 - \frac{\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{k} \tau_{i,k+1} \left(h(y_i) - h(y_{k+1})\right) + \tau_{\star,k+1} \left(h(x_\star) - h(y_{k+1})\right) + \tau_{k+1,k} \left(h(y_{k+1}) - h(y_k)\right) \\
&+ \frac{2\theta_k}{\theta_N^2(\theta_N^2-1)} \left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_{k+1}) \right\rangle - \frac{2\theta_k^2}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 \\
&+ \frac{2\theta_k(\theta_k^2-\theta_{k-1}^2)}{L\theta_N^2(\theta_N^2-1)}\|h'(y_{k+1})\|^2 - \frac{\theta_{k-1}^2+\theta_k+\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 + \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle h'(y_k), h'(y_{k+1})\rangle \\
&+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle \nabla f(x_k), h'(y_k)\rangle - \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle \\
&= \sum_{i=1}^{k} \tau_{i,k+1} \left(h(y_i) - h(y_{k+1})\right) + \tau_{\star,k+1} \left(h(x_\star) - h(y_{k+1})\right) + \tau_{k+1,k} \left(h(y_{k+1}) - h(y_k)\right) \\
&+ \frac{2\theta_k}{\theta_N^2(\theta_N^2-1)} \left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_{k+1}) \right\rangle - \frac{2\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 + \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle h'(y_k), h'(y_{k+1})\rangle \\
&+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle \nabla f(x_k), h'(y_k)\rangle - \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle
\end{aligned}
$$

□

Now we build our last part of the proof.

Lemma 24 *The following holds.*

$$\tilde{\theta}_{N-1}(4\theta_{N-1} - 2\tilde{\theta}_{N-1}) = \tilde{\theta}_{N-1} + \theta_{N-2}^2$$

Proof

$$
\begin{aligned}
&\tilde{\theta}_{N-1}(4\theta_{N-1} - 2\tilde{\theta}_{N-1}) - \tilde{\theta}_{N-1} - \theta_{N-2}^2 = 4\theta_{N-1}\tilde{\theta}_{N-1} - 2\tilde{\theta}_{N-1}^2 - \tilde{\theta}_{N-1} - \theta_{N-2}^2 \\
&= \frac{1}{2}\left(8\theta_{N-1}\tilde{\theta}_{N-1} - 4\tilde{\theta}_{N-1}^2 - 2\tilde{\theta}_{N-1} - 2\theta_{N-2}^2\right) \\
&= \frac{1}{2}\left(4\theta_{N-1}(2\theta_{N-1}+\theta_N-1) - (2\theta_{N-1}+\theta_N-1)^2 - 2\tilde{\theta}_{N-1} - 2\theta_{N-2}^2\right) \\
&= \frac{1}{2}\left(4\theta_{N-1}^2 + 2\theta_N - \theta_N^2 - 1 - 2\tilde{\theta}_{N-1} - 2\theta_{N-2}^2\right) \\
&= \frac{1}{2}\left(2\theta_{N-1}^2 + 2\theta_{N-1} + 2\theta_N - \theta_N^2 - 1 - 2\tilde{\theta}_{N-1}\right) = \frac{1}{2}\left(2\theta_{N-1} + \theta_N - 1 - \tilde{\theta}_{N-1}\right) = 0.
\end{aligned}
$$

□

Lemma 25 *The following holds for fixed $j \in [1:N]$.*

$$\tau_{\star,j}\langle h'(y_j), x_\star - y_j\rangle + \sum_{i=1}^{j-1}\tau_{i,j}\langle h'(y_j), y_i - y_j\rangle = \left\langle h'(y_j)\ ,\ \tau_{\star,j}(x_\star - x_0) + \sum_{i=0}^{j-1}\frac{4\theta_i\tilde{\theta}_{j-1}}{L\theta_N^2}\left(\nabla f(x_i) + h'(y_{i+1})\right)\right\rangle$$

Proof Using Lemma 17, we have:

$$
\begin{aligned}
&\tau_{\star,j}\langle h'(y_j), x_\star - y_j\rangle + \sum_{i=1}^{j-1}\tau_{i,j}\langle h'(y_j), y_i - y_j\rangle = \tau_{\star,j}\left\langle h'(y_j), x_\star - x_0 + \sum_{i=0}^{j-1}\frac{\gamma_i}{L}(\nabla f(x_i) + h'(y_{i+1}))\right\rangle + \sum_{i=1}^{j-1}\tau_{i,j}\langle h'(y_j), y_i - y_j\rangle \\
&= \tau_{\star,j}\left\langle h'(y_j), x_\star - x_0 + \sum_{i=0}^{j-1}\frac{\gamma_i}{L}(\nabla f(x_i) + h'(y_{i+1}))\right\rangle + \sum_{i=1}^{j-1}\tau_{i,j}\left\langle h'(y_j), \sum_{k=i}^{j-1}\frac{\gamma_k}{L}(\nabla f(x_k) + h'(y_{k+1}))\right\rangle \\
&= \langle h'(y_j), \tau_{\star,j}(x_\star - x_0)\rangle + \left\langle h'(y_j), \sum_{i=0}^{j-1}\tau_{\star,j}\frac{\gamma_i}{L}(\nabla f(x_i) + h'(y_{i+1}))\right\rangle + \left\langle h'(y_j), \sum_{i=1}^{j-1}\tau_{i,j}\sum_{k=i}^{j-1}\frac{\gamma_k}{L}(\nabla f(x_k) + h'(y_{k+1}))\right\rangle \\
&= \langle h'(y_j), \tau_{\star,j}(x_\star - x_0)\rangle + \left\langle h'(y_j), \sum_{i=0}^{j-1}\left(\tau_{\star,j} + \sum_{k=1}^{i}\tau_{k,j}\right)\frac{\gamma_i}{L}(\nabla f(x_i) + h'(y_{i+1}))\right\rangle \\
&= \langle h'(y_j), \tau_{\star,j}(x_\star - x_0)\rangle + \left\langle h'(y_j), \sum_{i=0}^{j-1}\frac{2\tilde{\theta}_{j-1}}{\theta_N^2 - 2\theta_i^2 + \theta_i}\frac{\gamma_i}{L}(\nabla f(x_i) + h'(y_{i+1}))\right\rangle \qquad \triangleright \text{telescopic sum of } \tau_{i,j} \\
&= \langle h'(y_j), \tau_{\star,j}(x_\star - x_0)\rangle + \left\langle h'(y_j), \sum_{i=0}^{j-1}\frac{4\theta_i\tilde{\theta}_{j-1}}{L\theta_N^2}(\nabla f(x_i) + h'(y_{i+1}))\right\rangle
\end{aligned}
$$

□

Proof of Theorem 1 For $k = N-1$,

$$\begin{aligned}
\mathcal{U}_{N-1} - \mathcal{U}_N &= \mathcal{F}_{N-1} - \mathcal{F}_N + \mathcal{H}_{N-1} - \mathcal{H}_N \\
&\geq \frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle w_N - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_N)\right\rangle + \cancel{\frac{\tilde{\theta}_{N-1}(4\theta_{N-1} - 2\tilde{\theta}_{N-1})}{L\theta_N^2}\|h'(y_N)\|^2} \qquad \triangleright \text{ using Lemma 24} \\
&+ \sum_{i=1}^{N-1} \tau_{i,N}\left(h(y_i) - h(y_N)\right) + \tau_{\star,N}\left(h(x_\star) - h(y_N)\right) + \tau_{N,N-1}\left(h(y_N) - h(y_{N-1})\right) \\
&+ \frac{2\tilde{\theta}_{N-1}}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_N)\right\rangle - \cancel{\frac{\tilde{\theta}_{N-1} + \theta_{N-2}^2}{L\theta_N^2}\|h'(y_N)\|^2} + \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle h'(y_{N-1}), h'(y_N) + \nabla f(x_{N-1})\rangle \\
&= \frac{2\tilde{\theta}_{N-1}}{\theta_N^2}\left\langle x_0 - x_\star + \frac{1}{L}\nabla f(x_\star) - \sum_{i=0}^{N-1}\frac{2\theta_i}{L}\nabla f(x_i) - \sum_{i=0}^{N-1}\frac{2\theta_i}{L}h'(y_{i+1}), h'(y_N)\right\rangle \qquad \triangleright\ w_{i+1} = w_i - \frac{2\theta_i}{L}\nabla f(x_i) - \frac{2\theta_i}{L}h'(y_{i+1}) \\
&+ \sum_{i=1}^{N-1} \tau_{i,N}\left(h(y_i) - h(y_N)\right) + \tau_{\star,N}\left(h(x_\star) - h(y_N)\right) + \tau_{N,N-1}\left(h(y_N) - h(y_{N-1})\right) \\
&+ \frac{2\tilde{\theta}_{N-1}}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_N)\right\rangle + \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle h'(y_{N-1}), h'(y_N) + \nabla f(x_{N-1})\rangle \\
&= \sum_{i=1}^{N-1} \tau_{i,N}\left(h(y_i) - h(y_N)\right) + \tau_{\star,N}\left(h(x_\star) - h(y_N)\right) + \tau_{N,N-1}\left(h(y_N) - h(y_{N-1})\right) \\
&+ \frac{2\tilde{\theta}_{N-1}}{\theta_N^2-1}\left\langle x_0 - x_\star, h'(y_N)\right\rangle - \sum_{i=0}^{N-1}\frac{4\tilde{\theta}_{N-1}\theta_i}{L\theta_N^2}\left\langle \nabla f(x_i) + h'(y_{i+1}), h'(y_N)\right\rangle \\
&+ \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle h'(y_{N-1}), h'(y_N) + \nabla f(x_{N-1})\rangle \\
&= \sum_{i=1}^{N-1} \tau_{i,N}\left(h(y_i) - h(y_N)\right) + \tau_{\star,N}\left(h(x_\star) - h(y_N)\right) + \tau_{\star,N}\left\langle x_0 - x_\star, h'(y_N)\right\rangle - \sum_{i=0}^{N-1}\frac{4\tilde{\theta}_{N-1}\theta_i}{L\theta_N^2}\left\langle \nabla f(x_i) + h'(y_{i+1}), h'(y_N)\right\rangle \\
&+ \frac{2\theta_{N-2}^2}{L\theta_N^2}\langle h'(y_{N-1}), h'(y_N) + \nabla f(x_{N-1})\rangle + \tau_{N,N-1}\left(h(y_N) - h(y_{N-1})\right) \\
&= \sum_{i=1}^{N-1} \tau_{i,N}\left(h(y_i) - h(y_N)\right) + \tau_{\star,N}\left(h(x_\star) - h(y_N)\right) - \tau_{\star,N}\langle h'(y_N), x_\star - y_N\rangle - \sum_{i=1}^{N-1}\tau_{i,N}\langle h'(y_N), y_i - y_N\rangle \qquad \triangleright \text{ using Lemma 25} \\
&+ \tau_{N,N-1}\left(\left\langle h'(y_{N-1}), \frac{\gamma_{N-1}}{L}h'(y_N) + \frac{\gamma_{N-1}}{L}\nabla f(x_{N-1})\right\rangle + h(y_N) - h(y_{N-1})\right) \\
&= \sum_{i=1}^{N-1} \tau_{i,N}\left(h(y_i) - h(y_N) - \langle h'(y_N), y_i - y_N\rangle\right) + \tau_{\star,N}\left(h(x_\star) - h(y_N) - \langle h'(y_N), x_\star - y_N\rangle\right) \\
&+ \tau_{N,N-1}\left(\langle h'(y_{N-1}), y_{N-1} - y_N\rangle + h(y_N) - h(y_{N-1})\right) \geq 0 \qquad \triangleright \text{ convexity of } h \text{ and by Lemma 17}
\end{aligned}$$

For $k \in [0 : N-2]$,

$$\begin{aligned}
\mathcal{U}_k - \mathcal{U}_{k+1} &= \mathcal{F}_k - \mathcal{F}_{k+1} + \mathcal{H}_k - \mathcal{H}_{k+1} \\
&\geq \frac{2\theta_k}{\theta_N^2}\left\langle w_{k+1} - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle + \cancel{\frac{2\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2} + \cancel{\frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle} \\
&+ \sum_{i=1}^{k} \tau_{i,k+1}\left(h(y_i) - h(y_{k+1})\right) + \tau_{\star,k+1}\left(h(x_\star) - h(y_{k+1})\right) + \tau_{k+1,k}\left(h(y_{k+1}) - h(y_k)\right) \\
&+ \frac{2\theta_k}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle - \cancel{\frac{2\theta_k^2}{L\theta_N^2}\|h'(y_{k+1})\|^2} + \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle h'(y_k), h'(y_{k+1}) + \nabla f(x_k)\rangle \\
&\cancel{- \frac{2\theta_k^2}{L\theta_N^2}\langle h'(y_{k+1}), \nabla f(x_{k+1})\rangle}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2\theta_k}{\theta_N^2}\left\langle x_0 - x_\star + \frac{1}{L}\nabla f(x_\star) - \sum_{i=0}^{k}\frac{2\theta_i}{L}\nabla f(x_i) - \sum_{i=0}^{k}\frac{2\theta_i}{L}h'(y_{i+1}), h'(y_{k+1})\right\rangle \qquad \triangleright\ w_{i+1} = w_i - \frac{2\theta_i}{L}\nabla f(x_i) - \frac{2\theta_i}{L}h'(y_{i+1}) \\
&+ \sum_{i=1}^{k}\tau_{i,k+1}\left(h(y_i) - h(y_{k+1})\right) + \tau_{\star,k+1}\left(h(x_\star) - h(y_{k+1})\right) + \tau_{k+1,k}\left(h(y_{k+1}) - h(y_k)\right) \\
&+ \frac{2\theta_k}{\theta_N^2(\theta_N^2-1)}\left\langle x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star), h'(y_{k+1})\right\rangle + \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle h'(y_k), h'(y_{k+1}) + \nabla f(x_k)\rangle \\
&= \sum_{i=1}^{k}\tau_{i,k+1}\left(h(y_i) - h(y_{k+1})\right) + \tau_{\star,k+1}\left(h(x_\star) - h(y_{k+1})\right) + \tau_{k+1,k}\left(h(y_{k+1}) - h(y_k)\right) \\
&+ \frac{2\theta_k}{\theta_N^2-1}\left\langle x_0 - x_\star, h'(y_{k+1})\right\rangle - \sum_{i=0}^{k}\frac{4\theta_k\theta_i}{L\theta_N^2}\left\langle \nabla f(x_i) + h'(y_{i+1}), h'(y_{k+1})\right\rangle + \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle h'(y_k), h'(y_{k+1}) + \nabla f(x_k)\rangle \\
&= \sum_{i=1}^{k}\tau_{i,k+1}\left(h(y_i) - h(y_{k+1})\right) + \tau_{\star,k+1}\left(h(x_\star) - h(y_{k+1})\right) + \tau_{\star,k+1}\left\langle x_0 - x_\star, h'(y_{k+1})\right\rangle - \sum_{i=0}^{k}\frac{4\theta_k\theta_i}{L\theta_N^2}\left\langle \nabla f(x_i) + h'(y_{i+1}), h'(y_{k+1})\right\rangle \\
&+ \frac{2\theta_{k-1}^2}{L\theta_N^2}\langle h'(y_k), h'(y_{k+1}) + \nabla f(x_k)\rangle + \tau_{k+1,k}\left(h(y_{k+1}) - h(y_k)\right) \\
&= \sum_{i=1}^{k}\tau_{i,k+1}\left(h(y_i) - h(y_{k+1})\right) + \tau_{\star,k+1}\left(h(x_\star) - h(y_{k+1})\right) - \tau_{\star,k+1}\langle h'(y_{k+1}), x_\star - y_{k+1}\rangle - \sum_{i=1}^{k}\tau_{i,k+1}\langle h'(y_{k+1}), y_i - y_{k+1}\rangle \\
&+ \tau_{k+1,k}\left(\left\langle h'(y_k), \frac{\gamma_k}{L}h'(y_{k+1}) + \frac{\gamma_k}{L}\nabla f(x_k)\right\rangle + h(y_{k+1}) - h(y_k)\right) \qquad \triangleright\ \text{using Lemma 25} \\
&= \sum_{i=1}^{k}\tau_{i,k+1}\left(h(y_i) - h(y_{k+1}) - \langle h'(y_{k+1}), y_i - y_{k+1}\rangle\right) + \tau_{\star,k+1}\left(h(x_\star) - h(y_{k+1}) - \langle h'(y_{k+1}), x_\star - y_{k+1}\rangle\right) \\
&+ \tau_{k+1,k}\left(\left\langle h'(y_k), y_k - y_{k+1}\right\rangle - h(y_{k+1}) + h(y_k)\right) \geq 0 \qquad \triangleright\ \text{convexity of } h \text{ and by Lemma 17}
\end{aligned}
$$

For $k = -1$, recall that $\theta_{-1} = 0$,

$$
\mathcal{U}_{-1} - \mathcal{U}_0 = \mathcal{F}_{-1} - \mathcal{F}_0 + \mathcal{H}_{-1} - \mathcal{H}_0 \geq \frac{2\theta_{-1}}{\theta_N^2}\left\langle w_0 - x_\star + \frac{1}{L}\nabla f(x_\star), h'(y_0)\right\rangle + \frac{2\theta_{-1}^2}{L\theta_N^2}\|h'(y_{k+1})\|^2 + \frac{2\theta_{-1}^2}{L\theta_N^2}\langle h'(y_0)), \nabla f(x_0)\rangle = 0.
$$

Finally, we have

$$
\begin{aligned}
\mathcal{U}_{-1} &= \mathcal{F}_{-1} + \mathcal{H}_{-1} \\
&= \frac{L}{2\theta_N^2}\left\|x_0 - x_\star + \frac{1}{L}\nabla f(x_\star)\right\|^2 - \frac{1}{2L}\|\nabla f(x_\star)\|^2 + \frac{L}{2\theta_N^2(\theta_N^2-1)}\left\|x_0 - x_\star - \frac{\theta_N^2-1}{L}\nabla f(x_\star)\right\|^2 \\
&= \frac{L}{2\theta_N^2}\|x_0 - x_\star\|^2 + \frac{L}{2\theta_N^2(\theta_N^2-1)}\|x_0 - x_\star\|^2 + \frac{1}{\theta_N^2}\left\langle x_0 - x_\star, \nabla f(x_\star)\right\rangle - \frac{1}{\theta_N^2}\left\langle x_0 - x_\star, \nabla f(x_\star)\right\rangle \\
&- \frac{1}{2L}\|\nabla f(x_\star)\|^2 - \frac{\theta_N^2-1}{2L\theta_N^2}\|\nabla f(x_\star)\|^2 + \frac{1}{2L\theta_N^2}\|\nabla f(x_\star)\|^2 \\
&= \frac{L}{2(\theta_N^2-1)}\|x_0 - x_\star\|^2.
\end{aligned}
$$

□

D Omitted proof of Theorem 2

Now we show that the following choice of $\sigma_0, \dots, \sigma_N$, $a_0, \dots, a_N$, $x_0, \dots, x_N, x_\star$, $g_\star$, $f_0, \dots, f_N$, and $f_\star$ satisfies the conditions of Lemma 5.

$$\begin{aligned}
&\sigma_i = \tfrac{2\theta_i}{\theta_N^2},\ i \in [0:N-1], \qquad \sigma_N = \tfrac{1}{\theta_N},\\
&\zeta_{N+1} = \tfrac{(\theta_N-1)R^2}{\theta_N^2(2\theta_N-1)}, \quad \zeta_N = \tfrac{\theta_N}{\theta_N-1}\zeta_{N+1}, \quad \zeta_i = \tfrac{2\theta_i}{2\theta_i-1}\zeta_{i+1}\ , i \in [0:N-1],\\
&a_i = \tfrac{1}{\theta_N^2-1} \cdot \tfrac{\zeta_i}{\sigma_i\sqrt{\zeta_i-\zeta_{i+1}}},\ i \in [0:N], \qquad x_i = -(\theta_N^2-1)\sum_{k=0}^{i-1}\sigma_k a_k e_k,\\
&f_i = \tfrac{L}{2}a_i^2(4\theta_i-1) - \tfrac{LR^2}{2(\theta_N^2-1)^2}\ , i \in [0:N-1], \qquad f_N = \tfrac{LR^2}{2(\theta_N^2-1)},\\
&x_\star = -(\theta_N^2-1)\sum_{k=0}^{N}\sigma_k a_k e_k, \qquad g_\star = -\tfrac{L}{\theta_N^2-1}x_\star, \qquad f_\star = 0.
\end{aligned} \tag{28}$$

At this point, we already have (3), and (4). The remaining conditions are:

$$f_j - \frac{1}{L}\langle g_i, g_j\rangle + \frac{1}{2L}\|g_j - Lx_j\|^2 - \frac{L}{2}\|x_j\|^2 \geq f_k - \frac{1}{L}\langle g_i, g_k\rangle + \frac{1}{2L}\|g_k - Lx_k\|^2 - \frac{L}{2}\|x_k\|^2 \quad \text{for } \ i \in [0:N],\ j \in [0:N-1],\ k \in [j+1:N], \tag{5}$$

$$\sigma_i \geq 0,\ \sum_{i=0}^{N}\sigma_i = 1, \tag{6}$$

$$\sum_{i=0}^{N}\sigma_i\left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2\right) = f_\star + \tfrac{1}{2L}\|g_\star - Lx_\star\|^2 - \tfrac{L}{2}\|x_\star\|^2. \tag{7}$$

We provide the proofs of equations (5), (6), and (7) in separate lemmas. The following two lemma states the sufficient conditions to ensure (5).

Lemma 26 *Choice of f_N in* (28) *can be rearranged as*

$$f_N = \frac{LR^2}{2(\theta_N^2-1)} = \frac{L}{2}a_N^2(2\theta_N-1) - \frac{LR^2}{2(\theta_N^2-1)^2}.$$

Proof The right hand side can be expanded as

$$\begin{aligned}
\frac{L}{2}a_N^2(2\theta_N-1) - \frac{LR^2}{2(\theta_N^2-1)^2} &= \frac{L}{2}\cdot\frac{1}{(\theta_N^2-1)^2}\frac{\zeta_N^2}{\sigma_N^2(\zeta_N-\zeta_{N+1})}(2\theta_N-1) - \frac{LR^2}{2(\theta_N^2-1)^2}\\
&= \frac{L}{2}\cdot\frac{\theta_N^2}{(\theta_N^2-1)^2}\frac{\zeta_N^2}{\zeta_N-\zeta_{N+1}}(2\theta_N-1) - \frac{LR^2}{2(\theta_N^2-1)^2}.
\end{aligned}$$

Furthermore, $\frac{\zeta_N^2}{\zeta_N-\zeta_{N+1}}$ can be further simplified as

$$\frac{\zeta_N^2}{\zeta_N-\zeta_{N+1}} = \frac{\zeta_N^2}{\left(1-\frac{\theta_N-1}{\theta_N}\right)\zeta_N} = \theta_N\zeta_N = \frac{\theta_N^2}{\theta_N-1}\zeta_{N+1} = \frac{\theta_N^2}{\theta_N-1}\frac{\theta_N-1}{\theta_N^2(2\theta_N-1)}R^2 = \frac{R^2}{2\theta_N-1}.$$

Hence,

$$\begin{aligned}
\frac{L}{2}\cdot\frac{\theta_N^2}{(\theta_N^2-1)^2}\frac{\zeta_N^2}{\zeta_N-\zeta_{N+1}}(2\theta_N-1) - \frac{LR^2}{2(\theta_N^2-1)^2} &= \frac{L}{2}\cdot\frac{\theta_N^2}{(\theta_N^2-1)^2}(2\theta_N-1)\frac{R^2}{2\theta_N-1} - \frac{LR^2}{2(\theta_N^2-1)^2}\\
&= \frac{LR^2}{2}\cdot\frac{\theta_N^2}{(\theta_N^2-1)^2} - \frac{LR^2}{2(\theta_N^2-1)^2}\\
&= \frac{(\theta_N^2-1)LR^2}{2(\theta_N^2-1)^2} = \frac{LR^2}{2(\theta_N^2-1)}.
\end{aligned}$$

□

Lemma 27 *If*

$$2a_{j+1}^2\theta_{j+1} - 2a_j^2\theta_j + a_j^2 \le 0 \ , j \in [0:N-2],$$
$$a_N^2\theta_N - 2a_{N-1}^2\theta_{N-1} + a_{N-1}^2 \le 0$$

for (28)*, then* (5) *holds.*

Proof It is enough to prove (5) for $i \in [0:N],\ j \in [0:N-1]$, and $k = j+1$. Also note that $\langle g_i, x_i\rangle = 0$ for $i \in [0:N]$. We first plug in (28) to (5). Then,

$$f_j - \frac{1}{L}\langle g_i, g_j\rangle + \frac{1}{2L}\|g_j - Lx_j\|^2 - \frac{L}{2}\|x_j\|^2 \ge f_{j+1} - \frac{1}{L}\langle g_i, g_{j+1}\rangle + \frac{1}{2L}\|g_{j+1} - Lx_{j+1}\|^2 - \frac{L}{2}\|x_{j+1}\|^2$$

reduces to

$$f_j - \frac{1}{L}\langle g_i, g_j\rangle + \frac{1}{2L}\|g_j\|^2 \ge f_{j+1} - \frac{1}{L}\langle g_i, g_{j+1}\rangle + \frac{1}{2L}\|g_{j+1}\|^2 \tag{29}$$

Now divide the problem into smaller cases:

- *Case1-1*: $j \in [0:N-2],\ \ i < j$ or $i > j+1$

Then (29) is equivalent to

$$\frac{L}{2}a_j^2(4\theta_j - 1) - \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{L}{2}a_j^2 \ge \frac{L}{2}a_{j+1}^2(4\theta_{j+1} - 1) - \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{L}{2}a_{j+1}^2.$$

It reduces to

$$2a_j^2\theta_j \ge 2a_{j+1}^2\theta_{j+1},$$

which is true since $2a_j^2\theta_j - 2a_{j+1}^2\theta_{j+1} \ge a_j^2 \ge 0$.

- *Case1-2*: $j \in [0:N-2],\ \ i = j$

Then (29) is equivalent to

$$\frac{L}{2}a_j^2(4\theta_j - 1) - \frac{LR^2}{2(\theta_N^2-1)^2} - La_j^2 + \frac{L}{2}a_j^2 \ge \frac{L}{2}a_{j+1}^2(4\theta_{j+1} - 1) - \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{L}{2}a_{j+1}^2.$$

It reduces to

$$2a_j^2\theta_j - a_j^2 \ge 2a_{j+1}^2\theta_{j+1},$$

which is true by the assumption of the lemma.

- *Case1-3*: $j \in [0:N-2],\ \ i = j+1$

Then (29) is equivalent to

$$\frac{L}{2}a_j^2(4\theta_j - 1) - \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{L}{2}a_j^2 \ge \frac{L}{2}a_{j+1}^2(4\theta_{j+1} - 1) - La_{j+1}^2 - \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{L}{2}a_{j+1}^2.$$

It reduces to

$$2a_j^2\theta_j \ge 2a_{j+1}^2\theta_{j+1} - a_{j+1}^2,$$

which is true by the assumption of the lemma.

- *Case2-1*: $j = N-1,\ \ i < j$

Using Lemma 26, (29) is equivalent to

$$\frac{L}{2}a_{N-1}^2(4\theta_{N-1} - 1) - \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{L}{2}a_{N-1}^2 \ge \frac{L}{2}a_N^2(2\theta_N - 1) - \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{L}{2}a_N^2.$$

It reduces to

$$2a_{N-1}^2\theta_{N-1} \ge a_N^2\theta_N,$$

which is true since $2a_{N-1}^2\theta_{N-1} - a_N^2\theta_N \ge a_{N-1}^2 \ge 0$.

- *Case2-2*: $j = N-1,\ \ i = N-1$

Using Lemma 26, (29) is equivalent to

$$\frac{L}{2}a_{N-1}^2(4\theta_{N-1} - 1) - La_{N-1}^2 - \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{L}{2}a_{N-1}^2 \ge \frac{L}{2}a_N^2(2\theta_N - 1) - \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{L}{2}a_N^2.$$

It reduces to

$$2a_{N-1}^2\theta_{N-1} - a_{N-1}^2 \ge a_N^2\theta_N,$$

which is true by the assumption of the lemma.

- *Case2-3*: $j = N-1, \quad i = N$

Using Lemma 26, (29) is equivalent to

$$\frac{L}{2}a_{N-1}^2(4\theta_{N-1}-1) - \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{L}{2}a_{N-1}^2 \geq \frac{L}{2}a_N^2(2\theta_N-1) - \frac{LR^2}{2(\theta_N^2-1)^2} - La_N^2 + \frac{L}{2}a_N^2.$$

It reduces to

$$2a_{N-1}^2\theta_{N-1} \geq a_N^2\theta_N - a_N^2,$$

which is true by the assumption of the lemma. □

The following lemma proves (5).

Lemma 28 *Choice of a_i in* (28) *satisfies*

$$2a_{j+1}^2\theta_{j+1} - 2a_j^2\theta_j + a_j^2 \leq 0 \ , j \in [0:N-2],$$

$$a_N^2\theta_N - 2a_{N-1}^2\theta_{N-1} + a_{N-1}^2 \leq 0.$$

Proof Recall that

$$a_i = \frac{1}{\theta_N^2-1} \cdot \frac{\zeta_i}{\sigma_i\sqrt{\zeta_i - \zeta_{i+1}}} \text{ for } i \in [0:N].$$

Then for $j \in [0:N-2]$,

$$2a_{j+1}^2\theta_{j+1} - 2a_j^2\theta_j + a_j^2 = \frac{2\theta_{j+1}}{(\theta_N^2-1)^2} \cdot \frac{\zeta_{j+1}^2}{\sigma_{j+1}^2(\zeta_{j+1}-\zeta_{j+2})} - \frac{2\theta_j}{(\theta_N^2-1)^2} \cdot \frac{\zeta_j^2}{\sigma_j^2(\zeta_j-\zeta_{j+1})} + \frac{1}{(\theta_N^2-1)^2} \cdot \frac{\zeta_j^2}{\sigma_j^2(\zeta_j-\zeta_{j+1})}.$$

For convenience, drop the constant term $\frac{1}{(\theta_N^2-1)^2} > 0$ to get:

$$\frac{2\theta_{j+1}\zeta_{j+1}^2}{\sigma_{j+1}^2(\zeta_{j+1}-\zeta_{j+2})} - \frac{2\theta_j\zeta_j^2}{\sigma_j^2(\zeta_j-\zeta_{j+1})} + \frac{\zeta_j^2}{\sigma_j^2(\zeta_j-\zeta_{j+1})}. \tag{30}$$

Note that

$$\frac{\zeta_j^2}{\zeta_j-\zeta_{j+1}} = \frac{\left(\frac{2\theta_j}{2\theta_j-1}\right)^2\zeta_{j+1}^2}{\left(\frac{2\theta_j}{2\theta_j-1}-1\right)\zeta_{j+1}} = \frac{4\theta_j^2\zeta_{j+1}}{2\theta_j-1},$$

$$\frac{\zeta_{j+1}^2}{\zeta_{j+1}-\zeta_{j+2}} = \frac{\zeta_{j+1}^2}{\left(1-\frac{2\theta_{j+1}-1}{2\theta_{j+1}}\right)\zeta_{j+1}} = 2\theta_{j+1}\zeta_{j+1}.$$

Plugging into (30) gives:

$$\begin{aligned}\frac{2\theta_{j+1}\zeta_{j+1}^2}{\sigma_{j+1}^2(\zeta_{j+1}-\zeta_{j+2})} - \frac{2\theta_j\zeta_j^2}{\sigma_j^2(\zeta_j-\zeta_{j+1})} + \frac{\zeta_j^2}{\sigma_j^2(\zeta_j-\zeta_{j+1})} &= \frac{4\theta_{j+1}^2\zeta_{j+1}}{\sigma_{j+1}^2} - \frac{8\theta_j^3\zeta_{j+1}}{\sigma_j^2(2\theta_j-1)} + \frac{4\theta_j^2\zeta_{j+1}}{\sigma_j^2(2\theta_j-1)} \\ &= \theta_N^4\zeta_{j+1} - \theta_N^4 \cdot \frac{2\theta_j\zeta_{j+1}}{2\theta_j-1} + \theta_N^4 \cdot \frac{\zeta_{j+1}}{2\theta_j-1} \\ &= \theta_N^4\zeta_{j+1}\left(1-\frac{2\theta_j-1}{2\theta_j-1}\right) = 0.\end{aligned}$$

For $j = N-1$,

$$a_N^2\theta_N - 2a_{N-1}^2\theta_{N-1} + a_{N-1}^2 = \frac{\theta_N}{(\theta_N^2-1)^2} \cdot \frac{\zeta_N^2}{\sigma_N^2(\zeta_N-\zeta_{N+1})} - \frac{2\theta_{N-1}}{(\theta_N^2-1)^2} \cdot \frac{\zeta_{N-1}^2}{\sigma_{N-1}^2(\zeta_{N-1}-\zeta_N)} + \frac{1}{(\theta_N^2-1)^2} \cdot \frac{\zeta_{N-1}^2}{\sigma_{N-1}^2(\zeta_{N-1}-\zeta_N)}.$$

Similarly, drop the constant term to get:

$$\frac{\theta_N\zeta_N^2}{\sigma_N^2(\zeta_N-\zeta_{N+1})} - \frac{2\theta_{N-1}\zeta_{N-1}^2}{\sigma_{N-1}^2(\zeta_{N-1}-\zeta_N)} + \frac{\zeta_{N-1}^2}{\sigma_{N-1}^2(\zeta_{N-1}-\zeta_N)}. \tag{31}$$

Note that

$$\frac{\zeta_{N-1}^2}{\zeta_{N-1}-\zeta_N} = \frac{\left(\frac{2\theta_{N-1}}{2\theta_{N-1}-1}\right)^2\zeta_N^2}{\left(\frac{2\theta_{N-1}}{2\theta_{N-1}-1}-1\right)\zeta_N} = \frac{4\theta_{N-1}^2\zeta_N}{2\theta_{N-1}-1},$$

$$\frac{\zeta_N^2}{\zeta_N - \zeta_{N+1}} = \frac{\zeta_N^2}{\left(1 - \frac{\theta_{N-1}}{\theta_N}\right)\zeta_N} = \theta_N \zeta_N.$$

Plugging into (31) gives:

$$\begin{aligned}\frac{\theta_N \zeta_N^2}{\sigma_N^2(\zeta_N - \zeta_{N+1})} - \frac{2\theta_{N-1}\zeta_{N-1}^2}{\sigma_{N-1}^2(\zeta_{N-1} - \zeta_N)} + \frac{\zeta_{N-1}^2}{\sigma_{N-1}^2(\zeta_{N-1} - \zeta_N)} &= \frac{\theta_N^2 \zeta_N^2}{\sigma_N^2} - \frac{8\theta_{N-1}^3 \zeta_N}{\sigma_{N-1}^2(2\theta_{N-1} - 1)} + \frac{4\theta_{N-1}^2 \zeta_N}{\sigma_{N-1}^2(2\theta_{N-1} - 1)} \\ &= \theta_N^4 \zeta_N^2 - \theta_N^4 \cdot \frac{2\theta_{N-1}\zeta_N}{2\theta_{N-1} - 1} + \theta_N^4 \cdot \frac{\zeta_N}{2\theta_{N-1} - 1} \\ &= \theta_N^4 \zeta_N \left(1 - \frac{2\theta_{N-1} - 1}{2\theta_{N-1} - 1}\right) = 0.\end{aligned}$$

□

The next lemma proves (6) directly.

Lemma 29 *The following holds for* (28).

$$\sum_{i=0}^{N} \sigma_i = 1,\ \sigma_i \geq 0\ , i \in [0:N].$$

Proof

$$\begin{aligned}\sum_{i=0}^{N} \sigma_i = \sum_{i=0}^{N-1} \sigma_i + \sigma_N &= \sum_{i=0}^{N-1} \frac{2\theta_i}{\theta_N^2} + \frac{1}{\theta_N} \\ &= \frac{1}{\theta_N^2} \cdot 2\theta_{N-1}^2 + \frac{1}{\theta_N} \qquad \triangleright \text{using (23) of Lemma 19} \\ &= \frac{2\theta_{N-1}^2 + \theta_N}{\theta_N^2} = 1. \qquad \triangleright \text{using (22) of Lemma 19}\end{aligned}$$

Also, positivity of σ_i for $i \in [0:N]$ is clear from the definition of $\{\theta\}_{i\in[0:N]}$. □

The following two lemma proves (7).

Lemma 30 *The following holds for* (28).

$$\|x_0 - x_\star\|^2 = \|x_\star\|^2 = \sum_{i=0}^{N} \frac{\zeta_i^2}{\zeta_i - \zeta_{i+1}} = R^2.$$

Proof We refer the reader to [23, Lemma 3]. □

Lemma 31 *The following holds for* (28).

$$\sum_{i=0}^{N} \sigma_i \left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2\right) = f_\star + \tfrac{1}{2L}\|g_\star - Lx_\star\|^2 - \tfrac{L}{2}\|x_\star\|^2.$$

Proof Recall $\langle x_i, g_i\rangle = 0$ for $i \in [0:N]$. Then,

$$\begin{aligned}&\sum_{i=0}^{N} \sigma_i \left(f_i + \frac{1}{2L}\|g_i - Lx_i\|^2 - \frac{L}{2}\|x_i\|^2\right) = \sum_{i=0}^{N} \sigma_i \left(f_i + \frac{1}{2L}\|g_i\|^2\right) \qquad \triangleright \text{expanding } \|g_i - Lx_i\|^2 \\ &= \sum_{i=0}^{N-1} \sigma_i \left(f_i + \frac{1}{2L}\|g_i\|^2\right) + \sigma_N \left(f_N + \frac{1}{2L}\|g_N\|^2\right) \\ &= \sum_{i=0}^{N-1} \sigma_i \left(\frac{L}{2}a_i^2(4\theta_i - 1) - \frac{LR^2}{2(\theta_N^2 - 1)^2} + \frac{L}{2}a_i^2\right) \\ &+ \sigma_N \left(\frac{L}{2}a_N^2(2\theta_N - 1) - \frac{LR^2}{2(\theta_N^2 - 1)^2} + \frac{L}{2}a_N^2\right) \qquad \triangleright \text{using Lemma 26} \\ &= \sum_{i=0}^{N-1} 2L\sigma_i a_i^2 \theta_i - \sum_{i=0}^{N-1} \sigma_i \frac{LR^2}{2(\theta_N^2 - 1)^2} + L\sigma_N a_N^2 \theta_N - \sigma_N \frac{LR^2}{2(\theta_N^2 - 1)^2} \\ &= L\theta_N^2 \sum_{i=0}^{N-1} \sigma_i^2 a_i^2 + L\sigma_N a_N^2 \theta_N - \frac{LR^2}{2(\theta_N^2 - 1)^2}. \qquad \triangleright \text{using Lemma 29}\end{aligned}$$

Here, $\sum_{i=0}^{N-1} \sigma_i^2 a_i^2$ is reduced to

$$\sum_{i=0}^{N-1} \sigma_i^2 a_i^2 = \sum_{i=0}^{N} \sigma_i^2 a_i^2 - \sigma_N a_N = \frac{1}{(\theta_N^2-1)^2} \sum_{i=0}^{N} \sigma_i^2 \frac{\zeta_i^2}{\sigma_i^2(\zeta_i - \zeta_{i+1})} - \sigma_N a_N = \frac{R^2}{(\theta_N^2-1)^2} - \sigma_N a_N,$$

where the last equality is from Lemma 30. So,

$$\begin{aligned} L\theta_N^2 \sum_{i=0}^{N-1} \sigma_i^2 a_i^2 + L\sigma_N a_N^2 \theta_N - \frac{LR^2}{2(\theta_N^2-1)^2} &= \theta_N^2 \cdot \frac{LR^2}{(\theta_N^2-1)^2} - L\theta_N^2 \sigma_N^2 a_N^2 + L\sigma_N a_N^2 \theta_N - \frac{LR^2}{2(\theta_N^2-1)^2} \\ &= \theta_N^2 \cdot \frac{LR^2}{(\theta_N^2-1)^2} - L\theta_N^2 \frac{1}{\theta_N^2} a_N^2 + L\frac{1}{\theta_N} a_N^2 \theta_N - \frac{LR^2}{2(\theta_N^2-1)^2} \\ &= \frac{2\theta_N^2 - 1}{2(\theta_N^2-1)^2} LR^2. \end{aligned}$$

Meanwhile,

$$\begin{aligned} f_\star + \frac{1}{2L}\|g_\star - Lx_\star\|^2 - \frac{L}{2}\|x_\star\|^2 = \frac{1}{2L}\|g_\star\|^2 - \langle g_\star, x_\star \rangle &= \frac{1}{2L}\left\| -\frac{L}{\theta_N^2 - 1} x_\star \right\|^2 - \left\langle -\frac{L}{\theta_N^2-1} x_\star, x_\star \right\rangle \\ &= \frac{LR^2}{2(\theta_N^2-1)^2} + \frac{LR^2}{\theta_N^2 - 1} = \frac{2\theta_N^2 - 1}{2(\theta_N^2-1)^2} LR^2. \end{aligned}$$

□

Therefore, by the aforementioned lemmas, we are able to apply Lemma 5 to (28). We now show semi-interpolating conditions to establish the lower bound of Theorem 2.

Lemma 32 *Let*

$$f(x) = \max_{\alpha \in \Delta_I} v(x, \alpha) = \max_{\alpha \in \Delta_I} \tfrac{L}{2}\|x\|^2 - \tfrac{L}{2}\|x - \tfrac{1}{L}\sum_{i \in I} \alpha_i g_i\|^2 + \sum_{i \in I} \alpha_i \left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2 \right),$$

and for some $i \in I$, suppose

$$f_i \geq f_j + \langle g_j, x_i - x_j \rangle + \frac{1}{2L}\|g_i - g_j\|^2 \quad \text{for all } j \in I.$$

Then $f(x_i) = f_i$ and $\nabla f(x_i) = g_i$

Proof We refer the reader to [25, Theorem 1]. □

Lemma 33 *If $\{g_i\}_{i \in [0:N]}$ are orthogonal and*

$$g_\star = \sum_{i=0}^{N} \sigma_i g_i \text{ for some } \sigma_i \geq 0,$$

then, $-g_\star \in \partial h(x_\star)$.

Proof Since $x_\star \in C$,

$$\partial h(x_\star) = \{y \,|\, \langle y, c - x_\star \rangle \leq 0 \quad \forall c \in C\}.$$

Let $\tilde{z} = x_\star + \sum_{i=0}^{N} c_i g_i$ for $c_i \geq 0$. Then,

$$\langle -g_\star, \tilde{z} - x_\star \rangle = \left\langle -g_\star, x_\star + \sum_{i=0}^{N} c_i g_i - x_\star \right\rangle = \left\langle -\sum_{i=0}^{N} \sigma_i g_i \ , \ \sum_{i=0}^{N} c_i g_i \right\rangle = -\sum_{i=0}^{N} \sigma_i c_i \|g_i\|^2 \leq 0.$$

□

Lemma 34 *The choice of* (28) *satisfies the following:*

$$f_\star \geq f_i + \langle g_i, x_\star - x_i \rangle + \frac{1}{2L}\|g_\star - g_i\|^2 \text{ for } i \in I,$$

and therefore $f(x_\star) = f_\star$ and $\nabla f(x_\star) = g_\star$ by Lemma 32.

Proof In fact, we will show that:

$$f_i = f_\star - \langle g_i, x_\star - x_i\rangle - \frac{1}{2L}\|g_\star - g_i\|^2 \text{ for } i \in I.$$

For $i \in [0:N]$,

$$\begin{aligned}
&f_\star - \langle g_i, x_\star - x_i\rangle - \frac{1}{2L}\|g_\star - g_i\|^2\\
&= 0 - \left\langle La_ie_i, -(\theta_N^2-1)\sum_{k=0}^{N}\sigma_k a_k e_k + (\theta_N^2-1)\sum_{k=0}^{i-1}\sigma_k a_k e_k\right\rangle - \frac{1}{2L}\left\|L\sum_{k=0}^{N}\sigma_k a_k e_k - La_ie_i\right\|^2\\
&= \left\langle La_ie_i, (\theta_N^2-1)\sum_{k=i}^{N}\sigma_k a_k e_k\right\rangle - \frac{1}{2L}\left\|L\sum_{k=0}^{N}\sigma_k a_k e_k\right\|^2 + \frac{1}{L}\left\langle L\sum_{k=0}^{N}\sigma_k a_k e_k, La_ie_i\right\rangle - \frac{1}{2L}\|La_ie_i\|^2\\
&= L(\theta_N^2-1)\sigma_i a_i^2 - \frac{1}{2L}\left\|\frac{L}{\theta_N^2-1}x_\star\right\|^2 + L\sigma_i a_i^2 - \frac{L}{2}a_i^2\\
&= \frac{L}{2}a_i^2\left(2\sigma_i\left(\theta_N^2-1\right)+2\sigma_i-1\right) - \frac{LR^2}{2(\theta_N^2-1)^2}. \qquad \triangleright \left\|\frac{L}{\theta_N^2-1}x_\star\right\|^2 = \frac{L^2R^2}{(\theta_N^2-1)^2}
\end{aligned}$$

If $i \in [0:N-1]$,

$$\begin{aligned}
\frac{L}{2}a_i^2\left(2\sigma_i\left(\theta_N^2-1\right)+2\sigma_i-1\right) - \frac{LR^2}{2(\theta_N^2-1)^2} &= \frac{L}{2}a_i^2\left(\frac{4\theta_i}{\theta_N^2}\left(\theta_N^2-1\right)+\frac{4\theta_i}{\theta_N^2}-1\right) - \frac{LR^2}{2(\theta_N^2-1)^2}\\
&= \frac{L}{2}a_i^2\left(4\theta_i-1\right) - \frac{LR^2}{2(\theta_N^2-1)^2} = f_i.
\end{aligned}$$

If $i = N$,

$$\begin{aligned}
\frac{L}{2}a_N^2\left(2\sigma_N\left(\theta_N^2-1\right)+2\sigma_N-1\right) - \frac{LR^2}{2(\theta_N^2-1)^2} &= \frac{L}{2}a_N^2\left(\frac{2}{\theta_N}\left(\theta_N^2-1\right)+\frac{2}{\theta_N}-1\right) - \frac{LR^2}{2(\theta_N^2-1)^2}\\
&= \frac{L}{2}a_N^2\left(2\theta_N-1\right) - \frac{LR^2}{2(\theta_N^2-1)^2} = f_N. \qquad \triangleright \text{using Lemma 26}
\end{aligned}$$

□

The following lemma is a restriction of Theorem 2 in the sense that it has a fixed starting point $x_0 = z_0 = 0$ and fixed dimension $d = N+1$.

Lemma 35 *Let $L > 0$, $R > 0$, $N > 0$, and let $x_0 = z_0 = 0 \in \mathbb{R}^{N+1}$. Let L-smooth and convex function $f\colon \mathbb{R}^{N+1} \to \mathbb{R}$ as*

$$f(x) = \max_{\alpha\in\Delta_I} \tfrac{L}{2}\|x\|^2 - \tfrac{L}{2}\|x - \tfrac{1}{L}\sum_{i\in I}\alpha_i g_i\|^2 + \sum_{i\in I}\alpha_i\left(f_i + \tfrac{1}{2L}\|g_i - Lx_i\|^2 - \tfrac{L}{2}\|x_i\|^2\right),$$

and closed, convex, and proper function $h\colon \mathbb{R}^{N+1} \to \mathbb{R}\cup\{\infty\}$ as $h(x) = \delta_C(x)$, where δ_C is the indicator function of the convex set $C = x_\star + \operatorname{cone}\{g_0, g_1, \ldots, g_N\}$, with the choice of (28).

Then $x_\star \in \operatorname{argmin}(f+h)$ and satisfies $\|x_0 - x_\star\| = R$ and

$$f(x_N) + h(x_N) - f(x_\star) - h(x_\star) \geq \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2-1)}.$$

for any sequence $\{\{(z_i,\delta_i,\gamma_i)\}_{i\in[0:2N-1]}, x_N\}$ satisfying the following double-function span condition:

$$\begin{aligned}
&\delta_i \in \{0,1\}, && \text{for } i = 0,\ldots,2N-1,\\
&\sum_{i=0}^{2N-1}\delta_i = \sum_{i=0}^{2N-1}(1-\delta_i) = N,\\
&d_i = \begin{cases}\nabla f(z_i) & \text{if } \delta_i = 0,\\ z_i - \mathbf{prox}_{\gamma_i h}(z_i) & \text{if } \delta_i = 1,\end{cases} && \text{for } i = 0,\ldots,2N-1,\\
&z_i \in x_0 + \operatorname{span}\{d_0,\ldots,d_{i-1}\}, && \text{for } i = 1,\ldots,2N-1,\\
&x_N \in x_0 + \operatorname{span}\{d_0,\ldots,d_{2N-1}\}.
\end{aligned} \tag{32}$$

To clarify, x_N in (32) *is unrelated to x_N in* (28).

Proof By Lemma 29 and Lemma 33, we have $-g_\star \in \partial h(x_\star)$. By Lemma 34, we have $g_\star = \nabla f(x_\star)$. Therefore, $x_\star \in \operatorname{argmin}(f+h)$ with $f(x_\star) + h(x_\star) = f_\star + 0 = 0$, and $\|x_0 - x_\star\| = R$ by Lemma 30. Now by Lemma 5, when the stating point is $x_0 = z_0 = 0$, each gradient evaluation will give at most one new next coordinate, and proximal evaluations does not introduce any new coordinate. So, after N gradient and N proximal evaluations respectively (order does not matter), the output x_N will be in the span of $\{e_0, \dots, e_{N-1}\}$ under the condition (32) of the lemma. Now we apply the second result of Lemma 5 to conclude that

$$\begin{aligned} f(x_N) + h(x_N) - f(x_\star) - h(x_\star) &\geq \inf_{x \in \operatorname{span}\{e_0,\dots,e_{N-1}\}} (f(x) + h(x)) - 0 \\ &\geq \inf_{x \in \operatorname{span}\{e_0,\dots,e_{N-1}\}} f(x) \geq f_N = \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2 - 1)}. \end{aligned}$$

□

Then we expand the condition of Lemma 35 to arrive at Theorem 2.

Proof of Theorem 2 Assume $d \geq N+1$. Take $f_0 \in \mathcal{F}_{0,L}$ and $h_0 \in \mathcal{F}_{0,\infty}$ be functions defined in Lemma 35, which are embedded in $\mathbb{R}^d$. Call $\tilde{x}_\star$ to be the element of $\operatorname{argmin}_{x \in \mathbb{R}^d}(f_0 + h_0)$ in Lemma 35. Now for arbitrary $x_0 = z_0$, let $f \colon \mathbb{R}^d \to \mathbb{R}$ and $h \colon \mathbb{R}^d \to \mathbb{R}$ be translation of f_0 and h_0 by x_0:

$$f(x) = f_0(x - x_0), \quad h(x) = h_0(x - x_0).$$

Then f is L-smooth and convex, h is an indicator function of nonempty closed convex set, and $x_\star \triangleq \tilde{x}_\star + x_0 \in \operatorname{argmin}_{x \in \mathbb{R}^d}(f+h)$. Now assume the sequence $\{\{(z_i, \delta_i, \gamma_i)\}_{i \in [0:2N-1]}, x_N\}$ is produced from an method that satisfies the double-function span condition. That is,

$$z_i \in z_0 + \operatorname{span}\{d_0, \dots, d_{i-1}\},$$

$$x_N \in z_0 + \operatorname{span}\{d_0, \dots, d_{2N-1}\},$$

where

$$d_i = \begin{cases} \nabla f(z_i) & \text{if } \delta_i = 0\ , \\ z_i - \mathbf{prox}_{\gamma_i h}(z_i) & \text{if } \delta_i = 1. \end{cases}$$

Then for $\tilde{z}_i \triangleq z_i - x_0 = z_i - z_0$ and $\tilde{x}_N \triangleq x_N - x_0 = x_N - z_0$, $\{\{(\tilde{z}_i, \delta_i, \gamma_i)\}_{i \in [0:2N-1]}, \tilde{x}_N\}$ satisfies:

$$\tilde{z}_i \in \tilde{z}_0 + \operatorname{span}\left\{\tilde{d}_0, \dots, \tilde{d}_{i-1}\right\},$$

$$\tilde{x}_N \in \tilde{z}_0 + \operatorname{span}\left\{\tilde{d}_0, \dots, \tilde{d}_{2N-1}, \right\}$$

where

$$\tilde{d}_i = \begin{cases} \nabla f_0(\tilde{z}_i) & \text{if } \delta_i = 0\ , \\ \tilde{z}_i - \mathbf{prox}_{\gamma_i h_0}(\tilde{z}_i) & \text{if } \delta_i = 1. \end{cases}$$

This is because $\tilde{z}_0 = 0$ and

$$\nabla f(z_i) = \nabla f(\tilde{z}_i + x_0) = \nabla f_0(\tilde{z}_i + x_0 - x_0) = \nabla f_0(\tilde{z}_i),$$
$$\mathbf{prox}_{\gamma_i h}(z_i) - x_0 = \mathbf{prox}_{\gamma_i h_0}(z_i - x_0) = \mathbf{prox}_{\gamma_i h_0}(\tilde{z}_i).$$

So,

$$\tilde{z}_i - \mathbf{prox}_{\gamma_i h_0}(\tilde{z}_i) = z_i - x_0 - \mathbf{prox}_{\gamma_i h}(z_i) + x_0 = z_i - \mathbf{prox}_{\gamma_i h}(z_i).$$

Hence, we can apply Lemma 35 on $\{\{(\tilde{z}_i, \delta_i, \gamma_i)\}_{i \in [0:2N-1]}, \tilde{x}_N\}$ to get

$$f_0(\tilde{x}_N) + h_0(\tilde{x}_N) - f_0(\tilde{x}_\star) - h_0(\tilde{x}_\star) \geq \frac{L\|0 - \tilde{x}_\star\|^2}{2(\theta_N^2 - 1)} = \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2 - 1)} = \frac{LR^2}{2(\theta_N^2 - 1)}.$$

Then we finally get

$$\begin{aligned} f(x_N) + h(x_N) - f(\tilde{x}_\star + x_0) - h(\tilde{x}_\star + x_0) &= f_0(\tilde{x}_N) + h_0(\tilde{x}_N) - f_0(\tilde{x}_\star) - h_0(\tilde{x}_\star) \\ &\geq \frac{L\|0 - \tilde{x}_\star\|^2}{2(\theta_N^2 - 1)} = \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2 - 1)} = \frac{LR^2}{2(\theta_N^2 - 1)}, \end{aligned}$$

and f and h are our desired functions. □

E Omitted proof of Theorem 4

We now make formal definition of deterministic N-step double-oracle method.

A *deterministic N-step double-oracle method* $\mathbf{A}$ is a mapping from initial point z_0 and a tuple of function (f,h) to $\mathbf{A}(z_0,(f,h)) = \{\{(z_i,\delta_i,\gamma_i)\}_{i\in[0:2N-1]}, x_N\}$. Here, we use x_N for consistency with Section 3, and we call x_N the *approximate solution*, or simply the *output*. The sequence $\mathbf{A}(z_0,(f,h)) = \{\{(z_i,\delta_i,\gamma_i)\}_{i\in[0:2N-1]}, x_N\}$ depends on (f,h) only through N queries to gradient oracle $\mathcal{O}_f(z_i) = (\nabla f(z_i), f(z_i))$ and N to proximal oracle $\mathcal{O}_h(z_i,\gamma_i) = (\mathbf{prox}_{\gamma_i h}(z_i), h(z_i))$. Here, z_i denotes the next (gradient or proximal) query point to an oracle, $\delta_i \in \{0,1\}$ is the indicator that tells an method to which oracle to query next. Specifically, if $\delta_i = 0$, z_i is fed into $\mathcal{O}_f(\cdot)$, otherwise z_i is fed into $\mathcal{O}_h(\cdot,\gamma_i)$, where and $\gamma_i > 0$ is i-th proximal stepsize. For further precision, we define

$$\begin{aligned}
(\delta_0,\gamma_0) &= \mathbf{A}_0(z_0),\\
(z_i,\delta_i,\gamma_i) &= \mathbf{A}_i\left(\{(z_j,\delta_j,\gamma_j)\}_{j=0}^{i-1}, \{(1-\delta_j)\mathcal{O}_f(z_j), \delta_i\mathcal{O}_h(z_j,\gamma_j)\}_{j=0}^{i-1}\right) \qquad \text{for } i\in[1:2N-1],\\
x_N &= \mathbf{A}_{2N}\left(\{(z_j,\delta_j,\gamma_j)\}_{j=0}^{2N-1}, \{(1-\delta_j)\mathcal{O}_f(z_i), \delta_j\mathcal{O}_h(z_j,\gamma_j)\}_{j=0}^{2N-1}\right) \qquad \text{for } i = 2N,
\end{aligned}$$

where we use, for notational shorthand,

- $0\cdot\mathcal{O}_f = \emptyset$ ($\mathcal{O}_f$ is not evaluated)
- $0\cdot\mathcal{O}_h = \emptyset$ ($\mathcal{O}_h$ is not evaluated)
- $1\cdot\mathcal{O}_f = \mathcal{O}_f$ ($\mathcal{O}_f$ is evaluated)
- $1\cdot\mathcal{O}_h = \mathcal{O}_h$ ($\mathcal{O}_h$ is evaluated).

Additionally, the number of queries to $\mathcal{O}_f$ and $\mathcal{O}_f$ must each be exactly N. i.e.,

$$\sum_{j=0}^{2N-1}\delta_j = \sum_{j=0}^{2N-1}(1-\delta_j) = N.$$

Theorem 4 expands the result of Theorem 2 to any deterministic N-step double-oracle methods using resisting oracle technique [54,12,61]. To get the desired result, we use *double-function zero-respecting* sequences, which is similar but slightly general than the double-function span condition. The sequence $\{\{(z_i,\delta_i,\gamma_i)\}_{i\in[0:2N-1]}, x_N\}$ is said to be *double-function zero-respecting* with respect to (f,h) if

$$\begin{aligned}
&\delta_i \in \{0,1\}, && \text{for } i = 0,\dots,2N-1,\\
&\sum_{i=0}^{2N-1}\delta_i = \sum_{i=0}^{2N-1}(1-\delta_i) = N,\\
&d_i = \begin{cases}\nabla f(z_i) & \text{if } \delta_i = 0,\\ z_i - \mathbf{prox}_{\gamma_i h}(z_i) & \text{if } \delta_i = 1\end{cases}, && \text{for } i = 0,\dots,2N-1,\\
&\operatorname{supp}\{z_i\} \subseteq \operatorname{supp}\{d_0,\dots,d_{i-1}\}, && \text{for } i = 1,\dots,2N-1,\\
&\operatorname{supp}\{x_N\} \subseteq \operatorname{supp}\{d_0,\dots,d_{2N-1}\},
\end{aligned}$$

where $\operatorname{supp}\{z_i\} \triangleq \{j\in[0:d-1] \mid \langle e_j, z_i\rangle \neq 0\}$ and $\operatorname{supp}\{d_0,\dots,d_i\} \triangleq \bigcup_{j=0}^{i}\operatorname{supp}\{d_j\}$. Note that by definition, $z_0 = 0$ for any double-function zero-respecting sequences.

We say a matrix $U\in\mathbb{R}^{m\times n}$ is *orthogonal* if columns $\{u_i\}_{i\in[0:m-1]}\in\mathbb{R}^n$ of U are mutually orthonormal, or equivalently, $U^\intercal U = I_n$.

The following lemmas are building blocks for resisting oracle technique.

Lemma 36 *For any orthogonal matrix $U\in\mathbb{R}^{d'\times d}$ with $d'\geq d$ and any vector $x_0\in\mathbb{R}^{d'}$, if $f\colon\mathbb{R}^d\to\mathbb{R}$ is L-smooth and convex, then $f_U\colon\mathbb{R}^{d'}\to\mathbb{R}$ defined by*

$$f_U(x) \triangleq f(U^\intercal(x-x_0))$$

is L-smooth and convex.

Proof Note that $\nabla f_U(x) = U\nabla f(U^\intercal(x-x_0))$. For any $x,y\in\mathbb{R}^{d'}$,

$$\begin{aligned}
\|\nabla f_U(x) - \nabla f_U(y)\| &= \|U\nabla f(U^\intercal(x-x_0)) - U\nabla f(U^\intercal(y-x_0))\|\\
&= \|\nabla f(U^\intercal(x-x_0)) - \nabla f(U^\intercal(y-x_0))\| \qquad \triangleright\, U \text{ is orthogonal}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&f_U(y) - f_U(x) - \langle\nabla f_U(x), y-x\rangle - \frac{1}{2L}\|\nabla f_U(x) - \nabla f_U(y)\|^2 = f(U^\intercal(y-x_0)) - f(U^\intercal(x-x_0))\\
&- \langle U\nabla f(U^\intercal(x-x_0)), y-x\rangle - \frac{1}{2L}\|\nabla f(U^\intercal(x-x_0)) - \nabla f(U^\intercal(y-x_0))\|^2\\
&= f(U^\intercal(y-x_0)) - f(U^\intercal(x-x_0)) - \langle\nabla f(U^\intercal(x-x_0)), U^\intercal(y-x)\rangle\\
&- \frac{1}{2L}\|\nabla f(U^\intercal(x-x_0)) - \nabla f(U^\intercal(y-x_0))\|^2 \geq 0,
\end{aligned}$$

where the last inequality is from L-smoothness and convexity of f. □

Lemma 37 *For any orthogonal matrix $U \in \mathbb{R}^{d' \times d}$ with $d' \geq d$ and any vector x_0, if $h \colon \mathbb{R}^d \to \mathbb{R}$ is an indicator function of nonempty closed convex set, then $h_U \colon \mathbb{R}^{d'} \to \mathbb{R}$ defined by*

$$h_U(x) \triangleq h(U^\intercal(x - x_0))$$

is an indicator function of nonempty closed convex set.

Proof Assume $h = \delta_C$ for nonempty closed convex set $C \subseteq \mathbb{R}^d$ and let

$$\tilde{C} \triangleq \{x \,|\, U^\intercal(x - x_0) \in C\} \subseteq \mathbb{R}^{d'}.$$

Then, $h_U = \delta_{\tilde{C}}$. Now we claim that $\tilde{C}$ is nomempty, closed, and convex. Take $c \in C$. Then since $Uc + x_0 \in \tilde{C}$, $\tilde{C} \neq \emptyset$. $\tilde{C}$ is closed since it is preimage of C under continuous transformation $U^\intercal(\cdot - x_0)$. For convexity, suppose $x, y \in \tilde{C}$ and $t \in [0, 1]$. Then,

$$\begin{aligned} U^\intercal\left((1-t)x + ty - x_0\right) &= U^\intercal\left((1-t)(x - x_0) + t(y - x_0)\right) \\ &= (1-t)U^\intercal(x - x_0) + tU^\intercal(y - x_0) \in C \end{aligned}$$

where the last inclusion is from the convexity of C. Therefore we conclude $(1-t)x + ty \in \tilde{C}$. □

From now, $U \in \mathbb{R}^{d' \times d}$ is orthogonal, and f_U and h_U are as defined in lemma 36 and 37 respectively.

Lemma 38 *Suppose $f \colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth convex and $h \colon \mathbb{R}^d \to \mathbb{R}$ is indicator function of nonempty closed convex set. If $\tilde{x}_\star \in \operatorname{argmin}_{x \in \mathbb{R}^d}(f + h)$ exists, then $x^\star \triangleq U\tilde{x}_\star + x_0 \in \operatorname{argmin}_{x \in \mathbb{R}^{d'}}(f_U + h_U)$.*

Proof Note that $U^\intercal U = I_d$. For any $x \in \mathbb{R}^{d'}$,

$$\begin{aligned} f_U(x_\star) + h_U(x_\star) &= f(U^\intercal(U\tilde{x}_\star + x_0 - x_0)) + h(U^\intercal(U\tilde{x}_\star + x_0 - x_0)) \\ &= f(\tilde{x}_\star) + h(\tilde{x}_\star) \\ &\leq f(U^\intercal(x - x_0)) + h(U^\intercal(x - x_0)) \qquad \triangleright\, \tilde{x}_\star \in \operatorname*{argmin}_{x \in \mathbb{R}^d}(f + h) \\ &= f_U(x) + h_U(x). \end{aligned}$$

□

Lemma 39 *Take $f \colon \mathbb{R}^d \to \mathbb{R}$ and $h \colon \mathbb{R}^d \to \mathbb{R}$ to be the functions that are defined in Theorem 2. If $\{\{(z_i, \delta_i, \gamma_i)\}_{i \in [0:2N-1]}, x_N\}$ is double-function zero-respecting with respect to (f, h), then it satisfies the double-function span condition* (32).

Proof Assume $\{\{(z_i, \delta_i, \gamma_i)\}_{i \in [0:2N-1]}, x_N\}$ is double-function zero-respecting with respect to (f, h). Then, we have $z_0 = 0$ by the double-function zero-respecting assumption. From this, we have $d_0 \in \operatorname{span}\{e_0\}$ by the property of (f, h) in Theorem 2 when $z_0 = 0$. Now suppose $d_i \in \operatorname{span}\{e_0, \ldots, e_i\}$ for all $0 \leq i < k$ where $0 < k \leq 2N - 1$. Then, $\operatorname{supp}\{z_k\} \subseteq \operatorname{supp}\{d_0, \ldots, d_{k-1}\} \subseteq \{0, \ldots, k-1\}$ by the double-function zero-respecting assumption. By the property of (f, h), we then have $d_k \in \{e_0, \ldots, e_k\}$. Therefore, $d_i \in \operatorname{span}\{e_0, \ldots, e_i\}$ holds for all $i \in [0 : 2N - 1]$. This proves the equivalence of double-function span condition and double-function zero-respecting sequence, under the choice of (f, h) from Theorem 2. □

Lemma 40 *Assume $L > 0$, $N > 0$, and $d' \geq d + 2N$. Let $\mathbf{A}$ be any deterministic N-step double-oracle method, $x_0 = z_0 \in \mathbb{R}^{d'}$ be any vector, $f \colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex, $h \colon \mathbb{R}^d \to \mathbb{R}$ is an indicator function of nonempty closed convex set. Then there exists an orthogonal matrix $U \in \mathbb{R}^{d' \times d}$ and $\mathbf{A}(z_0, (f_U, h_U)) = \{\{(z_i, \delta_i, \gamma_i)\}_{i \in [0:2N-1]}, x_N\}$ such that if $\tilde{z}_i \triangleq U^\intercal(z_i - z_0) \in \mathbb{R}^d$, then $\{\{(\tilde{z}_i, \delta_i, \gamma_i)\}_{i \in [0:2N-1]}, \tilde{x}_N\}$ is double-function zero-respecting with respect to (f, h).*

Proof Recall the support condition of double-function zero-respecting sequences.

$$\operatorname{supp}\{z_i\} \subseteq \bigcup_{j=0}^{i-1} \operatorname{supp}\left\{\nabla f(z_j) \text{ or } z_j - \mathbf{prox}_{\gamma_j h}(z_j)\right\} = \operatorname{supp}\{d_0, \ldots, d_{i-1}\},$$

$$\operatorname{supp}\{x_N\} = \operatorname{supp}\{d_0, \ldots, d_{2N-1}\},$$

where

$$d_i = \begin{cases} \nabla f(z_i) & \text{if } \delta_i = 0, \\ z_i - \mathbf{prox}_{\gamma_i h}(z_i) & \text{if } \delta_i = 1. \end{cases}$$

For $i \in [0 : 2N]$, let $I_i \subseteq [0 : d-1]$ be

$$I_i = \operatorname{supp}\left\{\tilde{d}_0, \ldots, \tilde{d}_{i-1}\right\},$$

where

$$\tilde{d}_i = \begin{cases} \nabla f(\tilde{z}_i) & \text{if } \delta_i = 0, \\ \tilde{z}_i - \mathbf{prox}_{\gamma_i h}(\tilde{z}_i) & \text{if } \delta_i = 1. \end{cases}$$

Then $\{\{(\tilde{z}_i, \delta_i, \gamma_i)\}_{i\in[0:2N-1]}, \tilde{x}_N\}$ being double-function zero-respecting with respect to (f, h) is equivalent to

$$\mathrm{supp}\{\tilde{z}_i\} \subseteq I_i,\ i \in [0:2N-1] \text{ and } \mathrm{supp}\{\tilde{x}_N\} \subseteq I_{2N}.$$

Also note that

$$\emptyset = I_0 \subseteq I_1 \subseteq \cdots \subseteq I_{2N-1} \subseteq I_{2N} \subseteq [0:d-1].$$

Now we construct the matrix $U = [u_0 \mid \cdots \mid u_{d-1}]$ by choosing appropriate $u_i \in \mathbb{R}^{d'}$ inductively. For $i = 0$, $\tilde{z}_0 = 0$ is clear from the definition. Since $\tilde{z}_i = U^\intercal(z_i - z_0)$, $\mathrm{supp}\{\tilde{z}_i\} \subseteq I_i$ is again equivalent to

$$\langle u_j, z_i - z_0\rangle = 0 \quad \forall j \notin I_i,\ i \in [0:2N-1] \text{ and } \langle u_j, x_N - z_0\rangle = 0,\ \forall j \notin I_{2N}.$$

Now for $i \in [1:2N]$, choose $\{u_j\}_{j\in I_i \setminus I_{i-1}}$ from the orthogonal complement of

$$S_{i-1} \triangleq \mathrm{span}\left\{\{z_1 - z_0, \cdots, z_{i-1} - z_0\} \cup \{u_j\}_{j\in I_{i-1}}\right\}$$

and let them be mutually orthonormal. At the last step, choose $\{u_j\}_{j\in[0:d-1]\setminus I_{2N}}$ from the orthogonal complement of

$$S_{2N} \triangleq \mathrm{span}\left\{\{z_1 - z_0, \cdots, z_{2N-1} - z_0, x_N - z_0\} \cup \{u_j\}_{j\in I_{2N}}\right\}.$$

Here, we need to verify that the dimension is large enough to choose such orthonormal set of vectors. More precisely, we have to check

$$\dim S_{i-1}^\perp \geq |I_i \setminus I_{i-1}| \text{ for } i \in [1:2N] \text{ and } \dim S_{2N}^\perp \geq d - |I_{2N}|,$$

which is true from our assumption $d' \geq d + 2N$. To be more precise,

$$\dim S_{i-1}^\perp = d' - (i-1) - |I_{i-1}| \geq d' - 2N - |I_{i-1}| \geq d - |I_{i-1}| \geq |I_i \setminus I_{i-1}|$$

and

$$\dim S_{2N}^\perp = d' - 2N - |I_{2N}| \geq d - |I_{2N}|.$$

Then $\{u_i\}_{i=0}^{d-1}$ are orthonormal and $\langle u_j, z_i - z_0\rangle = 0$ for $j \notin S_i$. Therefore $U = [u_0 \mid \cdots \mid u_{d-1}]$ satisfies the statement of the lemma. □

Now we prove Theorem 4.

Proof of Theorem 4 Take (f, h) to be tuple of L-smooth convex function on $\mathbb{R}^{N+1}$ and an indicator function of nonempty closed convex set on $\mathbb{R}^{N+1}$, which is defined in Theorem 2. Let $\tilde{x}_\star \in \mathrm{argmin}_{x\in\mathbb{R}^{N+1}}(f+h)$. By lemma 40, there exists an orthogonal matrix $U \in \mathbb{R}^{d\times(N+1)}$ with $d \geq 2N + N + 1 = 3N + 1$ and $\mathbf{A}(z_0, (f_U, h_U)) = \{\{(z_i, \delta_i, \gamma_i)\}_{i\in[0:2N-1]}, x_N\}$ such that if $\tilde{z}_i \triangleq U^\intercal(z_i - z_0) \in \mathbb{R}^d$, then $\{\{(\tilde{z}_i, \delta_i, \gamma_i)\}_{i\in[0:2N-1]}, \tilde{x}_N\}$ is double-function zero-respecting with respect to (f, h). Then by Lemma 39, $\{\{(\tilde{z}_i, \delta_i, \gamma_i)\}_{i\in[0:2N-1]}, \tilde{x}_N\}$ satisfies the double-function span condition. Therefore, by Theorem 2,

$$f(\tilde{x}_N) + h(\tilde{x}_N) - f(\tilde{x}_\star) - h(\tilde{x}_\star) \geq \frac{L\|0 - \tilde{x}_\star\|^2}{2(\theta_N^2 - 1)}.$$

Note that $x_\star \triangleq U\tilde{x}_\star + x_0 \in \mathrm{argmin}_{x\in\mathbb{R}^d}(f_U + h_U)$ by Lemma 38. Then we finally have

$$\begin{aligned} f_U(x_N) + h_U(x_N) - f_U(x_\star) - h_U(x_\star) &= f_U(x_N) + h_U(x_N) - f_U(U\tilde{x}_\star + x_0) - h_U(U\tilde{x}_\star + x_0) \\ &= f(\tilde{x}_N) + h(\tilde{x}_N) - f(\tilde{x}_\star) - h(\tilde{x}_\star) \\ &\geq \frac{L\|0 - \tilde{x}_\star\|^2}{2(\theta_N^2 - 1)} = \frac{L\|U\tilde{x}_\star\|^2}{2(\theta_N^2 - 1)} = \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2 - 1)}. \end{aligned}$$

Therefore, f_U and h_U are our desired functions. □

F Omitted proof of Theorem 5

We now make formal definition of deterministic N-step proximal method.

We assume proximal stepsizes $\{\gamma_i\}_{i\in[0:N-1]}$ are given and fixed. A *deterministic N-step proximal method* $\mathbf{A}$ is a mapping from initial point x_0 and a function h to $\mathbf{A}(x_0,h)=\{x_i\}_{i\in[0:N]}$, and we call x_N the *approximate solution*, or simply the *output*. The sequence $\mathbf{A}(x_0,h)=\{x_i\}_{i\in[0:N]}$ depends on h only through N queries to proximal oracle $\mathcal{O}_h(x_i,\gamma_i)=(\mathbf{prox}_{\gamma_i h}(x_i),h(x_i))$. Here, x_i denotes the next query point to an oracle. More precisely, for $i\in[1:N]$,

$$x_i=\mathbf{A}_i(x_0,h)=\mathbf{A}(\{x_j\}_{j=0}^{i-1},\mathcal{O}_h(x_0,\gamma_0),\dots,\mathcal{O}_h(x_{j-1},\gamma_{j-1})).$$

Similar to the previous section, we expand the result of Therorem 3 to any deterministic N-step proximal methods by using resisting oracle technique [54,12].

We use *proximal zero-respecting* sequences, which is similar but slightly general than the proximal span condition. The sequence $\{x_i\}_{i\in[0:N]}$ is said to be *proximal zero-respecting* with respect to h if

$$\operatorname{supp}\{x_i\}\subseteq\operatorname{supp}\{x_0-\mathbf{prox}_{\gamma_0 h}(x_0),\dots,x_{i-1}-\mathbf{prox}_{\gamma_{i-1}h}(x_{i-1})\}\qquad\text{for } i=1,\dots,N.$$

where $\operatorname{supp}\{x\}\triangleq\{i\in[0:d-1]\,|\,\langle e_i,x\rangle\neq 0\}$. Note that by definition, $x_0=0$ for any proximal zero-respecting sequences.

The following lemmas are building blocks for proximal version of resisting oracle technique.

Lemma 41 *For any orthogonal matrix $U\in\mathbb{R}^{d'\times d}$ with $d'\geq d$ and any vector $x_0\in\mathbb{R}^{d'}$, if $h\colon\mathbb{R}^d\to\mathbb{R}$ is convex, then $h_U\colon\mathbb{R}^{d'}\to\mathbb{R}$ defined by*

$$h_U(x)=h(U^\intercal(x-x_0))$$

is convex.

Proof Suppose $x,y\in\mathbb{R}^{d'}$. Then for $t\in[0,1]$,

$$\begin{aligned}h_U\left((1-t)\,x+ty\right)&=h\left(U^\intercal\left((1-t)\,x+ty-x_0\right)\right)\\&=h\left((1-t)\,U^\intercal\left(x-x_0\right)+tU^\intercal\left(y-x_0\right)\right)\\&\leq(1-t)h\left(U^\intercal\left(x-x_0\right)\right)+th\left(U^\intercal\left(y-x_0\right)\right)\\&=(1-t)h_U(x)+th_U(y).\end{aligned}$$

where the inequality is from convexity of h. Therefore, h_U is convex. □

Lemma 42 *Suppose $h\colon\mathbb{R}^d\to\mathbb{R}$ is convex. Let h_U be as defined in lemma 41. If $\tilde{x}_\star\in\operatorname{argmin}_{x\in\mathbb{R}^d}h$ exists, then $x^\star\triangleq U\tilde{x}_\star+x_0\in\operatorname{argmin}_{x\in\mathbb{R}^{d'}}h_U$.*

Proof For any $x\in\mathbb{R}^{d'}$,

$$h_U(x_\star)=h(U^\intercal(U\tilde{x}_\star+x_0-x_0))=h(\tilde{x}_\star)\leq h(U^\intercal(x-x_0))=h_U(x),$$

where the inequality is from $\tilde{x}_\star\in\operatorname{argmin}_{\mathbb{R}^d}h$. □

Lemma 43 *Take $h\colon\mathbb{R}^d\to\mathbb{R}$ to be the function defined in Theorem 3. If $\{x_i\}_{i\in[0:N]}$ is proximal zero-respecting with respect to h, then it satisfies the proximal span condition* (16).

Proof Assume $\{x_i\}_{i\in[0:N]}$ is proximal zero-respecting with respect to h. Then, we have $x_0=0$ by the double-function zero-respecting assumption. From this, we have $d_0\in\operatorname{span}\{e_0\}$ by the property of h in Theorem 3 when $x_0=0$. Now suppose $x_i-\mathbf{prox}_{\gamma_i h}(x_i)\in\operatorname{span}\{e_0,\dots,e_i\}$ for all $0\leq i<k$ where $0<k\leq N-1$. Then, $\operatorname{supp}\{x_k\}\subseteq\operatorname{supp}\{x_0-\mathbf{prox}_{\gamma_0 h}(x_0),\dots,x_{k-1}-\mathbf{prox}_{\gamma_{k-1}h}(x_{k-1})\}\subseteq\{0,\dots,k-1\}$ by the proximal zero-respecting assumption. By the property of h, we then have $x_k-\mathbf{prox}_{\gamma_k h}(x_k)\in\{e_0,\dots,e_k\}$. Therefore, $x_i-\mathbf{prox}_{\gamma_i h}(x_i)\in\operatorname{span}\{e_0,\dots,e_i\}$ holds for all $i\in[0:N-1]$. This proves the equivalence of proximal span condition and proximal zero-respecting sequence, under the choice of h from Theorem 3. □

Lemma 44 *Assume $N>0$, and $d'\geq d+N$. Let $\mathbf{A}$ be any deterministic N-step proximal method, $x_0\in\mathbb{R}^{d'}$ be any vector, $h\colon\mathbb{R}^d\to\mathbb{R}$ is convex. Then there exists an orthogonal matrix $U\in\mathbb{R}^{d'\times d}$ and $\mathbf{A}(x_0,h_U)=\{x_i\}_{i\in[0:N]}$ such that if $\tilde{x}_i\triangleq U^\intercal(x_i-x_0)\in\mathbb{R}^d$, then $\{\tilde{x}_i\}_{i\in[0:N]}$ is proximal zero-respecting with respect to h.*

Proof Recall the support condition of proximal zero-respecting sequences.

$$\operatorname{supp}\{x_i\}\subseteq\operatorname{supp}\left\{x_0-\mathbf{prox}_{\gamma_0 h}(x_0),\dots,x_{i-1}-\mathbf{prox}_{\gamma_{i-1}h}(x_{i-1})\right\}.$$

For $i\in[0:N]$, let $I_i\subseteq[0:d-1]$ be

$$I_i=\operatorname{supp}\left\{\tilde{x}_0-\mathbf{prox}_{\gamma_0 h}(\tilde{x}_0),\dots,\tilde{x}_{i-1}-\mathbf{prox}_{\gamma_{i-1}h}(\tilde{x}_{i-1})\right\}.$$

Then $\{\tilde{x}_i\}_{i\in[0:N]}$ being proximal zero-respecting with respect to h is equivalent to

$$\mathrm{supp}\{\tilde{x}_i\} \subseteq I_i,\ i \in [0:N].$$

Also note that

$$\emptyset = I_0 \subseteq I_1 \subseteq \cdots \subseteq I_{N-1} \subseteq I_N \subseteq [0:d-1].$$

Now we construct the matrix $U = [u_0 \mid \cdots \mid u_{d-1}]$ by choosing appropriate $u_i \in \mathbb{R}^{d'}$ inductively. For $i = 0$, $\tilde{x}_0 = 0$ is clear from the definition. Since $\tilde{x}_i = U^\intercal(x_i - z_0)$, $\mathrm{supp}\{\tilde{x}_i\} \subseteq I_i$ is again equivalent to

$$\langle u_j, x_i - x_0\rangle = 0 \quad \forall j \notin I_i,\ i \in [0:N].$$

Now for $i \in [1:N]$, choose $\{u_j\}_{j\in I_i\setminus I_{i-1}}$ from the orthogonal complement of

$$S_{i-1} \triangleq \mathrm{span}\left\{\{x_1 - x_0, \cdots, x_{i-1} - x_0\} \cup \{u_j\}_{j\in I_{i-1}}\right\},$$

so that they are mutually orthonormal. At the last step, choose $\{u_j\}_{j\in[0:d-1]\setminus I_N}$ from the orthogonal complement of

$$S_N \triangleq \mathrm{span}\left\{\{x_1 - x_0, \cdots, x_N - x_0\} \cup \{u_j\}_{j\in I_N}\right\}.$$

Here, we need to verify that the dimension is large enough to choose such orthonormal set of vectors. More precisely, we have to check

$$\mathrm{dim}S_{i-1}^\perp \geq |I_i \setminus I_{i-1}| \text{ for } i \in [1:N],$$

which is true from our assumption $d' \geq d + N$. To be more precise,

$$\mathrm{dim}S_{i-1}^\perp = d' - (i-1) - |I_{i-1}| \geq d' - N - |I_{i-1}| \geq d - |I_{i-1}| \geq |I_i \setminus I_{i-1}|,$$

$$\mathrm{dim}S_N^\perp = d' - N - |I_{2N}| \geq d - |I_{2N}|.$$

Then $\{u_i\}_{i=0}^{d-1}$ are orthonormal and $\langle u_j, y_i - y_0\rangle = 0$ for $j \notin S_i$. Therefore $U = [u_0 \mid \cdots \mid u_{d-1}]$ satisfies the statement of the lemma. □

Now we prove Theorem 5.

Proof of Theorem 5. Take h be closed, convex, and proper function on $\mathbb{R}^{N+1}$ defined in Theorem 3. Let $\tilde{x}_\star \in \mathrm{argmin}_{x\in\mathbb{R}^{N+1}} h$. By lemma 44, there exists an orthogonal matrix $U \in \mathbb{R}^{d\times(N+1)}$ with $d \geq N + N + 1 = 2N + 1$ and $\mathbf{A}(x_0, h_U) = \{x_i\}_{i\in[0:N]}$ such that if $\tilde{x}_i \triangleq U^\intercal(x_i - x_0) \in \mathbb{R}^d$, then $\{x_i\}_{i\in[0:N]}$ is proximal zero-respecting with respect to h. Then by Lemma 43, $\{x_i\}_{i\in[0:N]}$ satisfies the proximal span condition. Therefore by Theorem 3,

$$h(\tilde{x}_N) - h(\tilde{x}_\star) \geq \frac{\gamma_{N-1}\|0 - \tilde{x}_\star\|^2}{4\gamma_0^2\eta_{N-1}^2} - \varepsilon.$$

Note that $x_\star \triangleq U\tilde{x}_\star + x_0 \in \mathrm{argmin}_{x\in\mathbb{R}^d} h_U$ by Lemma 42. Then we finally have

$$\begin{aligned} h_U(x_N) - h_U(x_\star) &= h_U(x_N) - h_U(U\tilde{x}_\star + x_0) \\ &= h(\tilde{x}_N) - h(\tilde{x}_\star) \\ &\geq \frac{\gamma_{N-1}\|0 - \tilde{x}_\star\|^2}{4\gamma_0^2\eta_{N-1}^2} - \varepsilon \\ &= \frac{\gamma_{N-1}\|U\tilde{x}_\star\|^2}{4\gamma_0^2\eta_{N-1}^2} - \varepsilon \\ &= \frac{\gamma_{N-1}\|x_0 - x_\star\|^2}{4\gamma_0^2\eta_{N-1}^2} - \varepsilon. \end{aligned}$$

Therefore, h_U is our desired function. □

References

1. Achterberg, T., Towle, E.: Non-Convex Quadratic Optimization: Gurobi 9.0 (2020). https://www.gurobi.com/resource/non-convex-quadratic-optimization/
2. Barré, M., Taylor, A.B., Bach, F.: Principled analyses and design of first-order methods with inexact proximal operators. Mathematical Programming **201**, 185–230 (2023)
3. Barré, M., Taylor, A.B., d'Aspremont, A.: Complexity guarantees for Polyak steps with momentum. Conference on Learning Theory (2020)
4. Beck, A.: First-Order Methods in Optimization. SIAM (2017)
5. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Transactions on Image Processing **18**(11), 2419–2434 (2009)
6. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009)
7. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3**(1), 1–122 (2011)
8. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
9. Braun, G., Carderera, A., Combettes, C.W., Hassani, H., Karbasi, A., Mokhtari, A., Pokutta, S.: Conditional gradient methods. arXiv:2211.14103 (2022). URL https://conditional-gradients.org/
10. Brézis, H., Lions, P.L.: Produits infinis de résolvantes. Israel Journal of Mathematics **29**, 329–345 (1978)
11. Bruck, R.E.: 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)
12. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points I. Mathematical Programming **184**, 71–120 (2020)
13. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points ii: first-order methods. Mathematical Programming **185**(1-2), 315–355 (2021)
14. Clarke, F.: Optimization and Nonsmooth Analysis. Wiley New York (1983)
15. Das Gupta, S., Freund, R.M., Sun, X.A., Taylor, A.: Nonlinear conjugate gradient methods: worst-case convergence rates via computer-assisted analyses. Mathematical Programming pp. 1–49 (2024)
16. Das Gupta, S., Van Parys, B.P., Ryu, E.K.: Branch-and-bound performance estimation programming: A unified methodology for constructing optimal optimization methods. Mathematical Programming **204**(1), 567–639 (2024)
17. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences **57**(11), 1413–1457 (2004)
18. De Klerk, E., Glineur, F., Taylor, A.B.: On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. Optimization Letters **11**, 1185–1199 (2017)
19. De Klerk, E., Glineur, F., Taylor, A.B.: Worst-case convergence analysis of inexact gradient and newton methods through semidefinite programming performance estimation. SIAM Journal on Optimization **30**(3), 2053–2082 (2020)
20. Diakonikolas, J., Wang, P.: Potential function-based framework for making the gradients small in convex and min-max optimization. SIAM Journal on Optimization **32**(3), 1668–1697 (2022)
21. Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two and three space variables. Transactions of the American Mathematical Society **82**(2), 421–439 (1956)
22. Dragomir, R.A., Taylor, A.B., d'Aspremont, A., Bolte, J.: Optimal complexity and certification of bregman first-order methods. Mathematical Programming **194**, 41–83 (2022)
23. Drori, Y.: The exact information-based complexity of smooth convex minimization. Journal of Complexity **39**, 1–16 (2017)
24. Drori, Y., Shamir, O.: The complexity of finding stationary points with stochastic gradient descent. International Conference on Machine Learning (2020)
25. Drori, Y., Taylor, A.B.: On the oracle complexity of smooth strongly convex minimization. Journal of Complexity **68**, 101590 (2022)
26. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: A novel approach. Mathematical Programming **145**(1-2), 451–482 (2014)
27. Drori, Y., Teboulle, M.: An optimal variant of Kelley's cutting-plane method. Mathematical Programming **160**, 321–351 (2016)
28. Elad, M.: Why simple shrinkage is still relevant for redundant representations? IEEE Transactions on Information Theory **52**(12), 5559–5569 (2006)
29. Fortin, M., Glowinski, R.: On decomposition-coordination methods using an augmented Lagrangian. In: M. Fortin, R. Glowinski (eds.) Studies in Mathematics and Its Applications, vol. 15, pp. 97–146. Elsevier (1983)
30. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Research Logistics Quarterly **3**(1–2), 95–110 (1956)
31. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Computers and Mathematics with Applications **2**(1), 17–40 (1976)
32. Garrigos, G., Gower, R.M.: Handbook of convergence theorems for (stochastic) gradient methods. arXiv:2301.11235 (2023)
33. Glowinski, R., Marrocco, A.: Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. Revue Française d'Automatique, Informatique, Recherche Opérationnelle. Analyse Numérique **9**(2), 41–76 (1975)

34. Güler, O.: On the convergence of the proximal point algorithm for convex minimization. SIAM Journal on Control and Optimization **29**(2), 403–419 (1991)
35. Güler, O.: New proximal point algorithms for convex minimization. SIAM Journal on Optimization **2**(4), 649–664 (1992)
36. Hale, E.T., Yin, W., Zhang, Y.: Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. SIAM Journal on Optimization **19**(3), 1107–1130 (2008)
37. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press (2012)
38. Horst, R., Tuy, H.: Global Optimization: Deterministic Approaches. Springer Science & Business Media (2013)
39. Kim, D.: Accelerated proximal point method for maximally monotone operators. Mathematical Programming **190**(1–2), 57–87 (2021)
40. Kim, D., Fessler, J.A.: Optimized first-order methods for smooth convex minimization. Mathematical Programming **159**(1), 81–107 (2016)
41. Kim, D., Fessler, J.A.: On the convergence analysis of the optimized gradient method. Journal of Optimization Theory and Applications **172**(1), 187–205 (2017)
42. Kim, D., Fessler, J.A.: Another look at the fast iterative shrinkage/thresholding algorithm (FISTA). SIAM Journal on Optimization **28**(1), 223–250 (2018)
43. Kim, D., Fessler, J.A.: 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)
44. Kim, J., Ozdaglar, A., C.Park, Ryu, E.K.: Time-reversed dissipation induces duality between minimizing gradient norm and function value. Neural Information Processing Systems (2023)
45. Lee, J., Park, C., Ryu, E.K.: A geometric structure of acceleration and its role in making gradients small fast. Neural Information Processing Systems (2021)
46. Lieder, F.: On the convergence rate of the halpern-iteration. Optimization letters **15**(2), 405–418 (2021)
47. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM Journal on Numerical Analysis **16**(6), 964–979 (1979)
48. Locatelli, M., Schoen, F.: Global Optimization: Theory, Algorithms, and Applications. SIAM (2013)
49. Malitsky, Y., Mishchenko, K.: Adaptive gradient descent without descent. International Conference on Machine Learning (2020)
50. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. Recherche Opérationnelle **4**, 154–158 (1970)
51. Martinet, B.: Détermination approchée d'un point fixe d'une application pseudo-contractante. CR Acad. Sci. Paris **274**(2), 163–165 (1972)
52. Monteiro, R.D.C., Svaiter, B.F.: An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. SIAM Journal on Optimization **23**, 1092–1125 (2013)
53. Nemirovski, A.: Information-based complexity of convex programming. Lecture Notes, Technion - Israel Institute of Technology (1995)
54. Nemirovski, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience (1983)
55. Nemirovsky, A.: Information-based complexity of linear operator equations. Journal of Complexity **8**(2), 153–175 (1992)
56. Nemirovsky, A.S.: On optimality of Krylov's information when solving linear operator equations. Journal of Complexity **7**(2), 121–130 (1991)
57. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady **27**(2), 372–376 (1983)
58. Nesterov, Y.: Lectures on Convex Optimization, 2nd edn. Springer (2018)
59. Park, C., Park, J., Ryu, E.K.: Factor-$\sqrt{2}$ acceleration of accelerated gradient methods. Applied Mathematics & Optimization **88**(3), 77 (2023)
60. Park, C., Ryu, E.K.: Optimal first-order algorithms as a function of inequalities. Journal of Machine Learning Research **25**, 1–69 (2024)
61. Park, J., Ryu, E.K.: Exact optimal accelerated complexity for fixed-point iterations. International Conference on Machine Learning (2022)
62. Passty, G.B.: 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)
63. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM Journal on Control and Optimization **14**(5), 877–898 (1976)
64. Ryu, E.K., Taylor, A.B., Bergeling, C., Giselsson, P.: Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. SIAM Journal on Optimization **30**(3), 2251–2271 (2020)
65. Ryu, E.K., Yin, W.: Large-Scale Convex Optimization via Monotone Operators. Cambridge University Press (2022)
66. Taylor, A.B.: Convex interpolation and performance estimation of first-order methods for convex optimization. Ph.D. thesis, Catholic University of Louvain, Louvain-la-Neuve, Belgium (2017)
67. Taylor, A.B., Drori, Y.: An optimal gradient method for smooth strongly convex minimization. Mathematical Programming **199**, 557–594 (2023)
68. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Exact worst-case performance of first-order methods for composite convex optimization. SIAM Journal on Optimization **27**(3), 1283–1313 (2017)
69. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Mathematical Programming **161**(1–2), 307–345 (2017)

70. Taylor, A.B., Van Scoy, B., Lessard, L.: Lyapunov functions for first-order methods: Tight automated convergence guarantees. International Conference on Machine Learning (2018)
71. Teboulle, M., Vaisbourd, Y.: An elementary approach to tight worst case complexity analysis of gradient based methods. Mathematical Programming **201**(1), 63–96 (2023)
72. Wright, S.J., Nowak, R.D., Figueiredo, M.A.: Sparse reconstruction by separable approximation. IEEE Transactions on Signal Processing **57**(7), 2479–2493 (2009)
73. Yoon, T., Ryu, E.K.: Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. International Conference on Machine Learning (2021)