Journal of Machine Learning Research 25 (2024) 1-69 Submitted 10/21; Revised 3/23; Published 2/24

Optimal First-Order Algorithms as a Function of Inequalities

Chanwoo Park cpark97@mit.edu
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Massachusetts, United States of America

Ernest K. Ryu eryu@snu.ac.kr
Department of Mathematical Sciences
Interdisciplinary Program in Artificial Intelligence
Seoul National University
Seoul, Korea

Editor: Martin Jaggi

Abstract

In this work, we present a novel algorithm design methodology that finds the optimal algorithm as a function of inequalities. Specifically, we restrict convergence analyses of algorithms to use a prespecified subset of inequalities, rather than utilizing all true inequalities, and find the optimal algorithm subject to this restriction. This methodology allows us to design algorithms with certain desired characteristics. As concrete demonstrations of this methodology, we find new state-of-the-art accelerated first-order gradient methods using randomized coordinate updates and backtracking line searches.

1. Introduction

Nesterov's seminal work presented the fast gradient method (FGM) with rate $\mathcal{O}(1/k^2)$ (Nesterov, 1983), and Nemirovsky and Yudin established a complexity lower bound matching the rate up to a constant (Nemirovsky and Yudin, 1983). A rich line of research following these footsteps flourished in the following decades, and FGM became the prototypical "optimal" method. Recently, however, it was discovered that the Nesterov's FGM can be improved by a constant; the optimized gradient method (OGM) (Drori and Teboulle, 2014; Kim and Fessler, 2016) outperforms FGM by a factor of 2. Furthermore, the prior complexity lower bound was also improved by a constant factor to exactly match the rate of OGM (Drori, 2017). Thus, the search for the exact optimal first-order gradient is now complete, and OGM, not FGM, emerges as the victor.

That FGM can be improved was, in our view, a surprising discovery, and it leads us to ask the following questions. First, can we also improve the variants of Nesterov's FGM in related setups? FGM's acceleration has been extended to utilize randomized coordinate updates (Nesterov, 2012; Allen-Zhu et al., 2016; Nesterov and Stich, 2017) and backtracking line searches (Beck and Teboulle, 2009). Second, is there some sense in which Nesterov's FGM is exactly optimal?

In this work, we address these two questions by examining the inequalities used in the analyses of the algorithms. We introduce the notion of $\mathcal{A}^\star$-optimality, which defines optimality of an algorithm *conditioned* on a set of inequalities. OGM is the $\mathcal{A}^\star$-optimal algorithm conditioned on all true inequalities and therefore is the exact optimal algorithm in the classical sense. However, other algorithms become $\mathcal{A}^\star$-optimal when conditioned on a different restrictive subset of inequalities. By restricting convergence analyses of algorithms to use only a prespecified subset of inequalities with good properties, rather than utilizing all true inequalities, we obtain algorithms with better capacity for extensions. Specifically, we obtain new algorithms utilizing randomized coordinate updates and backtracking linesearches that improve upon the prior state-of-the-art rates. Moreover, we show that FGM is the optimal algorithm roughly in the sense that it is the best algorithm that admits the use of randomized coordinate updates and backtracking line searches.

Contributions. As the main contribution of this work, we present an algorithm design methodology based on $\mathcal{A}^\star$-optimality and the performance estimation problem (PEP) (Drori and Teboulle, 2014; Taylor et al., 2017b) and demonstrate the strength of the methodology by finding new $\mathcal{A}^\star$-optimal algorithms that improve the state-of-the-art rates achieved by variants of FGM. As a minor contribution, we establish the optimality of FGM in the following sense: FGM is the $\mathcal{A}^\star$-optimal algorithm that relies on a certain set of inequalities that are amenable to both randomized coordinate updates and backtracking line searches.

1.1 Preliminaries and notations

In this section, we review standard definitions and set up the notation.

Problem setting and L-smoothness. For $L > 0$, $f\colon \mathbb{R}^n \to \mathbb{R}$ is L-smooth if f is differentiable and

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \qquad \forall\, x, y \in \mathbb{R}^n.$$

Throughout this paper, we consider the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x)$$

with the following assumptions

(A1) f is convex

(A2) f is L-smooth

(A3) f has a minimizer (not necessarily unique)

(A4) $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$.

We write $x_\star$ to denote a minimizer of f if one exists and $f_\star = \inf_{x\in\mathbb{R}^n} f(x)$ for the optimal value.

Fixed-step first-order algorithms (FSFO). The class of fixed-step first-order algorithms (FSFO) has the following form: given a differentiable f, total iteration count N, and starting point $x_0 \in \mathbb{R}^d$, the iterates are defined by

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

for $i = 0, 1, \dots, N-1$. The coefficients $\{h_{i,k}\}_{0\le k<i\le N}$ may depend on N and prior information about the function f, such as the smoothness coefficient L, but are otherwise predetermined. In particular, $\{h_{i,k}\}_{0\le k<i\le N}$ may not depend on function values or gradients observed throughout the algorithm. The classical algorithms such as FGM, OGM, and the heavy-ball method are all FSFO.

Nesterov's fast gradient method (FGM). The celebrated FGM is

$$\begin{aligned}
y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k)\\
z_{k+1} &= z_k - \frac{\theta_k}{L}\nabla f(x_k)\\
x_{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1},
\end{aligned}$$

where $z_0 = x_0$, $\theta_0 = 1$, and $\{\theta_k\}_{k=0}^{\infty}$ is defined as $\theta_{k+1} = \frac{1+\sqrt{4\theta_k^2+1}}{2}$ for $k = 0, 1, \dots$ (Nesterov, 1983). FGM has the rate

$$f(y_N) - f_\star \le \frac{2L\left\|x_0 - x_\star\right\|^2}{N^2} + o\left(\frac{1}{N^2}\right),$$

which is optimal *up to a constant*. Many extensions and variants of FGM have been presented, including versions utilizing randomized coordinate updates (Lee and Sidford, 2013; Allen-Zhu et al., 2016; Nesterov and Stich, 2017) and backtracking linesearches (Beck and Teboulle, 2009).

Optimized gradient method (OGM). Let N be the total iteration count. OGM (Drori and Teboulle, 2014; Kim and Fessler, 2016) is

$$\begin{aligned}
y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k)\\
z_{k+1} &= z_k - \frac{2\theta_k}{L}\nabla f(x_k)\\
x_{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1},
\end{aligned}$$

for $k = 0, 1, \dots, N-2$, where $z_0 = y_0 = x_0$ and θ_k is the same as with FGM. Different from FGM, OGM has what we refer to as the *last-step modification*

$$x_N = \left(1 - \frac{1}{\tilde{\theta}_N}\right) y_N + \frac{1}{\tilde{\theta}_N} z_N,$$

where $\{\tilde{\theta}_k\}_{k=1}^{\infty}$ is defined as $\tilde{\theta}_k = \frac{1+\sqrt{8\theta_{k-1}^2+1}}{2}$. OGM exhibits the rate

$$f(x_N) - f_\star \leq \frac{L\,\|x_0 - x_\star\|^2}{N^2} + o\left(\frac{1}{N^2}\right),$$

which is faster than FGM by a factor of 2 and is in fact exactly optimal (Drori, 2017). This remarkable discovery was made using a computer-assisted methodology, the performance estimation problem (PEP) (Drori and Teboulle, 2014; Kim and Fessler, 2016). Variants of OGM with randomized coordinate updates or backtracking linesearches had not been discovered.

Optimized gradient method - Gradient norm (OGM-G). Let N be the total iteration count. The method OGM-G has what we refer to as the *first-step modification*

$$\begin{aligned}
y_1 &= x_0 - \frac{1}{L}\nabla f(x_0) \\
z_1 &= z_0 - \frac{1+\tilde{\theta}_N}{2L}\nabla f(x_0) \\
x_1 &= \frac{\theta_{N-1}^4}{\tilde{\theta}_N^4} y_1 + \left(1 - \frac{\theta_{N-1}^4}{\tilde{\theta}_N^4}\right) z_1,
\end{aligned}$$

where $z_0 = y_0 = x_0$ and $\{\theta_k\}_{k=0}^{\infty}$, $\{\tilde{\theta}_k\}_{k=1}^{\infty}$ are defined as OGM. The remaining iterates of OGM-G are defined as

$$\begin{aligned}
y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\
z_{k+1} &= z_k - \frac{\theta_{N-k+1}}{L}\nabla f(x_k) \\
x_{k+1} &= \frac{\theta_{N-k-1}^4}{\theta_{N-k}^4} y_{k+1} + \left(1 - \frac{\theta_{N-k-1}^4}{\theta_{N-k}^4}\right) z_{k+1},
\end{aligned}$$

for $k = 1, 2, \dots, N-1$ (Kim and Fessler, 2021; Lee et al., 2021). Note that the indices of the θ-coefficients are decreasing as the iteration count increases. Different to FGM and OGM, the guarantee of OGM-G is on the gradient magnitude:

$$\|\nabla f(x_N)\|^2 \leq \mathcal{O}\left(\frac{L\,(f(x_0) - f_\star)}{N^2}\right).$$

An important use of OGM-G is that when combined with FGM (or OGM), one can achieve the rate (Nesterov et al., 2020, Remark 2.1)

$$\|\nabla f(x_N)\|^2 \leq \mathcal{O}\left(\frac{L^2\|x_0 - x_\star\|^2}{N^4}\right).$$

Recently, analyses of OGM-G based on potential function approaches have been presented (Lee et al., 2021; Diakonikolas and Wang, 2022).

Coordinate-wise smoothness. We say $f\colon \mathbb{R}^n \to \mathbb{R}$ is coordinate-wise smooth with parameters $(L_1, \dots, L_n)$ if it is differentiable and

$$\|\nabla_i f(x + \delta e_i) - \nabla_i f(x)\| \leq L_i \delta \tag{2}$$

for all $x \in \mathbb{R}^n$, $\delta > 0$, and $i = 1, \dots, n$, where $\nabla_i f = \frac{\partial f}{\partial x_i} e_i$ is the i-th partial derivative vector and e_i is the i-th unit vector for $i = 1, \dots, n$.

Fast gradient method - randomized coordinate updates (FGM-RC). There are several randomized coordinate update variants of Nesterov's FGM (Nesterov, 2012; Allen-Zhu et al., 2016; Nesterov and Stich, 2017). We discuss the version of Allen-Zhu et al. (2016), which we call FGM-RC, as it has the smallest (best) constant. Assume $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex and coordinate-wise smooth with parameters $(L_1, L_2, \dots, L_n)$. FGM-RC is

$$\begin{aligned}
&\text{Sample } i(k) \text{ from } \{1, 2, \dots, n\} \text{ with } \mathbb{P}(i(k) = t) = \frac{\sqrt{L_t}}{S} \\
&y_{k+1} = x_k - \frac{1}{L_{i(k)}} \nabla_{i(k)} \nabla f(x_k) \\
&z_{k+1} = z_k - \frac{k+2}{2S^2} \frac{1}{p_{i(k)}} \nabla_{i(k)} f(x_k) \\
&x_{k+1} = \frac{k+1}{k+3} y_{k+1} + \frac{2}{k+3} z_{k+1}
\end{aligned}$$

for $k = 0, 1, \dots$, where $z_0 = y_0 = x_0$, $p_t = \mathbb{P}(i = t)$, and $S = \sum_{k=1}^n \sqrt{L_k}$. FGM-RC exhibits the rate

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

While Allen-Zhu et al.'s FGM-RC uses convenient rational coefficients, their algorithm can be slightly sharpened (through straightforward modifications of their presented analysis) to use the θ-coefficients of Nesterov. We refer to this refinement as FGM-RC$^\sharp$:

$$\begin{aligned}
&\text{Sample } i(k) \text{ from } \{1, 2, \dots, n\} \text{ with } \mathbb{P}(i(k) = t) = \frac{\sqrt{L_t}}{S} \\
&y_{k+1} = x_k - \frac{1}{L_{i(k)}} \nabla_{i(k)} f(x_k) \\
&z_{k+1} = z_k - \frac{\theta_k}{S^2} \frac{1}{p_{i(k)}} \nabla_{i(k)} f(x_k) \\
&x_{k+1} = \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1}
\end{aligned}$$

for $k = 0, 1, \dots$, where $z_0 = y_0 = x_0$ and $S = \sum_{k=1}^n \sqrt{L_k}$. FGM-RC$^\sharp$ exhibits the rate

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

Compared to first-order methods utilizing the full gradient, randomized coordinate updates methods have a lower cost per iteration and can be significantly faster. In particular, FGM-RC$^\sharp$ can be significantly faster than FGM or OGM.

Fast gradient method - backtracking linesearch (FGM-BL). Beck and Teboulle (2009) provides a version of Nesterov's FGM that uses backtracking linesearches (FGM-BL). Define $z_0 = y_0 = x_0$, $\eta > 1$, and $L_0 > 0$. Consider the backtracking linesearch that finds the smallest nonnegative integer i_k such that with $\bar{L} = \eta^{i_k} L_{k-1}$

$$f\left(x_k - \frac{1}{\bar{L}}\nabla f(x_k)\right) \leq f(x_k) - \frac{1}{2\bar{L}}\left\|\nabla f(x_k)\right\|^2$$

holds, for each step k. In FGM-BL, we set $L_k = \eta^{i_k} L_{k-1}$ and define

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L_k}\nabla f(x_k) \\ z_{k+1} &= z_k - \frac{\theta_k}{L_k}\nabla f(x_k) \\ x_{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1} \end{aligned}$$

for $k = 0, 1, \ldots$. FGM-BL exhibits the rate

$$f(y_N) - f_\star \leq \frac{L_N}{2\theta_{N-1}^2}\left\|x_0 - x_\star\right\|^2.$$

The backtracking linesearch is useful when we do not know the smoothness parameter L. FGM-BL obtains an estimate of L while making progress with the accelerated gradient method.

Computer-assisted algorithm design. The performance estimation problem (PEP) is a computer-assisted proof methodology that analyzes the worst-case performance of optimization algorithms through semidefinite programs (Drori and Teboulle, 2014; Taylor et al., 2017b,a). The use of the PEP has led to many discoveries that would have otherwise been difficult without the assistance (Kim and Fessler, 2018a; Taylor et al., 2018; Taylor and Bach, 2019; **?**; De Klerk et al., 2020; Gu and Yang, 2020; Lieder, 2021; Ryu et al., 2020; **?**; Kim, 2021; Yoon and Ryu, 2021). Notably, the algorithms OGM (Drori and Teboulle, 2014; Kim and Fessler, 2016, 2018b), OGM-G (Kim and Fessler, 2021), and ITEM (Taylor and Drori, 2023) were obtained by using the PEP for the setup of minimizing a smooth convex (possibly strongly convex) function. OGM and ITEM improve the rates of Nesterov's FGM by constants and have an exact matching complexity lower bound (Drori, 2017; Drori and Taylor, 2022). The integral quadratic constraints (IQC) is another technique based on control-theoretic notions for computer-assisted algorithm design (Lessard et al., 2016; Hu and Lessard, 2017; Van Scoy et al., 2017; Fazlyab et al., 2018; Seidman et al., 2019; Zhang et al., 2021). This work builds upon the PEP methodology.

1.2 Organization

This paper is organized as follows. Section 2 defines the notion of handy inequalities and discusses how it will be utilized for designing accelerated algorithms with randomized coordinate updates and backtracking linesearches. Section 3 defines the notion of $\mathcal{A}^\star$-optimality. Section 4 exhibits the main methodology of this work by using it to obtain an $\mathcal{A}^\star$-optimal algorithm and a variant utilizing randomized coordinate updates. Section 5 presents several other $\mathcal{A}^\star$-optimal algorithms obtained with our algorithm design methodology. The proofs of $\mathcal{A}^\star$-optimality of the algorithms of Section 5 are deferred to Sections A and B of the appendix.

As the proofs of $\mathcal{A}^\star$-optimality require lengthy calculations, we provide Matlab scripts verifying them. Specifically, the following scripts show that the derived analytical results agree with the numerical solutions of the SDPs:
`https://github.com/chanwoo-park-official/A-star-map/`.

2. Handy inequalities for deriving variants of FSOM

In this section, we define the notion of *handy inequalities*. The definition captures the empirical observation that the use of some inequalities makes the algorithm more amenable to modifications (and are therefore “handy”) while other inequalities make the analysis brittle and not amenable to modifications. In particular, FGM has extensions using randomized coordinate updates and backtracking linesearches, as discussed in the preliminaries, while such modifications seem difficult with OGM. By identifying the notion of handy inequalities, we point out that the fault is in the inequalities being utilized.

2.1 Inequalities for smooth convex functions

In this section, we quickly review and name a few commonly used inequalities for smooth convex functions.

If $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex, L-smoothness is equivalent to

$$f(x) \geq f(y) + \langle \nabla f(y), x-y\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \tag{3}$$

for all $x, y \in \mathbb{R}^n$. We call (3) the *cocoercivity inequality on* (x, y). As a particular case, when $y = x - \frac{1}{L}\nabla f(x)$, the cocoercivity inequality on (x, y) becomes

$$f(x) \geq f(y) + \frac{1}{2L}\left\|\nabla f(x)\right\|^2 + \frac{1}{2L}\|\nabla f(y)\|^2,$$

and dropping the last term leads to

$$f(x) \geq f(y) + \frac{1}{2L}\left\|\nabla f(x)\right\|^2.$$

We call this the *gradient-step inequality at* x. Dropping $\frac{1}{2L}\left\|\nabla f(x) - \nabla f(y)\right\|^2$ in (3), we get

$$f(x) \geq f(y) + \langle \nabla f(y), x-y\rangle.$$

We call this the *convexity inequality on* (x, y). Note that the gradient-step or convexity inequalities are weaker than the cocoercivity inequality in the sense that they were obtained by dropping a nonnegative term from the cocercivity inequality.

If $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex and coordinate-wise smooth with parameters $(L_1, \dots, L_n)$, then

$$f(x) \geq f(y) + \langle \nabla f(y), x-y\rangle + \frac{1}{2L_i}\left\|\nabla_i f(x) - \nabla_i f(y)\right\|^2 \tag{4}$$

for all $x, y \in \mathbb{R}^n$ and $i = 1, \dots, n$. We call this the *coordinate-wise cocoercivity inequality on* (x, y, i). As a particular case, when $y = x - \frac{1}{L_i}\nabla_i f(x)$, the coordinate-wise cocoercivity inequality on (x, y, i) becomes

$$f(x) \geq f(y) + \frac{1}{2L_i} \|\nabla_i f(x)\|^2 + \frac{1}{2L_i}\|\nabla_i f(y)\|^2.$$

Dropping the last term, we get

$$f(x) \geq f(y) + \frac{1}{2L_i} \|\nabla_i f(x)\|^2 .$$

We call this the *coordinate-wise gradient-step inequality at* (x, i).

Proof of (4)**.** While we suspect the coordinate-wise cocoercivity inequality to be commonly known, we are unaware of a written proof to reference. We therefore quickly provide the following proof.

Let g be convex, coordinate-wise smooth with parameters $(L_1, \dots, L_n)$, and $y = x - \frac{1}{L_i}\nabla_i g(x)$. Then, we have

$$\begin{aligned} g_\star - g(x) &\leq g(y) - g(x) \\ &= \int_{t=0}^{1} \langle \nabla g(x + t(y-x)), y - x\rangle dt \\ &= \langle \nabla g(x), y - x\rangle + \int_{t=0}^{1} \langle \nabla g(x + t(y-x)) - \nabla g(x), y - x\rangle dt \\ &= \langle \nabla g(x), y - x\rangle + \int_{t=0}^{1} (\nabla_i g(x + t(y_i - x_i)e_i) - \nabla_i g(x))(y_i - x_i) dt \\ &\leq \langle \nabla g(x), y - x\rangle + \int_{t=0}^{1} t L_i (y_i - x_i)^2 dt = -\frac{1}{2L_i} \|\nabla_i g(x)\|^2 . \end{aligned}$$

For all y, the function $f(x) - f(y) - \langle \nabla f(y), x - y\rangle$ is convex and coordinate-wise smooth with parameters $(L_1, \dots, L_n)$, as a function of x. Therefore, for all y, setting $g(x) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle$, we conclude

$$f(x) - f(y) - \langle \nabla f(y), x - y\rangle \geq \frac{1}{2L_i} \|\nabla_i f(x) - \nabla_i f(y)\|^2 .$$

2.2 The inequalities of FGM and OGM

The analyses of FGM and OGM crucially differ in the inequalities they use. The common convergence analysis of FGM defines the Lyapunov function

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

and establishes the non-increasing property

$$\begin{aligned} U_k - U_{k+1} = \theta_k^2 &\left(f(x_k) - f(y_{k+1}) - \frac{1}{2L} \|\nabla f(x_k)\|^2 \right) \\ &+ \theta_{k-1}^2 \left(f(y_k) - f(x_k) - \langle \nabla f(x_k), y_k - x_k\rangle \right) \\ &+ \theta_k \left(f_\star - f(x_k) - \langle \nabla f(x_k), x_\star - x_k\rangle \right) \\ &\geq 0. \end{aligned}$$

In contrast, the analysis of OGM defines the Lyapunov function

$$U_k = 2\theta_k^2\left(f(x_k) - f_\star - \frac{1}{2L}\|\nabla f(x_k)\|^2\right) + \frac{L}{2}\|z_{k+1} - x_\star\|^2$$

and establishes the non-increasing property

$$\begin{aligned} &U_k - U_{k+1} \\ &= 2\theta_k^2\left(f(x_k) - f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle - \frac{1}{2L}\|\nabla f(x_k) - \nabla f(x_{k+1})\|^2\right) \\ &+ 2\theta_{k+1}\left(f_\star - f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_{k+1} - x_\star\rangle - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2\right) \\ &\geq 0. \end{aligned}$$

Note the difference in the inequalities being used; FGM uses gradient-step and convexity inequalities, while OGM uses cocoercivity inequalities. The fact that OGM uses the stronger cocoercivity inequalities partially explains why its guarantee is stronger than FGM's guarantee. These inequalities are all, of course, true inequalities, but the inequalities used by FGM are much more amenable to obtaining variants using randomized coordinate updates and backtracking line searches; the inequalities used by OGM do not have this property. In Sections 2.3 and 2.4, we define inequalities that admit such variants to be algorithmically *handy*.

2.3 Handy inequalities for randomized coordinate updates

We now discuss the notion of handy inequalities for randomized coordinate updates. Let us examine the analysis of FGM-RC$^\sharp$. For $k = 0, 1, \ldots,$ define

$$U_k = \frac{\theta_{k-1}^2}{S^2}(f(y_k) - f_\star) + \frac{1}{2}\|z_k - x_\star\|^2$$

and write $\mathbb{E}_{i(k)}$ for the expectation conditioned on information up to the k-th iteration. Then,

$$\begin{aligned} U_k - U_{k+1} = &\frac{\theta_k^2}{S^2}\left(f(x_k) - f(y_{k+1}) - \frac{1}{2L_{i(k)}}\left\|\nabla_{i(k)} f(x_k)\right\|^2\right) \\ &+ \frac{\theta_{k-1}^2}{S^2}\left(f(y_k) - f(x_k) - \frac{S}{\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), y_k - x_k\rangle\right) \\ &+ \frac{\theta_k}{S^2}\left(f_\star - f(x_k) - \frac{S}{\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), x_\star - x_k\rangle\right), \end{aligned}$$

and taking the conditional expectation $\mathbb{E}_{i(k)}$ gives us

$$\begin{aligned} U_k - \mathbb{E}_{i(k)} U_{k+1} = \mathbb{E}_{i(k)} \left[\frac{\theta_k^2}{S^2} \left(f(x_k) - f(y_{k+1}) - \frac{1}{2L_{i(k)}} \left\| \nabla_{i(k)} f(x_k) \right\|^2 \right) \right] \\ + \frac{\theta_{k-1}^2}{S^2} \left(f(y_k) - f(x_k) - \langle \nabla f(x_k), y_k - x_k \rangle \right) \\ + \frac{\theta_k}{S^2} \left(f_\star - f(x_k) - \langle \nabla f(x_k), x_\star - x_k \rangle \right) \\ \geq 0. \end{aligned}$$

Finally, taking the full expectation gives us

$$\frac{\theta_k^2}{S^2} (\mathbb{E}[f(y_{k+1})] - f_\star) \leq \mathbb{E} U_{k+1} \leq \cdots \leq U_0 \leq \frac{1}{2} \left\| x_0 - x_\star \right\|^2 .$$

We can interpret this convergence analysis as a direct modification of FGM's analysis by taking expectations of the inequalities. Utilizing the linearity of expectation to obtain the convexity inequality and having the coordinate-wise gradient-step inequality holds almost surely is crucial. The gradient-step inequality and the convexity inequality are handy for randomized coordinate updates as this analysis of FGM-RC$^\sharp$ demonstrates. The coordinate-wise cocoercivity inequality on $(x_\star, x_k)$ is also handy as the term $\frac{1}{2L_{i(k)}} \left\| \nabla_{i(k)} f(x_k) \right\|^2$ is one we can take the expectation of.

On the other hand, the cocoercivity inequalities on (y_k, x_k) or (x_k, y_{k+1}) do not seem to be handy as the terms $\| \nabla_{i(k)} f(x_k) - \nabla f(y_k) \|^2$ or $\| \nabla_{i(k)} f(x_k) - \nabla f(y_{k+1}) \|^2$ are not easily manipulated under expectations. For this reason, adapting OGM and its analysis to use randomized coordinate updates seems difficult.

2.4 Handy inequalities for backtracking linesearch

Next, we discuss the notion of handy inequalities for backtracking linesearches. Let us examine the analysis of FGM-BL. For $k = 0, 1, \dots,$ define

$$U_{k,L} = \frac{\theta_{k-1}^2}{L} (f(y_k) - f_\star) + \frac{1}{2} \left\| z_k - x_\star \right\|^2 .$$

Then,

$$\begin{aligned} U_{k,L_{k+1}} - U_{k+1,L_{k+1}} = \frac{1}{L_{k+1}} \Bigg(\theta_k^2 \left(f(x_k) - f(y_{k+1}) - \frac{1}{2L_{k+1}} \left\| \nabla f(x_k) \right\|^2 \right) \\ + \theta_{k-1}^2 \left(f(y_k) - f(x_k) - \langle \nabla f(x_k), y_k - x_k \rangle \right) \\ + \theta_k \left(f_\star - f(x_k) - \langle \nabla f(x_k), x_\star - x_k \rangle \right) \Bigg) \\ \geq 0. \end{aligned}$$

Note that the inequality $f(x_k) - f(y_{k+1}) - \frac{1}{2L_{k+1}} \left\| \nabla f(x_k) \right\|^2 \geq 0$ is enforced by the backtracking linesearch. Finally, we conclude

$$\frac{\theta_k^2}{L_{k+1}} (f(y_{k+1}) - f_\star) \leq U_{k+1,L_{k+1}} \leq U_{k,L_{k+1}} \leq U_{k,L_k} \leq \cdots \leq U_{0,L_0} \leq \frac{1}{2} \left\| x_0 - x_\star \right\|^2 .$$

We can interpret this convergence analysis as a direct modification of FGM's analysis with L replaced with L_{k+1}, an estimate of the unknown Lipschitz parameter L. The role of the backtracking linesearch is to verify the inequality involving L_k.

For a linesearch to be implementable, it is critical that it relies on quantities that are *algorithmically observable*. In the analysis of OGM, the inequalities

$$f(x_k) - f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle - \frac{1}{2L} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 \geq 0$$
$$f_\star - f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle - \frac{1}{2L} \|\nabla f(x_{k+1})\|^2 \geq 0$$

are used. The first inequality involves algorithmically observable quantities and is therefore handy for backtracking linesearches. However, the second is not handy as its verification requires the knowledge of $x_\star$. The convexity inequality on $(x_\star, x_{k+1})$ is handy as it does not involve L and hence does not require verification through a linesearch.

To clarify, we define the notion of "handy inequalities" informally through examples. The motivation is to avoid having certain problematic terms in the analysis. In the following sections, we demonstrate cases where we succeed in generating algorithms with the desired characteristic using the notion of handy inequalities.

3. Optimal algorithm map

In this section, we define the notion of $\mathcal{A}^\star$-optimality, the notion of optimality conditioned on a set of inequalities.

3.1 $\mathcal{A}^\star$-optimality

Define oracles $\mathcal{O} = (\mathcal{O}_0, \mathcal{O}_1)$ to take as input a function and a point and return zero and first-order information of the function, i.e., $\mathcal{O}_0(f, x) = f(x)$ and $\mathcal{O}_1(f, x) = \nabla f(x)$. Define the optimal oracle $\mathcal{O}^\star = (\mathcal{O}^\star_x, \mathcal{O}^\star_f)$, which takes as input a function and returns an optimal point, if one exists, and the optimal value, i.e., $\mathcal{O}^\star_x(f) = x^\star \in \arg\min f$ and $\mathcal{O}^\star_f(f) = f^\star = \inf f$. The optimal oracle $\mathcal{O}^\star$ is used in the minimax formulation, but is, of course, not used in the algorithms.

Let $\mathfrak{A}_N$ be the class of fixed-step first-order algorithms (FSFO) with N iterations, i.e., an algorithm in $\mathfrak{A}_N$ may access $\mathcal{O}_1$ up to N times. To further specify our notation, a first-order algorithm $\mathcal{A}_N(x_0, f)\colon \mathbb{R}^k \times \mathcal{F}_L \to \mathbb{R}^{k\times N}$ in $\mathfrak{A}_N$ generates the N iterates as follows:

$$\begin{aligned}
x_1 &= \mathcal{A}_{N,1}(x_0, \mathcal{O}_1(f, x_0)) \\
x_2 &= \mathcal{A}_{N,2}(x_0, \mathcal{O}_1(f, x_0), \mathcal{O}_1(f, x_1)) \\
&\vdots \\
x_N &= \mathcal{A}_{N,N}(x_0, \mathcal{O}_1(f, x_0), \dots, \mathcal{O}_1(f, x_{N-1})),
\end{aligned}$$

where $\mathcal{A}_{N,i}$ is defined for $i \in \{1, 2, \dots, N\}$ as

$$\mathcal{A}_{N,i}(x_0, g_0, \dots, g_{i-1}) = x_0 - h_{i,0} g_0 - \cdots - h_{i,i-1} g_{i-1}.$$

Let $\mathcal{P}$ be a *performance criterion* that measures the performance of an algorithm $\mathcal{A}$ on a function f. To clarify, a performance criterion only depends on $f(x^\star), \{f(x_i), \nabla f(x_i)\}_{i=0}^N, x^\star, \{x_i\}_{i=0}^N$. For example, the function-value suboptimality

$$\mathcal{P}(\mathcal{A}_N(x_0, f), \mathcal{O}, \mathcal{O}^\star) = f(x_N) - f_\star = \mathcal{O}_0(f, x_N) - \mathcal{O}_f^\star(f)$$

or the squared gradient magnitude

$$\mathcal{P}(\mathcal{A}_N(x_0, f), \mathcal{O}, \mathcal{O}^\star) = \|\nabla f(x_N)\|^2 = \|\mathcal{O}_1(f, x_N)\|^2$$

of the last iterate, x_N are commonly used performance criteria.

Let $\mathcal{C}$ be an *initial condition*, a condition we impose or assume on the initial point x_0. To clarify, an initial condition only depends on $f(x^\star), (f(x_0), \nabla f(x_0)), x^\star, x_0$. For example, the initial distance to a solution

$$\mathcal{C}(x_0, \mathcal{O}, \mathcal{O}^\star) = \{\|x_0 - x_\star\| \leq R\} = \{\|x_0 - \mathcal{O}_x^\star(f)\| \leq R\}$$

or function value suboptimality

$$\mathcal{C}(x_0, \mathcal{O}, \mathcal{O}^\star) = \{f(x_0) - f_\star \leq R\} = \{\mathcal{O}_0(f, x_0) - \mathcal{O}_f^\star(f) \leq R\}$$

are commonly used initial conditions.

Let $\mathcal{I}$ be an *inequality collection*, a set of inequalities the output of the oracles $\mathcal{O}(f, x_0), \dots, \mathcal{O}(f, x_N), \mathcal{O}^\star(f)$ we assume satisfies. In prior work, convergence analyses were permitted to use all true inequalities. Unique to our work, we consider analyses based on a restricted inequality collection $\mathcal{I}$; convergence proofs may use inequalities in $\mathcal{I}$, a strict subset of the true inequalities.

Define the *rate* (or risk) of an algorithm conditioned on $\mathcal{I}$ as

$$\begin{aligned} \mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}) = \sup_{x_0, f, \mathcal{O}, \mathcal{O}^\star} \quad & \mathcal{P}(\mathcal{A}_N(x_0, f), \mathcal{O}, \mathcal{O}^\star) \\ \text{subject to} \quad & x_i = \mathcal{A}_{N,i}(x_0, g_0, \dots, g_{i-1}), \quad i \in \{1, 2, \dots, N\} \\ & (x_0, g_0, f_0, x_\star, f_\star) \text{ satisfies } \mathcal{C}(x_0, \mathcal{O}, \mathcal{O}^\star) \\ & \{(x_i, g_i, f_i)\}_{i=0}^N \text{ and } (x_\star, f_\star) \text{ satisfy } \mathcal{I} \\ & (f_i, g_i) = \mathcal{O}(f, x_i), \quad i \in \{0, \dots, N\} \\ & (x_\star, f_\star) = \mathcal{O}^\star(f). \end{aligned}$$

Note that we impose constraints on f only through the output of the oracles $\mathcal{O}$ and $\mathcal{O}^\star$. Define the minimax optimal rate conditioned on $\mathcal{I}$ as

$$\mathcal{R}^\star(\mathfrak{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}) = \inf_{\mathcal{A}_N \in \mathfrak{A}_N} \mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}). \tag{5}$$

If this infimum is attained, write $\mathcal{A}_N^\star$ to denote the optimal algorithm, and (with some abuse of notation) say the algorithm is $\mathcal{A}^\star$-optimal conditioned on $\mathcal{I}$. Conversely, write $\mathcal{A}_N^\star(\mathcal{P}, \mathcal{C}, \mathcal{I})$ to denote the $\mathcal{A}^\star$-optimal algorithm, and refer to this as the $\mathcal{A}^\star$-map. (An $\mathcal{A}^\star$-optimal algorithm may or may not be unique.)

3.2 Optimality of OGM

A series of work on OGM (Drori and Teboulle, 2014; Kim and Fessler, 2016; Drori, 2017) established that OGM is the exact optimal first-order gradient method. Using our notation, we can express these prior results as

$$\text{OGM} = \mathcal{A}^\star_N \left(f(x_N) - f_\star, \|x_0 - x_\star\| \le R, \mathcal{I}_{L-\text{smooth}} \right)$$

and

$$\mathcal{R}^* \left(\mathfrak{A}_N, f(x_N) - f_\star, \|x_0 - x_\star\| \le R, \mathcal{I}_{L-\text{smooth}} \right) = \frac{LR^2}{2\tilde{\theta}_N^2},$$

where

$$\mathcal{I}_{L-\text{smooth}} = \left\{ f(x_i) \ge f(x_j) - \langle \nabla f(x_j), x_j - x_i \rangle + \frac{1}{2L} \left\| \nabla f(x_i) - \nabla f(x_j) \right\|^2 \right\}_{i,j=0}^{N} \bigcup \left\{ f_\star \ge f(x_k) - \langle \nabla f(x_k), x_k - x_\star \rangle + \frac{1}{2L} \left\| \nabla f(x_k) \right\|^2 \right\}_{k=0}^{N}.$$

3.3 Inequality collection selection

When considering $\mathcal{A}^\star$-optimal algorithms, the choice of the inequality collection represents a tradeoff. On one extreme, if the inequality collection is empty, the performance criterion is the function-value suboptimality or the square gradient magnitude, and the initial condition is the initial distance condition or the function value suboptimality condition (i.e., $\mathcal{I} = \emptyset$, $\mathcal{P}(\mathcal{A}_N, \mathcal{O}, \mathcal{O}^\star) = \mathcal{O}_0(f, x_n) - \mathcal{O}^\star_f(f)$ or $\|\mathcal{O}_1(f, x_N)\|^2$, and $\mathcal{C}(x_0, \mathcal{O}, \mathcal{O}^\star) = \{\|x_0 - \mathcal{O}^\star_x(f)\| \le R\}$ or $\{\mathcal{O}_0(f, x_0) - \mathcal{O}^\star_f(f) \le R\}$), no convergence analysis can be done, and the "algorithm" that does not move from the starting point is $\mathcal{A}^\star$-optimal. On the other extreme, using all true inequalities in the smooth convex minimization setup makes OGM $\mathcal{A}^\star$-optimal. The inequality collections that we consider in later sections include handy inequalities that have the capacity to admit randomized coordinate updates or backtracking linesearches while being sufficiently powerful to establish good rates.[1]

4. ORC-F (Optimized randomized coordinate updates - function value)

In this section, we present ORC-F$_\flat$, an $\mathcal{A}^\star$-optimal algorithm. We first state the theorem precisely describing the $\mathcal{A}^\star$-optimality result, while deferring the proof to the end of this section. We then provide a direct Lyapunov analysis of ORC-F$_\flat$ and modify this Lyapunov analysis to obtain ORC-F, a randomized coordinate update version of ORC-F$_\flat$.

1. As a relevant negative result, we tried but did not succeed in finding a randomized coordinate update version of OGM-G. We tried to modify co-coercivity inequality to the randomized coordinate version. Proof of the OGM-G uses co-coercivity inequality on (x_k, x_{k+1}). Since we choose a random direction for each iterate, it is hard to utilize in $\|\nabla f(x_k) - \nabla f(x_{k+1})\|^2$ term for different direction partial differentiation. We considered several inequality collections that are handy for randomized coordinate updates, but the $\mathcal{A}^\star$-optimal algorithms conditioned on those inequality collections exhibited $\mathcal{O}(1/k)$ rates, i.e., the handy inequalities we considered were not sufficiently powerful to establish the accelerated $\mathcal{O}(1/k^2)$ rate of OGM-G.

4.1 Main results

Optimized randomized coordinate updates - function value$_\flat$ (**ORC-F$_\flat$**) is defined as

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$
$$z_{k+1} = z_k - \frac{\varphi_{k+1} - \varphi_k}{L}\nabla f(x_k)$$
$$x_{k+1} = \frac{\varphi_{k+1}}{\varphi_{k+2}} y_{k+1} + \left(1 - \frac{\varphi_{k+1}}{\varphi_{k+2}}\right) z_{k+1}$$

for $k = 0, 1, \ldots$ where $y_0 = z_0 = x_0$, $\varphi_0 = 0$, and the strictly increasing sequence $\{\varphi_k\}_{k=0}^{\infty}$ is defined by $(2\varphi_{k+1} - \varphi_k) = (\varphi_{k+1} - \varphi_k)^2$ for $k = 0, 1 \ldots$.

Theorem 1 ($\mathcal{A}^\star$-optimality of ORC-F$_\flat$) *ORC-F$_\flat$ is $\mathcal{A}^\star$-optimal in the sense that*

$$\text{ORC-F}_\flat = \mathcal{A}^\star_N \left(f(y_{N+1}) - f_\star, \|x_0 - x_\star\| \leq R, \mathcal{I}_{\text{ORC-F}_\flat} \right)$$

and has the minimax optimal rate

$$\mathcal{R}^\star \left(\mathfrak{A}_N, f(y_{N+1}) - f_\star, \|x_0 - x_\star\| \leq R, \mathcal{I}_{\text{ORC-F}_\flat} \right) = \frac{LR^2}{2\varphi_{N+1}}$$

with respect to the inequalities

$$\begin{aligned}
\mathcal{I}_{\text{ORC-F}_\flat} = &\left\{ f(x_k) \geq f(y_{k+1}) + \frac{1}{2L} \|\nabla f(x_k)\|^2 \right\}_{k=0}^{N} \\
&\bigcup \left\{ f(y_k) \geq f(x_k) + \langle \nabla f(x_k), y_k - x_k \rangle \right\}_{k=1}^{N} \\
&\bigcup \left\{ f_\star \geq f(x_k) + \langle \nabla f(x_k), x_\star - x_k \rangle + \frac{1}{2L} \|\nabla f(x_k)\|^2 \right\}_{k=0}^{N}.
\end{aligned}$$

Note that the inequalities in $\mathcal{I}_{\text{ORC-F}_\flat}$ are handy for randomized coordinate updates. We defer the proof of Theorem 1 to Section 4.3.

The following corollary is a consequence of Theorem 1, but we state it separately and present a standalone proof so that we can modify it for the proof of Theorem 3.

Corollary 2 *Assume (A1), (A2), and (A3). ORC-F$_\flat$'s y_k-sequence exhibits the rate*

$$f(y_{k+1}) - f_\star \leq \frac{L \|x_0 - x_\star\|^2}{2\varphi_{k+1}}$$

for $k = 1, 2, \ldots$.

Proof For $k = 0, 1, 2, \ldots$, define

$$U_k = \varphi_k (f(y_k) - f_\star) + \frac{L}{2} \|z_k - x_\star\|^2.$$

Then we have

$$\begin{aligned}
U_k - U_{k+1} &= \varphi_k(f(y_k) - f_\star) + \frac{L}{2}\|z_k - x_\star\|^2 - \varphi_{k+1}(f(y_{k+1}) - f_\star) - \frac{L}{2}\|z_{k+1} - x_\star\|^2 \\
&= \varphi_k(f(y_k) - f_\star) - \varphi_{k+1}(f(y_{k+1}) - f_\star) + \frac{L}{2}\langle z_k - z_{k+1}, z_k + z_{k+1} - 2x_\star\rangle \\
&= \varphi_k(f(y_k) - f_\star) - \varphi_{k+1}(f(y_{k+1}) - f_\star) \\
&+ \frac{L}{2}\left\langle \frac{\varphi_{k+1} - \varphi_k}{L}\nabla f(x_k), 2z_k - \frac{\varphi_{k+1} - \varphi_k}{L}\nabla f(x_k) - 2x_\star \right\rangle \\
&= \varphi_{k+1}\left(f(x_k) - f(y_{k+1}) - \frac{1}{2L}\|\nabla f(x_k)\|^2\right) \\
&+ \varphi_k\left(f(y_k) - f(x_k) - \langle \nabla f(x_k), y_k - x_k\rangle\right) \\
&+ (\varphi_{k+1} - \varphi_k)\left(f_\star - f(x_k) - \langle \nabla f(x_k), x_\star - x_k\rangle - \frac{1}{2L}\|\nabla f(x_k)\|^2\right) \\
&+ \frac{2\varphi_{k+1} - \varphi_k}{2L}\|\nabla f(x_k)\|^2 + \varphi_k\langle \nabla f(x_k), y_k - x_k\rangle + (\varphi_{k+1} - \varphi_k)\langle \nabla f(x_k), x_\star - x_k\rangle \\
&- \frac{(\varphi_{k+1} - \varphi_k)^2}{2L}\|\nabla f(x_k)\|^2 + (\varphi_{k+1} - \varphi_k)\langle \nabla f(x_k), z_k - x_\star\rangle.
\end{aligned}$$

Since

$$\begin{aligned}
(\varphi_{k+1} - \varphi_k)\langle \nabla f(x_k), z_k - x_\star\rangle &= (\varphi_{k+1} - \varphi_k)\langle \nabla f(x_k), z_k - x_k\rangle + (\varphi_{k+1} - \varphi_k)\langle \nabla f(x_k), x_k - x_\star\rangle \\
&= \varphi_k\langle \nabla f(x_k), x_k - y_k\rangle + (\varphi_{k+1} - \varphi_k)\langle \nabla f(x_k), x_k - x_\star\rangle
\end{aligned}$$

and $(2\varphi_{k+1} - \varphi_k) = (\varphi_{k+1} - \varphi_k)^2$, we get

$$\begin{aligned}
U_k - U_{k+1} &= \varphi_{k+1}\left(f(x_k) - f(y_{k+1}) - \frac{1}{2L}\|\nabla f(x_k)\|^2\right) \\
&+ \varphi_k\left(f(y_k) - f(x_k) - \langle \nabla f(x_k), y_k - x_k\rangle\right) \\
&+ (\varphi_{k+1} - \varphi_k)\left(f_\star - f(x_k) - \langle \nabla f(x_k), x_\star - x_k\rangle - \frac{1}{2L}\|\nabla f(x_k)\|^2\right) \\
&\geq 0.
\end{aligned}$$

We conclude $\varphi_{k+1}(f(y_{k+1}) - f^\star) \leq U_{k+1} \leq \cdots \leq U_{-1} = \frac{L}{2}\|x_0 - x_\star\|^2$. ■

Note that the proof only utilized inequalities in $\mathcal{I}_{\text{ORC-F}_\flat}$.

Randomized coordinate updates version. Assume f is a coordinate-wise smooth function with parameters $(L_1, \ldots, L_n)$. Define $S = \sum_{i=1}^n \sqrt{L_i}$. At iteration k, select the coordinate $i(k)$ with probability $\mathbb{P}(i(k) = t) = \frac{\sqrt{L_t}}{S}$. Define optimized randomized coordinate updates - function value (**ORC-F**), a randomized coordinate updates version of ORC-F$_\flat$, as

$$\begin{aligned}
y_{k+1} &= x_k - \frac{1}{L_{i(k)}}\nabla_{i(k)} f(x_k) \\
z_{k+1} &= z_k - \frac{\varphi_{k+1} - \varphi_k}{S\sqrt{L_{i(k)}}}\nabla_{i(k)} f(x_k) \\
x_{k+1} &= \frac{\varphi_{k+1}}{\varphi_{k+2}} y_{k+1} + \left(1 - \frac{\varphi_{k+1}}{\varphi_{k+2}}\right) z_{k+1}
\end{aligned}$$

for $k = 0, 1, \ldots$.

Theorem 3 *Assume (A1) and (A3). Assume f is a coordinate-wise smooth function with parameters $(L_1, \ldots, L_n)$. Then ORC-F exhibits the rate as*

$$\mathbb{E}\left[f(y_{k+1})\right] - f_\star \leq \frac{S^2 \left\|x_0 - x_\star\right\|^2}{2\varphi_{k+1}}$$

for $k = 0, 1, \ldots$.

Proof For $k = 0, 1, \ldots$, define

$$U_k = \frac{\varphi_k}{S^2}(f(y_k) - f_\star) + \frac{1}{2}\left\|z_k - x_\star\right\|^2$$

and define $\mathbb{E}_{i(k)}$ as the expectation conditioned on $i(0), \ldots, i(k-1)$. Then, we have

$$\begin{aligned}
U_k - U_{k+1} &= \frac{\varphi_k}{S^2}(f(y_k) - f_\star) + \frac{1}{2}\left\|z_k - x_\star\right\|^2 - \frac{\varphi_{k+1}}{S^2}(f(y_{k+1}) - f_\star) - \frac{1}{2}\left\|z_{k+1} - x_\star\right\|^2 \\
&= \frac{\varphi_k}{S^2}(f(y_k) - f_\star) - \frac{\varphi_{k+1}}{S^2}(f(y_{k+1}) - f_\star) + \frac{1}{2}\langle z_k - z_{k+1}, z_k + z_{k+1} - 2x_\star\rangle \\
&= \frac{\varphi_k}{S^2}(f(y_k) - f_\star) - \frac{\varphi_{k+1}}{S^2}(f(y_{k+1}) - f_\star) \\
&+ \frac{1}{2}\left\langle \frac{\varphi_{k+1} - \varphi_k}{S\sqrt{L_{i(k)}}}\nabla_{i(k)} f(x_k), 2z_k - \frac{\varphi_{k+1} - \varphi_k}{S\sqrt{L_{i(k)}}}\nabla_{i(k)} f(x_k) - 2x_\star \right\rangle \\
&= \frac{(\varphi_{k+1} - \varphi_k)}{S^2}\left(f_\star - f(x_k) - \frac{S}{\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), x_\star - x_k\rangle - \frac{1}{2L_{i(k)}}\left\|\nabla_{i(k)} f(x_k)\right\|^2\right) \\
&+ \frac{\varphi_k}{S^2}\left(f(y_k) - f(x_k) - \frac{S}{\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), y_k - x_k\rangle\right) \\
&+ \frac{\varphi_{k+1}}{S^2}\left(f(x_k) - f(y_{k+1}) - \frac{1}{2L_{i(k)}}\left\|\nabla_{i(k)} f(x_k)\right\|^2\right) \\
&+ \frac{2\varphi_{k+1} - \varphi_k}{2S^2 L_{i(k)}}\left\|\nabla_{i(k)} f(x_k)\right\|^2 + \frac{\varphi_k}{S\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), y_k - x_k\rangle + \frac{\varphi_{k+1} - \varphi_k}{S\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), x_\star - x_k\rangle \\
&- \frac{(\varphi_{k+1} - \varphi_k)^2}{2S^2 L_{i(k)}}\left\|\nabla_{i(k)} f(x_k)\right\|^2 + \frac{\varphi_{k+1} - \varphi_k}{S\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), z_k - x_\star\rangle.
\end{aligned}$$

Since

$$\begin{aligned}
\frac{\varphi_{k+1} - \varphi_k}{S\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), z_k - x_\star\rangle &= \frac{\varphi_{k+1} - \varphi_k}{S\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), z_k - x_k\rangle + \frac{\varphi_{k+1} - \varphi_k}{S\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), x_k - x_\star\rangle \\
&= \frac{\varphi_{k+1} - \varphi_k}{S\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), x_k - x_\star\rangle + \frac{\varphi_k}{S\sqrt{L_{i(k)}}}\langle \nabla_{i(k)} f(x_k), x_k - y_k\rangle
\end{aligned}$$

and $(2\varphi_{k+1} - \varphi_k) = (\varphi_{k+1} - \varphi_k)^2$, we get

$$
\begin{aligned}
U_k - U_{k+1} = \frac{(\varphi_{k+1} - \varphi_k)}{S^2} &\left(f_\star - f(x_k) - \frac{S}{\sqrt{L_{i(k)}}} \langle \nabla_{i(k)} f(x_k), x_\star - x_k \rangle - \frac{1}{2L_{i(k)}} \left\| \nabla_{i(k)} f(x_k) \right\|^2 \right) \\
&+ \frac{\varphi_k}{S^2} \left(f(y_k) - f(x_k) - \frac{S}{\sqrt{L_{i(k)}}} \langle \nabla_{i(k)} f(x_k), y_k - x_k \rangle \right) \\
&+ \frac{\varphi_{k+1}}{S^2} \left(f(x_k) - f(y_{k+1}) - \frac{1}{2L_{i(k)}} \left\| \nabla_{i(k)} f(x_k) \right\|^2 \right).
\end{aligned}
$$

and taking the conditional expectation, we have

$$
\begin{aligned}
U_k - \mathbb{E}_{i(k)} U_{k+1} = \mathbb{E}_{i(k)} &\left[\frac{(\varphi_{k+1} - \varphi_k)}{S^2} \left(f_\star - f(x_k) - \langle \nabla f(x_k), x_\star - x_k \rangle - \frac{1}{2L_{i(k)}} \left\| \nabla_{i(k)} f(x_k) \right\|^2 \right) \right] \\
&+ \frac{\varphi_k}{S^2} \left(f(y_k) - f(x_k) - \langle \nabla f(x_k), y_k - x_k \rangle \right) \\
&+ \mathbb{E}_{i(k)} \left[\frac{\varphi_{k+1}}{S^2} \left(f(x_k) - f(y_{k+1}) - \frac{1}{2L_{i(k)}} \left\| \nabla_{i(k)} f(x_k) \right\|^2 \right) \right] \\
&\geq 0.
\end{aligned}
$$

Taking the full expectation, we have $\mathbb{E} U_k \leq \cdots \leq U_0$, and we conclude the statement of the theorem. ■

Discussion. ORC-F has the smallest (best) constant among the randomized coordinate updates methods, to the best of our knowledge. In particular, ORC-F's rate is slightly faster than that of FGM-RC of Allen-Zhu et al. (2016) or FGM-RC$^\sharp$ since $\theta_k^2 \leq \varphi_{k+1}$ for $k = 0, 1, \ldots$, which follows from induction. However, the improvement is small as the leading-term constant is the same, i.e., $\theta_k^2 / \varphi_{k+1} \to 1$ as $k \to \infty$.

4.2 Brief review of Taylor et al. (2017b)

This section is closely follows Taylor et al. (2017b) with minor changes including different in notation and the addition of some variables. Consider the underlying space $\mathbb{R}^d$ with $d \geq N+2$. The assumption that d is sufficiently large is made to obtain dimension-independent results. See (Taylor et al., 2017b, Section 3.3) for further discussion of this matter. Denote $y_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$ and $y_0 = x_0$. Define $f_{i,0} = f(x_i)$, $f_{i,1} = f(y_i)$, $g_i = \nabla f(x_i)$, and

$$
\begin{aligned}
\mathbf{G} &= \begin{pmatrix}
\|x_0 - x_\star\|^2 & \langle g_0, x_0 - x_\star \rangle & \langle g_1, x_0 - x_\star \rangle & \dots & \langle g_N, x_0 - x_\star \rangle \\
\langle g_0, x_0 - x_\star \rangle & \|g_0\|^2 & \langle g_1, g_0 \rangle & \dots & \langle g_N, g_0 \rangle \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\langle g_N, x_0 - x_\star \rangle & \langle g_N, g_0 \rangle & \langle g_N, g_1 \rangle & \dots & \|g_N\|^2
\end{pmatrix}, \\
\mathbf{F}_0 &= \begin{pmatrix} f_{0,0} - f_\star \\ f_{1,0} - f_\star \\ \vdots \\ f_{N+1,0} - f_\star \end{pmatrix}, \qquad
\mathbf{F}_1 = \begin{pmatrix} f_{0,1} - f_\star \\ f_{1,1} - f_\star \\ \vdots \\ f_{N+1,1} - f_\star \end{pmatrix}.
\end{aligned}
\tag{6}
$$

Note that $\mathbf{G} \succeq 0$, $\mathbf{F}_0 \succeq 0$, and $\mathbf{F}_1 \succeq 0$, i.e., $\mathbf{G}$ is positive semidefinite and $\mathbf{F}_0$ and $\mathbf{F}_1$ are elementwise nonnegative. Since $d \geq N+2$, given $\mathbf{G}$ and $\mathbf{F}_0 \succeq 0$, we can take the Cholesky factorization[2]of $\mathbf{G}$ to recover the triplet $\{(x_i, g_i, f_i)\}_{i=0}^{N}$. Define

$$\boldsymbol{x}_0 = e_1 \in \mathbb{R}^{N+2}, \quad \boldsymbol{g}_i = e_{i+2} \in \mathbb{R}^{N+2}, \quad \boldsymbol{f}_i = e_{i+1} \in \mathbb{R}^{N+2} \tag{7}$$

for $i = 0, 1, \ldots, N$, where e_i are standard basis of $\mathbb{R}^{N+2}$ or $\mathbb{R}^{N+1}$. We define FSFO (1) with $(h_{i,j})$ and $(s_{i,j})$ as

$$\begin{aligned} \boldsymbol{x}_{i+1} &= \boldsymbol{x}_i - \sum_{k=0}^{i} \frac{h_{i+1,k}}{L} \boldsymbol{g}_k \\ &= \boldsymbol{x}_0 - \sum_{k=0}^{i} \frac{s_{i+1,k}}{L} \boldsymbol{g}_k \end{aligned} \tag{8}$$

for $i = 0, 1, \ldots, N-1$. (Note that $\boldsymbol{x}_i$ appears in the first expression while $\boldsymbol{x}_0$ appears in the second.) With this new notation, we can write

$$\begin{aligned} &f_{i,j} - f_\star = \boldsymbol{f}_i^\intercal \mathbf{F}_j, && i = 0, 1, \ldots, N+1, \quad j = 0, 1 \\ &\langle g_i, g_j \rangle = \boldsymbol{g}_i^\intercal \mathbf{G} \boldsymbol{g}_j, && i, j = 0, 1, \ldots, N \\ &\|x_i - x_\star\|^2 = \boldsymbol{x}_i^\intercal \mathbf{G} \boldsymbol{x}_i, && i = 0, 1, \ldots, N \\ &\langle g_i, x_j - x_\star \rangle = \boldsymbol{g}_i^\intercal \mathbf{G} \boldsymbol{x}_j, && i, j = 0, 1, \ldots, N. \end{aligned}$$

This allows us to express the optimization of the algorithm as an optimization problem with variables $\mathbf{G}, \mathbf{F}_0, \mathbf{F}_1$.

Let $\mathcal{F}_L$ be the class of L-smooth convex functions. Let I be an index set and consider the set of triplets $S = \{(x_i, g_i, f_i)\}_{i \in I}$, where $x_i, g_i \in \mathbb{R}^d$ and $f_i \in \mathbb{R}$ for all $i \in I$. We say S is $\mathcal{F}_L$-interpolable if and only if there exists a function $f \in \mathcal{F}_L$ that $g_i \in \partial f(x_i)$ and $f(x_i) = f_i$ for all $i \in I$.

Fact 1 (Taylor et al., 2017b, Theorem 4) *S is $\mathcal{F}_L$-interpolable if and only if*

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

4.2.1 Strong duality for the PEP

We propose a general PEP form. The original sdp-PEP of Taylor et al. (2017b) is

$$\begin{aligned} \underset{\mathbf{G}, \mathbf{F}_0}{\text{maximize}} \quad & b^\intercal \mathbf{F}_0 + \mathrm{Tr}(C\mathbf{G}) \\ \text{subject to} \quad & 0 \geq (\boldsymbol{f}_j - \boldsymbol{f}_i)^\intercal \mathbf{F}_0 + \mathrm{Tr}\Big(\mathbf{G}\big((\boldsymbol{x}_i - \boldsymbol{x}_j)\boldsymbol{g}_j^\intercal + \frac{1}{2L}(\boldsymbol{g}_i - \boldsymbol{g}_j)(\boldsymbol{g}_i - \boldsymbol{g}_j)^\intercal\big)\Big) \qquad i, j \in \{0, 1, \ldots, N\} \\ & 1 \geq \mathrm{Tr}(\mathbf{G}\boldsymbol{x}_0\boldsymbol{x}_0^\intercal) \\ & 0 \preceq \mathbf{G}, \end{aligned}$$

2. Since $\mathbf{G}$ is not strictly positive definite, the "Cholesky factorization" is not unique. In fact, any factorization of the form $\mathbf{G} = MM^\intercal$ suffices. See (Taylor et al., 2017b, Section 3) for further discussion on this matter.

for $b \in \mathbb{R}^{N+1}$ and C is a nonnegative definite matrix. For further details, refer to (Taylor et al., 2017b, Theorem 5). This original sdp-PEP is induced by the $\mathcal{F}_L$-interpolable condition. We extend this sdp-PEP to replace the constraints with relaxed inequalities. Our general sdp-PEP is

$$\begin{array}{ll} \underset{\mathbf{G},\mathbf{F}_0,\mathbf{F}_1}{\text{maximize}} & b_0^\intercal \mathbf{F}_0 + b_1^\intercal \mathbf{F}_1 + \mathrm{Tr}(C\mathbf{G}) \\ \text{subject to} & \text{conditions corresponding to inequality collection } \mathcal{I} \\ & 1 \geq \mathrm{Tr}(\mathbf{G}\boldsymbol{x}_0\boldsymbol{x}_0^\intercal) \\ & 0 \preceq \mathbf{G}, \end{array}$$

for $b_0, b_1 \in \mathbb{R}^{N+1}$ and C is a nonnegative definite matrix. Specific instances of this general sdp-PEP are considered in subsequent section. We call the convex-dual problem of general sdp-PEP as dual-sdp-PEP (Taylor et al., 2017b). Strong duality holds between the primal and dual SDPs.

Fact 2 *Assume the stepsizes of* (8) *satisfy* $s_{k,k-1} \neq 0$ *for* $k = 1, \ldots, N$. *In addition, inequality collection corresponds to the algorithms in Sections 4.3, 5.1, and 5.2. Then, the strong duality holds between general sdp-PEP and dual-sdp-PEP.*

The formal proof Fact 2, which we omit for the sake of brevity, follows from the same reasoning as that of (Taylor et al., 2017b, Theorem 5).

4.3 Proof of Theorem 1

In this section, we prove Theorem 1, i.e., $\mathcal{A}^\star$-optimality of ORC-F$_\flat$, using the PEP machinery. To verify the lengthy calculations, we provide Matlab scripts verifying the analytical solution of the SDP:
`https://github.com/chanwoo-park-official/A-star-map/`.
To obtain ORC-F as an $\mathcal{A}^\star$-optimal algorithm, set $f(y_{N+1}) - f_\star$ to be the performance measure and $\|x_0 - x_\star\| \leq R$ to be the initial condition. Since the constraints and the objective of the problem are homogeneous, we assume $R = 1$ without loss of generality. For the argument of homogeneous, we refer to (Drori and Teboulle, 2014; Kim and Fessler, 2016; Taylor et al., 2017b). We use the set of inequalities that are handy for randomized coordinate updates:

$$\mathcal{I}_{\text{ORC-F}_\flat} = \left\{ f_{k,0} \geq f_{k+1,1} + \frac{1}{2L}\|g_k\|^2 \right\}_{k=0}^{N} \bigcup \left\{ f_{k,1} \geq f_{k,0} + \langle g_k, y_k - x_k \rangle \right\}_{k=1}^{N} \\ \bigcup \left\{ f_\star \geq f_{k,0} + \langle g_k, x_\star - x_k \rangle + \frac{1}{2L}\|g_k\|^2 \right\}_{k=0}^{N}.$$

For calculating $\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{ORC-F}_\flat})$ with fixed $\mathcal{A}_N$, define the PEP with $\mathcal{I}_{\text{ORC-F}_\flat}$ as

$$
\begin{aligned}
&\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{ORC-F}_\flat}) \\
&= \left(\begin{array}{llll}
\text{maximize} & f_{N+1,1} & -f_\star & \\
\text{subject to} & 1 \;\; \geq & \|x_0 - x_\star\|^2 & \\
 & f_{k,0} \geq & f_{k+1,1} + \frac{1}{2L}\|g_k\|^2, & k \in \{0,1,\dots,N\} \\
 & f_{k,1} \geq & f_{k,0} + \langle g_k, y_k - x_k\rangle, & k \in \{1,\dots,N\} \\
 & f_\star \;\; \geq & f_{k,0} + \langle g_k, x_\star - x_k\rangle + \frac{1}{2L}\|g_k\|^2, & k \in \{0,1,\dots,N\} \\
 & x_k, y_k & \text{are following the algorithm } \mathcal{A}_N. &
\end{array}\right)
\end{aligned} \tag{9}
$$

Using the notation of Section 4.2, we reformulate the problem of computing the risk $\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{ORC-F}_\flat})$ as the following SDP:

$$
\begin{aligned}
\underset{\mathbf{G},\mathbf{F}_0,\mathbf{F}_1}{\text{maximize}} \quad & \boldsymbol{f}_{N+1}^\intercal \mathbf{F}_1 \\
\text{subject to} \quad & 1 \geq \boldsymbol{x}_0^\intercal \mathbf{G} \boldsymbol{x}_0 \\
& 0 \geq \boldsymbol{f}_{k+1}^\intercal \mathbf{F}_1 - \boldsymbol{f}_k^\intercal \mathbf{F}_0 + \frac{1}{2L}\boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{g}_k, && k \in \{0,1,\dots,N\} \\
& 0 \geq \boldsymbol{f}_k^\intercal (\mathbf{F}_0 - \mathbf{F}_1) + \boldsymbol{g}_k^\intercal \mathbf{G}(\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) - \frac{1}{L}\boldsymbol{g}_{k-1}^\intercal \mathbf{G} \boldsymbol{g}_k, && k \in \{1,2,\dots,N\} \\
& 0 \geq \boldsymbol{f}_k^\intercal \mathbf{F}_0 - \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{x}_k + \frac{1}{2L}\boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{g}_k, && k \in \{0,1,\dots,N\} \\
& \mathbf{G} \succcurlyeq 0, \mathbf{F}_0 \geq 0, \mathbf{F}_1 \geq 0.
\end{aligned}
$$

For above transformation, $d \geq N+2$ is used (Taylor et al., 2017b). The Lagrangian of the optimization problem becomes

$$
\begin{aligned}
&\Lambda(\mathbf{F}_0, \mathbf{F}_1, \mathbf{G}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau) \\
&\qquad = -\boldsymbol{f}_{N+1}^\intercal \mathbf{F}_1 + \tau(\boldsymbol{x}_0^\intercal \mathbf{G} \boldsymbol{x}_0 - 1) + \sum_{k=0}^{N} \alpha_k \left(\boldsymbol{f}_{k+1}^\intercal \mathbf{F}_1 - \boldsymbol{f}_k^\intercal \mathbf{F}_0 + \frac{1}{2L}\boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{g}_k\right) \\
&\qquad + \sum_{k=1}^{N} \lambda_k \left(\boldsymbol{f}_k^\intercal (\mathbf{F}_0 - \mathbf{F}_1) + \boldsymbol{g}_k^\intercal \mathbf{G}(\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) - \frac{1}{L}\boldsymbol{g}_{k-1}^\intercal \mathbf{G} \boldsymbol{g}_k\right) \\
&\qquad + \sum_{k=0}^{N} \beta_k \left(\boldsymbol{f}_k^\intercal \mathbf{F}_0 - \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{x}_k + \frac{1}{2L}\boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{g}_k\right)
\end{aligned}
$$

with dual variables $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N) \in \mathbb{R}_+^N$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_N) \in \mathbb{R}_+^{N+1}$, $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_N) \in \mathbb{R}_+^{N+1}$, and $\tau \geq 0$. Then the dual formulation of PEP problem is

$$\begin{array}{ll} \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau)\geq \mathbf{0}}{\text{maximize}} & -\tau \\ \text{subject to} & \mathbf{0} = -\sum_{k=0}^{N} \alpha_k \boldsymbol{f}_k + \sum_{k=1}^{N} \lambda_k \boldsymbol{f}_k + \sum_{k=0}^{N} \beta_k \boldsymbol{f}_k \\ & \mathbf{0} = -\boldsymbol{f}_{N+1} - \sum_{k=1}^{N} \lambda_k \boldsymbol{f}_k + \sum_{k=0}^{N} \alpha_k \boldsymbol{f}_{k+1} \\ & 0 \preceq S(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau), \end{array} \tag{10}$$

where S is defined as

$$\begin{aligned} S(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau) = {} & \tau \boldsymbol{x}_0 \boldsymbol{x}_0^{\intercal} + \sum_{k=0}^{N} (\alpha_k + \beta_k) \left(\frac{1}{2L} \boldsymbol{g}_k \boldsymbol{g}_k^{\intercal} \right) + \sum_{k=0}^{N} \frac{\beta_k}{2} \left(-\boldsymbol{g}_k \boldsymbol{x}_k^{\intercal} - \boldsymbol{x}_k \boldsymbol{g}_k^{\intercal} \right) \\ & + \sum_{k=1}^{N} \frac{\lambda_k}{2} \left(\boldsymbol{g}_k (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k)^{\intercal} + (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) \boldsymbol{g}_k^{\intercal} - \frac{1}{L} \boldsymbol{g}_{k-1} \boldsymbol{g}_k^{\intercal} - \frac{1}{L} \boldsymbol{g}_k \boldsymbol{g}_{k-1}^{\intercal} \right). \end{aligned}$$

Using the strong duality result of Fact 2 and a continuity argument that we justify at the end of this proof, we proceed with

$$\underset{h_{i,j}}{\arg\min} \underset{\mathbf{G}, \mathbf{F}_0, \mathbf{F}_1}{\text{maximize}} \ \boldsymbol{f}_{N+1}^{\intercal} \mathbf{F}_1 = \underset{h_{i,j}}{\arg\min} \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau)\geq \mathbf{0}}{\text{minimize}} \ \tau \tag{11}$$

i.e., it is sufficient to obtain $h_{i,j}$'s argmin value of (10). We omitted (11)'s constraints for ease of writing. Note that (10) finds the optimal proof for the algorithm. Minimizing (10) with respect to $(h_{i,j})$ corresponds to optimizing the algorithm:

$$\underset{h_{i,j}}{\text{minimize}} \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau)\geq \mathbf{0}}{\text{minimize}} \quad \tau \tag{12}$$

$$\text{subject to} \quad \mathbf{0} = -\sum_{k=0}^{N} \alpha_k \boldsymbol{f}_k + \sum_{k=1}^{N} \lambda_k \boldsymbol{f}_k + \sum_{k=0}^{N} \beta_k \boldsymbol{f}_k \tag{13}$$

$$\mathbf{0} = -\boldsymbol{f}_{N+1} - \sum_{k=1}^{N} \lambda_k \boldsymbol{f}_k + \sum_{k=0}^{N} \alpha_k \boldsymbol{f}_{k+1} \tag{14}$$

$$0 \preceq S(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau). \tag{15}$$

We note that $\boldsymbol{f}_i$ is a standard unit vector mentioned in (7) (not a variable), we can write (13) and (14) as

$$\begin{array}{c} \left(\begin{array}{ll} \beta_k = \alpha_k - \lambda_k = \lambda_{k+1} - \lambda_k, & k \in \{1, \ldots, N-1\} \\ \beta_0 = \alpha_0 = \lambda_1 & \\ \beta_N = \alpha_N - \lambda_N = 1 - \lambda_N. & \end{array} \right) \\ \left(\begin{array}{ll} \alpha_N = 1 & \\ \alpha_k = \lambda_{k+1}, & k \in \{0, 1, \ldots, N-1\} \end{array} \right) \end{array} \tag{16}$$

We consider (15) with (16) and FSFO's $h_{i,j}$. To be specific, we substitute $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to $\boldsymbol{\lambda}$ in $S(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau)$. To show the dependency of S to $(h_{i,j})$ since $\boldsymbol{x}_k$ are represented with $(h_{i,j})$, we will explicitly write S as $S(\boldsymbol{\lambda},\tau;(h_{i,j}))$. Then, we get

$$\begin{aligned}
S(\boldsymbol{\lambda},\tau;(h_{i,j})) &= \tau\boldsymbol{x}_0\boldsymbol{x}_0^\intercal + \frac{\lambda_1}{2L}\boldsymbol{g}_0\boldsymbol{g}_0^\intercal - \sum_{k=1}^{N-1}\frac{2\lambda_k-\lambda_{k+1}}{2L}\boldsymbol{g}_k\boldsymbol{g}_k^\intercal + \frac{2-2\lambda_N}{2L}\boldsymbol{g}_N\boldsymbol{g}_N^\intercal + \sum_{k=1}^{N}\frac{\lambda_k}{2L}(\boldsymbol{g}_{k-1}-\boldsymbol{g}_k)(\boldsymbol{g}_{k-1}-\boldsymbol{g}_k)^\intercal \\
&+ \sum_{k=1}^{N-1}\sum_{t=0}^{k-1}\left(\frac{\lambda_k}{2}\frac{h_{k,t}}{L} + \frac{\lambda_{k+1}-\lambda_k}{2}\sum_{j=t+1}^{k}\frac{h_{j,t}}{L}\right)\left(\boldsymbol{g}_k\boldsymbol{g}_t^\intercal + \boldsymbol{g}_t\boldsymbol{g}_k^\intercal\right) \\
&+ \sum_{t=0}^{N-1}\left(\frac{\lambda_N}{2}\frac{h_{N,t}}{L} + \frac{1-\lambda_N}{2}\sum_{j=t+1}^{N}\frac{h_{j,t}}{L}\right)\left(\boldsymbol{g}_N\boldsymbol{g}_t^\intercal + \boldsymbol{g}_t\boldsymbol{g}_N^\intercal\right) \\
&- \sum_{k=1}^{N-1}\frac{\lambda_{k+1}-\lambda_k}{2}\left(\boldsymbol{x}_0\boldsymbol{g}_k^\intercal + \boldsymbol{g}_k\boldsymbol{x}_0^\intercal\right) - \frac{\lambda_1}{2}\left(\boldsymbol{x}_0\boldsymbol{g}_0^\intercal + \boldsymbol{g}_0\boldsymbol{x}_0^\intercal\right) - \frac{1-\lambda_N}{2}\left(\boldsymbol{x}_0\boldsymbol{g}_N^\intercal + \boldsymbol{g}_N\boldsymbol{x}_0^\intercal\right).
\end{aligned}$$

Using the fact that $\boldsymbol{x}_0,\boldsymbol{g}_i,\boldsymbol{f}_i$ are unit vectors, we can represent $S(\boldsymbol{\lambda},\tau;(h_{i,j}))$ with $\boldsymbol{\gamma}(\boldsymbol{\lambda}) = -L\boldsymbol{\beta} = -L(\lambda_1,\lambda_2-\lambda_1,\dots,1-\lambda_N) = (\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}),\gamma_N(\boldsymbol{\lambda}))$ and $\tau' = 2L\tau$ as

$$S(\boldsymbol{\lambda},\tau';(h_{i,j})) = \frac{1}{L}\begin{pmatrix} \frac{1}{2}\tau' & \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})^\intercal & \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) \\ \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) & Q(\boldsymbol{\lambda};(h_{i,j})) & q(\boldsymbol{\lambda};(h_{i,j})) \\ \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) & q(\boldsymbol{\lambda};(h_{i,j}))^\intercal & \frac{2-\lambda_N}{2} \end{pmatrix} \succeq 0.$$

Here, Q and $\mathbf{q}$ are defined as

$$\begin{aligned}
Q(\boldsymbol{\lambda};(h_{i,j})) &= \frac{\lambda_1}{2}\boldsymbol{g}_0'\boldsymbol{g}_0'^\intercal + \sum_{k=1}^{N-1}\frac{\lambda_{k+1}-2\lambda_k}{2}\boldsymbol{g}_k'\boldsymbol{g}_k'^\intercal + \sum_{k=1}^{N-1}\frac{\lambda_k}{2}(\boldsymbol{g}_{k-1}'-\boldsymbol{g}_k')(\boldsymbol{g}_{k-1}'-\boldsymbol{g}_k)'^\intercal + \frac{\lambda_N}{2}\boldsymbol{g}_{N-1}'\boldsymbol{g}_{N-1}'^\intercal \\
&+ \sum_{k=1}^{N-1}\sum_{t=0}^{k-1}\left(\frac{\lambda_k}{2}h_{k,t} + \frac{\lambda_{k+1}-\lambda_k}{2}\sum_{j=t+1}^{k}h_{j,t}\right)\left(\boldsymbol{g}_k'\boldsymbol{g}_t'^\intercal + \boldsymbol{g}_t'\boldsymbol{g}_k'^\intercal\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{q}(\boldsymbol{\lambda};(h_{i,j})) &= -\frac{\lambda_N}{2}\boldsymbol{g}_{N-1}' + \sum_{t=0}^{N-1}\left(\frac{\lambda_N}{2}h_{N,t} + \frac{1-\lambda_N}{2}\sum_{j=t+1}^{N}h_{j,t}\right)\boldsymbol{g}_t' \\
&= \sum_{t=0}^{N-2}\left(\frac{\lambda_N}{2}h_{N,t} + \frac{1-\lambda_N}{2}\sum_{j=t+1}^{N}h_{j,t}\right)\boldsymbol{g}_t' + \left(\frac{1}{2}h_{N,N-1} - \frac{\lambda_N}{2}\right)\boldsymbol{g}_{N-1}'
\end{aligned}$$

where $\boldsymbol{g}_k' = e_{k+1} \in \mathbb{R}^{N+1}$. Note that (12) is equivalent to

$$\begin{aligned}
&\underset{h_{i,j}}{\text{minimize}}\ \underset{(\boldsymbol{\lambda},\tau')\geq\mathbf{0}}{\text{minimize}} \quad \tau' \\
&\text{subject to} \quad \begin{pmatrix} \frac{1}{2}\tau' & \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})^\intercal & \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) \\ \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) & Q(\boldsymbol{\lambda};(h_{i,j})) & q(\boldsymbol{\lambda};(h_{i,j})) \\ \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) & q(\boldsymbol{\lambda};(h_{i,j}))^\intercal & \frac{2-\lambda_N}{2} \end{pmatrix} \succeq 0
\end{aligned} \tag{17}$$

and dividing this optimized value with $2L$ gives the optimized value of (12). Using Schur complement (?), (we already know $0 \le \lambda_N \le 1$ by (16)) (17) can be converted to the problem as

$$\underset{h_{i,j}}{\text{minimize}}\ \underset{(\boldsymbol{\lambda},\tau')\ge\mathbf{0}}{\text{minimize}}\quad \tau' \tag{18}$$

$$\text{subject to}\quad \begin{pmatrix} Q - \frac{2\mathbf{q}\mathbf{q}^\intercal}{2-\lambda_N} & \frac{1}{2}(\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) - \frac{2\mathbf{q}\gamma_N(\boldsymbol{\lambda})}{2-\lambda_N}) \\ \frac{1}{2}(\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) - \frac{2\mathbf{q}\gamma_N(\boldsymbol{\lambda})}{2-\lambda_N})^\intercal & \frac{1}{2}(\tau' - \frac{\gamma_N(\boldsymbol{\lambda})^2}{2-\lambda_N}) \end{pmatrix} \succeq 0. \tag{19}$$

So far we simplified SDP. We will have three steps: finding variables that make (19)'s left hand side zero, showing that the solution from the first step satisfies Karuch-Kuhn-Tucker (KKT) condition, and finally showing that obtained algorithm is equivalent to ORC-F$_\flat$.

Claim 1 *There is a point that makes* (19)*'s left-hand side zero.*

Proof Defining the positive sequence $\{\varphi_k\}_{k=0}^\infty$ as

$$2\varphi_{k+1} - \varphi_k = (\varphi_{k+1} - \varphi_k)^2$$

for $k = 0, 1, \ldots$, $\varphi_0 = 0$, and $\{\varphi_k\}_{k=0}^\infty$ is a strictly increasing sequence. Defining $\{r_{k,t}\}_{k=1,2,\ldots,N,t=0,\ldots,k-1}$ as

$$r_{k,t} = \lambda_k h_{k,t} - \frac{\gamma_k}{L}\sum_{j=t+1}^{k} h_{j,t}.$$

Then, if $r_{i,j}$ is determined, (Drori and Teboulle, 2014, Theorem 3) indicates this uniquely determine $h_{i,j}$. We set $(\lambda_k)_{k=0}^N$ and $(r_{N,k})_{k=0}^{N-1}$ as

$$\begin{aligned} \lambda_k &= \frac{\varphi_k}{\varphi_{N+1}}, \qquad k \in \{0, 1, \ldots, N\} \\ r_{N,k} &= \frac{(\varphi_{k+1} - \varphi_k)(\varphi_{N+1} - \varphi_N)}{\varphi_{N+1}}, \qquad k \in \{0, 1, \ldots, N-2\} \\ r_{N,N-1} - \lambda_N &= \frac{(\varphi_{N+1} - \varphi_{N-1})(\varphi_{N+1} - \varphi_N)}{\varphi_{N+1}}. \end{aligned} \tag{20}$$

Moreover, we set

$$\begin{aligned} r_{k,t} &= \frac{1}{\varphi_{N+1}}(\varphi_{k+1} - \varphi_k)(\varphi_{t+1} - \varphi_t), \qquad k \in \{1, 2, \ldots, N-1\}, \quad t \in \{0, 1, \ldots, k-2\} \\ r_{k,k-1} - \lambda_k &= \frac{1}{\varphi_{N+1}}(\varphi_{k+1} - \varphi_k)(\varphi_k - \varphi_{k-1}), \qquad k \in \{1, 2 \ldots, N-1\}. \end{aligned} \tag{21}$$

In addition, we set $\hat{\boldsymbol{\gamma}}$ as

$$\begin{aligned} \gamma_t &= \frac{\gamma_N}{2 - \lambda_N} r_{N,t}, \qquad t \in \{0, 1, \ldots, N-2\} \\ \gamma_{N-1} &= \frac{\gamma_N(r_{N,N-1} - \lambda_N)}{2 - \lambda_N}. \end{aligned}$$

Lastly, we set τ' as

$$\tau' = \frac{L^2}{\varphi_{N+1}}, \tag{22}$$

and $\tau = \frac{L}{2\varphi_{N+1}}$. These variables make (19)'s left-hand side zero. ∎

Claim 2 (20)*,* (21) *and* (22) *are an optimal solution of* (18)*.*

Proof Let we represent S with the variable $(r_{i,j})$. We will denote this as $\mathbf{A}$. To be specific,

$$\mathbf{A}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j})) = S(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \tau'; (h_{i,j})) = \begin{pmatrix} \frac{1}{2}\tau' & -\frac{L}{2}\hat{\boldsymbol{\beta}}^\intercal & -\frac{L}{2}\beta_N \\ -\frac{L}{2}\hat{\boldsymbol{\beta}}^\intercal & Q(\boldsymbol{\lambda}; (r_{i,j})) & \mathbf{q}((r_{i,j})) \\ -\frac{L}{2}\beta_N & \mathbf{q}((r_{i,j}))^\intercal & \frac{2-\lambda_N}{2} \end{pmatrix} \succeq 0.$$

Here, $\boldsymbol{\beta} = (\hat{\boldsymbol{\beta}}^\intercal, \beta_N)^\intercal$,

$$\begin{aligned} Q(\boldsymbol{\lambda}; (r_{i,j})) &= \frac{\lambda_1}{2}\boldsymbol{g}_0\boldsymbol{g}_0^\intercal + \sum_{k=1}^{N-1} \frac{\lambda_{k+1} - 2\lambda_k}{2}\boldsymbol{g}_k\boldsymbol{g}_k^\intercal + \sum_{k=1}^{N-1} \frac{\lambda_k}{2}(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal \\ &\quad + \frac{\lambda_N}{2}\boldsymbol{g}_{N-1}\boldsymbol{g}_{N-1}^\intercal + \sum_{k=1}^{N-1}\sum_{t=0}^{k-1} \left(\frac{r_{k,t}}{2}\right)\left(\boldsymbol{g}_k\boldsymbol{g}_t^\intercal + \boldsymbol{g}_t\boldsymbol{g}_k^\intercal\right), \end{aligned}$$

and

$$\mathbf{q}((r_{i,j})) = \sum_{t=0}^{N-1} \frac{r_{N,t}}{2}\boldsymbol{g}_t - \frac{\lambda_N}{2}\boldsymbol{g}_{N-1}.$$

Define a linear SDP relaxation of (17) as

$$\begin{aligned} \underset{r_{i,j}}{\text{minimize}}\ \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau')\geq\mathbf{0}}{\text{minimize}} \quad & \tau' \\ \text{subject to} \quad & \mathbf{A}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j})) \succeq 0. \\ & \mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau') = \left(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau'\right) \geq 0 \\ & \mathbf{C}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = (-\alpha_0 + \beta_0, -\alpha_1 + \lambda_1 + \beta_1, \dots, -\alpha_N + \lambda_N + \beta_N) = 0 \\ & \mathbf{D}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = (-\lambda_1 + \alpha_0, -\lambda_2 + \alpha_1, \dots, -\lambda_N + \alpha_{N-1}, \alpha_N - 1) = 0. \end{aligned} \tag{23}$$

(Drori and Teboulle, 2014, Theorem 3) indicates that if we prove the choice in the previous claim satisfies KKT condition of (23), then this is also an optimal solution for the original problem since $(r_{i,j})$ uniquely determines $h_{i,j}$. The Lagrangian of the minimization problem is

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j}), \mathbf{K}, \mathbf{b}, \mathbf{c}, \mathbf{d}) \\ &\quad = \frac{1}{2}\tau' - \operatorname{tr}\left\{\mathbf{A}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j}))\mathbf{K}\right\} - \mathbf{b}^\intercal\mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau') - \mathbf{c}^\intercal\mathbf{C}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}) - \mathbf{d}^\intercal\mathbf{D}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \end{aligned}$$

and the KKT conditions of the minimization problems are

$$\begin{aligned}
&\mathbf{A}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau'; (r_{i,j})) \succeq 0, \mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau') \geq 0, \mathbf{C}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = 0, \mathbf{D}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = 0, \\
&\nabla_{(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j}))} \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j}), \mathbf{K}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = 0, \\
&\mathbf{K} \succeq 0, \mathbf{b} \geq 0, \\
&\operatorname{tr}\left\{\mathbf{A}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j}))\mathbf{K}\right\} = 0, \mathbf{b}^\intercal \mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau') = 0,
\end{aligned}$$

where $\mathbf{K}$ is a symmetric matrix. Here, $\mathbf{b} = (\mathbf{u}, \mathbf{v}, \mathbf{w}, s)$. We re-index K's column and row starting from -1 (so K's rows and columns index are $\{-1, 0, 1 \ldots, N\}$). Now, we will show that there exist a dual optimal solution $(\mathbf{K}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ that $(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j}), \mathbf{K}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ satisfies KKT condition, which proves a pair $(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j}))$ is an optimal solution for primal problem. The stationary condition $\nabla_{(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j}))} \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau', (r_{i,j}), \mathbf{K}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = 0$ can be rewritten as

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \lambda_k} = -\frac{1}{2}\left(2K_{k-1,k-1} - K_{k-1,k} - K_{k,k-1} - K_{k,k}\right) - u_k - c_k + d_{k-1} = 0, \qquad k \in \{1, 2, \ldots, N\} \\
&\frac{\partial \mathcal{L}}{\partial \beta_k} = \frac{L}{2}\left(K_{-1,k} + K_{k,-1}\right) - v_k - c_k = 0, \qquad k \in \{0, 1, \ldots, N\} \\
&\frac{\partial \mathcal{L}}{\partial \alpha_k} = -w_k + c_k - d_k = 0, \qquad k \in \{0, 1, \ldots, N\} \\
&\frac{\partial \mathcal{L}}{\partial \tau'} = \frac{1}{2} - \frac{1}{2}K_{-1,-1} - s = 0 \\
&\frac{\partial \mathcal{L}}{\partial r_{k,t}} = -\frac{1}{2}(K_{k,t} + K_{t,k}) = 0, \qquad k \in \{1, 2, \ldots, N\}, \quad t \in \{0, 1, \ldots, k-1\}.
\end{aligned} \tag{24}$$

We already know that $\mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau') \neq 0$, we can set $\mathbf{b} = 0$. Then, (24) reduces to

$$\begin{aligned}
&K_{k,t} = 0, \qquad k \in \{1, 2, \ldots, N\}, \quad t \in \{0, 1, \ldots, k-1\} \\
&-\frac{1}{2}\left(2K_{k-1,k-1} - K_{k,k}\right) - c_k + d_{k-1} = 0, \qquad k \in \{1, 2, \ldots, N\} \\
&LK_{-1,k} - c_k = 0, \qquad k \in \{0, 1, \ldots, N\} \\
&c_k - d_k = 0, \qquad k \in \{0, 1, \ldots, N\} \\
&K_{-1,-1} = 1.
\end{aligned}$$

Then, we have

$$\mathbf{K} = \begin{pmatrix}
1 & \frac{c_0}{L} & \frac{c_1}{L} & \ldots & \frac{c_{N-1}}{L} & \frac{c_N}{L} \\
\frac{c_0}{L} & K_{0,0} & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\frac{c_{N-1}}{L} & 0 & 0 & \ldots & K_{N-1,N-1} & 0 \\
\frac{c_N}{L} & 0 & 0 & \ldots & 0 & K_{N,N}
\end{pmatrix} \succeq 0$$

and since $\operatorname{tr}\{\mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))\mathbf{K}\} = 0$ with $\mathbf{A} \succeq 0$, we can replace this condition by $\mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))\mathbf{K} = 0$. Then the KKT condition for the given $(\boldsymbol{\lambda},\boldsymbol{\beta},\tau',(r_{i,j}))$ reduces to

$$\begin{aligned}
&\frac{1}{2}\tau' - \frac{1}{2}\boldsymbol{\beta}^\intercal \mathbf{c} = 0\\
&\frac{1}{2L}\tau'\mathbf{c} - \frac{L}{2}\operatorname{diag}(K_{0,0},\dots,K_{N-1,N-1},K_{N,N})\boldsymbol{\beta} = 0\\
&-\frac{1}{2}\boldsymbol{\beta}\mathbf{c}^\intercal + \begin{pmatrix} Q & \mathbf{q} \\ \mathbf{q}^\intercal & \frac{2-\lambda_N}{2}\end{pmatrix}\operatorname{diag}(K_{0,0},\dots,K_{N-1,N-1},K_{N,N}) = 0.
\end{aligned}$$

By solving the equation, we have $c_i = (\varphi_{i+1}-\varphi_i)K_{i,i}$ for $i = 0,1,\dots,N$ and we have $1 = \sum_{i=0}^N \frac{c_i^2}{L^2 K_{i,i}}$ by the first above equation. Therefore, $K \succeq 0$. ■

Claim 3 *The obtained algorithm is ORC-F$_\flat$.*

Proof By calculating $(h_{i,j})$ of ORC-F$_\flat$, we can prove the equivalence of the obtained solution and ORC-F$_\flat$. Indeed, ORC-F$_\flat$ is obtained by using (Lee et al., 2021)'s auxiliary sequences. We will show that obtained $(\hat{h}_{i,j})$ satisfies

$$\begin{aligned}
x_0 - \sum_{i=1}^{k+1}\sum_{j=0}^{i-1}\frac{\hat{h}_{i.j}}{L}\nabla f(x_j) = {}& \frac{\varphi_{k+1}}{\varphi_{k+2}}\left(x_0 - \sum_{i=1}^{k}\sum_{j=0}^{i-1}\frac{\hat{h}_{i,j}}{L}\nabla f(x_j) - \frac{1}{L}\nabla f(x_k)\right)\\
&+\left(1-\frac{\varphi_{k+1}}{\varphi_{k+2}}\right)\left(x_0 - \sum_{j=0}^{k}\frac{\varphi_{j+1}-\varphi_j}{L}\nabla f(x_j)\right),
\end{aligned}$$

which is re-written form of ORC-F$_\flat$. Comparing $\nabla f(x_j)$'s each coefficient, we should prove

$$\begin{aligned}
\sum_{i=j+1}^{k+1}\hat{h}_{i,j} &= \frac{\varphi_{k+1}}{\varphi_{k+2}}\sum_{i=j+1}^{k}\hat{h}_{i,j} + \left(1-\frac{\varphi_{k+1}}{\varphi_{k+2}}\right)(\varphi_{j+1}-\varphi_j) \qquad j \in \{0,1,\dots,k-1\}\\
\hat{h}_{k+1,k} &= \frac{\varphi_{k+1}}{\varphi_{k+2}} + \left(1-\frac{\varphi_{k+1}}{\varphi_{k+2}}\right)(\varphi_{k+1}-\varphi_k),
\end{aligned}$$

which is exactly equal to the recursive rule of (Drori and Teboulle, 2014, Theorem 3). ■

It now remains to justify (11). Write $\|(s_{i,j})-(s'_{i,j})\|_\infty \le \varepsilon$ if $\max_{i,j}|s_{i,j}-s'_{i,j}| \le \varepsilon$. Denote the optimal value of (9) as $p((s_{i,j}))$, i.e.,

$$p((s_{i,j})) = \underset{\mathbf{G},\mathbf{F}_0,\mathbf{F}_1}{\text{maximize}}\ \boldsymbol{f}_{N+1}^\intercal \mathbf{F}_1.$$

We show that p is a continuous function. If p is continuous, by Fact 2,

$$\underset{s_{i,i-1}}{\arg\min}\ \underset{\mathbf{G},\mathbf{F}_0,\mathbf{F}_1}{\text{maximize}}\ \boldsymbol{f}_{N+1}^\intercal \mathbf{F}_1 = \underset{s_{i,i-1}\neq 0}{\arg\min}\ \underset{\mathbf{G},\mathbf{F}_0,\mathbf{F}_1}{\text{maximize}}\ \boldsymbol{f}_{N+1}^\intercal \mathbf{F}_1 = \underset{s_{i,i-1}\neq 0}{\arg\min}\ \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau)\geq \mathbf{0}}{\text{minimize}}\ \tau.$$

Since our analytic solution for $\arg\min_{(s_{i,j})}\text{minimize}_{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau)\geq\mathbf{0}}\ \tau$ satisfies $s_{i,i-1} \neq 0$, the strong duality claim (11) is justified.

Finally, we establish continuity of $p((s_{i,j}))$ with the following claim.

Claim 4 *Assume* $\|(s_{i,j}) - (s'_{i,j})\|_\infty \leq \varepsilon$. *Let* $\{x_i, y_i\}_{i=0}^N$ *be points with the FSFO with coefficients* $(s_{i,j})$ *and* $\{x'_i, y'_i\}_{i=0}^N$ *be points with the FSFO with coefficients* $(s'_{i,j})$. *Assume* $\{(x_i, y_i, g_i, f_{i,0}, f_{i+1,1})\}_{i=0}^N$ *satisfies* $\mathcal{I}_{\text{ORC}}$. *Assume* $\|x_0 - x_\star\| \leq R$. *Moreover, we can find* $\{(x'_i, y'_i, g'_i, f'_{i,0}, f'_{i+1,1})\}_{i=0}^N = \{(x'_i, y'_i, g_i, f_{i,0} - \frac{i^2}{L}C\varepsilon, f_{i+1,1} - \frac{i^2}{L}C\varepsilon)\}_{i=0}^N$ *that satisfies* $\mathcal{I}_{\text{ORC}}$, *where* $C = C(\{(s_{i,j}), R, L\})$ *is a constant continuously depending only on* $(\{(s_{i,j}), R, L\})$.

Proof A continuous function C only depending on $\{(s_{i,j}), R, L\}$ bounds $\max_{i=0}^N \|g_i\|^2 \leq C$. We first show that such a constant exists. Under the initial condition $\|x_0 - x_\star\|^2 \leq R^2$,

$$\begin{aligned}\frac{1}{4L}\|g_k\|^2 + L\|x_k - x_\star\|^2 - \frac{1}{2L}\|g_k\|^2 &\geq \langle g_k, x_k - x_\star\rangle - \frac{1}{2L}\|g_k\|^2 \\ &\geq f_\star - f_{k,0} - \langle g_k, x_\star - x_k\rangle - \frac{1}{2L}\|g_k\|^2 \geq 0,\end{aligned}$$

where the first inequality follows from Young's inequality, the second inequality follows from the fact that $f_\star$ is the optimal value, and the third inequality is the cocoercivity inequality on $(x_\star, y_k)$. So $L\|x_k - x_\star\|^2 \geq \frac{1}{4L}\|g_k\|^2$, and if $\|x_k - x_\star\|^2$ is bounded by a continuous function then $\|g_k\|^2$ is also bounded by a continuous function. Inductively, if $\|g_0\|, \ldots, \|g_{k-1}\|$ are bounded by a continuous function ($\|g_k\|$ not included), then

$$\begin{aligned}\|x_k - x_\star\| &= \left\| x_0 - x_\star - \sum_{i=0}^{k-1} \frac{s_{k,i}}{L} g_i \right\| \\ &\leq \|x_0 - x_\star\| + \sum_{i=0}^{k-1} \frac{|s_{k,i}|}{L}\|g_i\|\end{aligned}$$

indicates that $\|x_k - x_\star\|$ is bounded by a continuous function. Chaining these arguments inductively while making sure to check that the "continuous function" only depend on $\{(s_{i,j}), R, L\}$, we conclude $\max_{i=0}^N \|g_i\|^2 \leq C$.

Without loss of generality, assume $f_\star = f'_\star = 0$. Then,

$$\begin{pmatrix} \Gamma_{k,1} = f_{k,0} - f_{k+1,1} - \frac{1}{2L}\|g_k\|^2 \geq 0, & k \in \{0, 1, \ldots, N\} \\ \Gamma_{k,2} = f_{k,1} - f_{k,0} - \langle g_k, y_k - x_k\rangle \geq 0 & k \in \{1, \ldots, N\} \\ \Gamma_{k,3} = -f_{k,0} - \langle g_k, x_\star - x_k\rangle - \frac{1}{2L}\|g_k\|^2 \geq 0, & k \in \{0, 1, \ldots, N\} \end{pmatrix}.$$

We will show

$$\begin{pmatrix} f'_{k,0} - f'_{k+1,1} - \frac{1}{2L}\|g'_k\|^2 \geq 0, & k \in \{0, 1, \ldots, N\} \\ f'_{k,1} - f'_{k,0} - \langle g'_k, y'_k - x'_k\rangle \geq 0 & k \in \{1, \ldots, N\} \\ -f'_{k,0} - \langle g'_k, x_\star - x'_k\rangle - \frac{1}{2L}\|g'_k\|^2 \geq 0, & k \in \{0, 1, \ldots, N\} \end{pmatrix}.$$

This is equivalent to

$$\begin{pmatrix} f_{k,0} - f_{k+1,1} - \frac{1}{2L}\|g_k\|^2 \geq 0, & k \in \{0, 1, \ldots, N\} \\ f_{k,1} - f_{k,0} + \frac{(2k-1)C}{L}\varepsilon - \langle g_k, (y'_k - x'_k) - (y_k - x_k) + (y_k - x_k)\rangle \geq 0 & k \in \{1, \ldots, N\} \\ -f_{k,0} + \frac{k^2}{L}C\varepsilon - \langle g_k, (x_\star - x_k) + (x_k - x'_k)\rangle - \frac{1}{2L}\|g_k\|^2 \geq 0, & k \in \{0, 1, \ldots, N\} \end{pmatrix}.$$

This can be reduced as

$$\begin{pmatrix} \Gamma_{k,1} \geq 0, & k \in \{0, 1, \dots, N\} \\ \Gamma_{k,2} + \frac{(2k-1)C}{L}\varepsilon - \langle g_k, (y_k' - x_k') - (y_k - x_k)\rangle \geq 0 & k \in \{1, \dots, N\} \\ \Gamma_{k,3} + \frac{k^2}{L}C\varepsilon - \langle g_k, (x_k - x_k')\rangle \geq 0, & k \in \{0, 1, \dots, N\} \end{pmatrix}.$$

For the second one,

$$\begin{aligned} &\langle g_k, (y_k' - y_k) - (x_k' - x_k)\rangle \\ &= \langle g_k, (x_{k-1}' - x_{k-1}) - (x_k' - x_k)\rangle \\ &= \langle g_k, -\frac{1}{L}\sum_{i=0}^{k-2}(s_{k-1,i}' - s_{k-1,i})g_i + \frac{1}{L}\sum_{i=0}^{k-1}(s_{k,i}' - s_{k,i})g_i\rangle \\ &\leq \frac{1}{L}\left\|g_k\right\|\left(\sum_{i=0}^{k-2}\varepsilon\left\|g_i\right\| + \sum_{i=0}^{k-1}\varepsilon\left\|g_i\right\|\right) \\ &\leq \frac{2k-1}{L}C\varepsilon. \end{aligned}$$

For the third one,

$$\langle g_k, (x_k' - x_k)\rangle = \langle g_k, -\frac{1}{L}\sum_{i=0}^{k-1}(s_{k,i}' - s_{k,i})g_i\rangle \leq \frac{1}{L}\left\|g_k\right\|\left(\sum_{i=0}^{k-1}\varepsilon\left\|g_i\right\|\right) \leq \frac{k}{L}C\varepsilon,$$

which shows the claim. ∎

Finally, we prove continuity of $p((s_{i,j}))$. For $0 < \varepsilon < 1$, fix any $(s_{i,j})$ and $(s_{i,j}')$ that $\|(s_{i,j}) - (s_{i,j}')\|_\infty \leq \varepsilon$. Define $\mathcal{A}((s_{i,j}))$ and $\mathcal{A}((s_{i,j}'))$ as the algorithms corresponding to $(s_{i,j})$ and $(s_{i,j}')$, respectively. For any $\{(x_i, y_i, g_i, f_{i,0}, f_{i+1,1})\}_{i=0}^N$ generated by $\mathcal{A}((s_{i,j}))$ and satisfying $\mathcal{I}_{\text{ORC-F}}$ and $\|x_0 - x_\star\| \leq R$, there exists $\{(x_i', y_i', g_i', f_{i,0}', f_{i+1,1}')\}_{i=0}^N$ generated by $\mathcal{A}((s_{i,j}'))$ satisfying $\mathcal{I}_{\text{ORC-F}}$ and $\|x_0 - x_\star\| \leq R$, such that the performance measures difference satisfies $(f_{N+1,1} - f_\star) - (f_{N+1,1}' - f_\star') = N^2C((s_{i,j}), R, L)\varepsilon$, where C depends only on $\{(s_{i,j}), R, L\}$. Therefore, $p((s_{i,j})) - p((s_{i,j}')) \leq N^2C((s_{i,j}), R, L)\varepsilon$. Conversely, for $\{(x_i', y_i', g_i', f_{i,0}', f_{i+1,1}')\}_{i=0}^N$ generated by $\mathcal{A}((s_{i,j}'))$ and satisfying $\mathcal{I}_{\text{ORC-F}}$ and $\|x_0 - x_\star\| \leq R$, there exists $\{(x_i, y_i, g_i, f_{i,0}, f_{i+1,1})\}_{i=0}^N$ generated by $\mathcal{A}((s_{i,j}))$ and satisfying $\mathcal{I}_{\text{ORC-F}}$ and $\|x_0 - x_\star\| \leq R$, such that the performance measures difference satisfies $(f_{N+1,1} - f_\star) - (f_{N+1,1}' - f_\star') = -N^2C((s_{i,j}'), R, L)\varepsilon$. Therefore, $p((s_{i,j}')) - p((s_{i,j})) \leq N^2C((s_{i,j}'), R, L)\varepsilon$. Since C is a continuous function of $(s_{i,j}')$, we conclude $|p((s_{i,j}')) - p((s_{i,j}))| \to 0$ as $\varepsilon \to 0$.

To summarize, the algorithm's performance criterion $f(y_{N+1}) - f_\star$ is bounded as

$$f(y_{N+1}) - f_\star \leq \frac{L}{2\varphi_{N+1}}\left\|x_0 - x_\star\right\|^2.$$

Overall, we showed that ORC-$\text{F}_\flat$ is $\mathcal{A}^\star$-optimal in the sense that $\mathcal{I}_{\text{ORC-F}_\flat}$

$$\text{ORC-F}_\flat = \mathcal{A}_N^\star(f(y_{N+1}) - f_\star, \|x_0 - x_\star\| \leq R, \mathcal{I}_{\text{ORC-F}_\flat}).$$

Furthermore,

$$\begin{aligned}&\mathcal{R}(\text{ORC-F}_\flat, f(y_{N+1}) - f_\star, \|x_0 - x_\star\| \le R, \mathcal{I}_{\text{ORC-F}_\flat})\\&\qquad = \mathcal{R}^\star(\mathfrak{A}_N, f(y_{N+1}) - f_\star, \|x_0 - x_\star\| \le R, \mathcal{I}_{\text{ORC-F}_\flat}) = \frac{LR^2}{2\varphi_{N+1}}.\end{aligned}$$

5. Other results

We also have two more $\mathcal{A}^\star$-optimal algorithms; OBL-$\text{F}_\flat$ and FGM, which will be explained in this section. Moreover, we give a conjecture about one $\mathcal{A}^\star$-optimal algorithm; OBL-$\text{G}_\flat$.

5.1 OBL-F

Optimized backtracking linesearch - function value$_\flat$ (**OBL-F$_\flat$**) is defined as

$$\begin{aligned}y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k)\\ z_{k+1} &= z_k - \frac{k+1}{L}\nabla f(x_k)\\ x_{k+1} &= \left(1 - \frac{2}{k+3}\right) y_{k+1} + \frac{2}{k+3} z_{k+1}\end{aligned}$$

for $k = 0, 1, \dots$ where $y_0 = z_0 = x_0$. The *last-step modification* for OBL-$\text{F}_\flat$ on secondary sequence is written as

$$\tilde{x}_k = \frac{1}{\sqrt{\frac{k(k+1)}{2}} + 1}\left(\sqrt{\frac{k(k+1)}{2}} y_k + z_k\right)$$

where $k = 0, 1, \dots$.

Theorem 4 ($\mathcal{A}^\star$-optimality of OBL-F$_\flat$) *OBL-$F_\flat$ is $\mathcal{A}^\star$-optimal in the sense that*

$$\text{OBL-F}_\flat = \mathcal{A}^\star_N(f(x_N) - f_\star, \|x_0 - x_\star\| \le R, \mathcal{I}_{\text{OBL-F}_\flat})$$

and has the minimax optimal rate

$$\mathcal{R}^\star(\mathfrak{A}_N, f(x_N) - f_\star, \|x_0 - x_\star\| \le R, \mathcal{I}_{\text{OBL-F}_\flat}) = \frac{LR^2}{k(k+1) + \sqrt{2k(k+1)}}$$

with respect to the inequalities

$$\begin{aligned}\mathcal{I}_{\text{OBL-F}_\flat} = &\left\{ f(x_{k-1}) \ge f(x_k) + \langle \nabla f(x_k), x_{k-1} - x_k \rangle + \frac{1}{2L}\left\|\nabla f(x_{k-1}) - \nabla f(x_k)\right\|^2 \right\}_{k=1}^N\\ &\bigcup \left\{ f_\star \ge f(x_k) + \langle \nabla f(x_k), x_\star - x_k \rangle \right\}_{k=0}^N.\end{aligned}$$

Note that the inequalities in $\mathcal{I}_{\text{OBL-F}_\flat}$ are handy for backtracking linesearches. We defer the proof of Theorem 4 to Appendix B.

The following corollary is a consequence of Theorem 4, but we state it separately and present a standalone proof so that we can modify it for the proof of Theorem 6.

Corollary 5 *Assume (A1), (A2), and (A3). OBL-F$_\flat$'s $\tilde{x}_k$-sequence and y_k-sequence exhibit the rate*

$$f(\tilde{x}_k) - f_\star \le \frac{L\|x_0 - x_\star\|^2}{k(k+1) + \sqrt{2k(k+1)}}$$

and

$$f(y_{k+1}) - f_\star \le \frac{L\|x_0 - x_\star\|^2}{(k+1)(k+2)}$$

for $k = 1, 2, \ldots$.

Proof Let $x_{-1} = x_0$. For $k = -1, 0, 1, \ldots$, define

$$U_k = \frac{(k+1)(k+2)}{2}\left(f(x_k) - f_\star - \frac{1}{2L}\|\nabla f(x_k)\|^2\right) + \frac{L}{2}\|z_{k+1} - x_\star\|^2$$

and

$$\tilde{U}_k = \left(\sqrt{\frac{k(k+1)}{2}} + \frac{k(k+1)}{2}\right)(f(\tilde{x}_k) - f_\star) + \frac{L}{2}\left\|z_k - \frac{1}{L}\frac{k(k+1)}{2}\nabla f(\tilde{x}_k) - x_\star\right\|^2.$$

Then we have $U_{k+1} \overset{(*)}{\le} U_k$ and $\tilde{U}_k \le U_{k-1}$, which implies

$$\left(\sqrt{\frac{k(k+1)}{2}} + \frac{k(k+1)}{2}\right)(f(\tilde{x}_k) - f_\star) \le \tilde{U}_k \le U_{k-1} \le \cdots \le U_{-1} = \frac{L}{2}\|x_0 - x_\star\|^2$$

and

$$\frac{(k+1)(k+2)}{2}(f(y_{k+1}) - f_\star) \le U_k \le U_{k-1} \le \cdots \le U_{-1} = \frac{L}{2}\|x_0 - x_\star\|^2.$$

To complete the proof, it remains to justify the (*) part. We defer the calculations to Appendix A.1. ■

Note that the proof only utilized inequalities in $\mathcal{I}_{\text{OBL-F}_\flat}$. The Lyapunov function in this proof was inspired by the Lyapunov function used in the analysis of OGM in (Park et al., 2023).

Backtracking linesearch version. Define optimized backtracking linesearch - function value (**OBL-F**), a line backtracking version of OBL-F$_\flat$, as follows. Initialize L_0 and $\eta > 1$. For $k = 0, 1, \ldots$, we define $x_{k+1}, y_{k+1}, z_{k+1}$ as

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L_{k+1}}\nabla f(x_k) \\ z_{k+1} &= z_k - \frac{k+1}{L_{k+1}}\nabla f(x_k) \\ x_{k+1} &= \left(1 - \frac{2}{k+3}\right)y_{k+1} + \frac{2}{k+3}z_{k+1} \end{aligned}$$

with $L_{k+1} = \eta^{i_{k+1}}L_k$ where $y_0 = z_0 = x_0$. The backtracking linesearch finds the smallest i_{k+1} such that

$$\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L_{k+1}}\|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 + \langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle\right) \ge 0.$$

Theorem 6 *Assume (A1), (A2), and (A3). OBL-F exhibits the rate as*

$$f(y_{N+1}) - f_\star \leq \frac{L_N}{(N+1)(N+2)} \left(\|x_0 - x_\star\|^2 + \sum_{k\in K} \frac{(k+1)(k+2)}{2} \left(\frac{1}{L_k^2} - \frac{1}{L_{k+1}^2} \right) \|\nabla f(x_{k+1})\|^2 \right).$$

Proof Let $x_{-1} = x_0$. For $k = -1, 0, 1, \ldots$, define

$$U_k = \frac{(k+1)(k+2)}{2L_k} \left(f(x_k) - f_\star - \frac{1}{2L_k} \|\nabla f(x_k)\|^2 \right) + \frac{1}{2} \|z_{k+1} - x_\star\|^2 .$$

Then we have

$$U_k - U_{k+1} \overset{(*)}{\geq} \frac{(k+1)(k+2)}{4} \left(\frac{1}{L_{k+1}^2} - \frac{1}{L_k^2} \right) \|\nabla f(x_k)\|^2 .$$

If $L_N \geq L$ which L is smoothness constant of L, then if we set y_{N+1} as gradient $1/L_N$-step of x_N (i.e. $y_{N+1} = x_N - \frac{1}{L_N} x_N$). We define K as the set of having smooth factor-jump, then by the above relationship, we have

$$\begin{aligned}
&\frac{1}{2} \|x_0 - x_\star\|^2 \\
&\geq \frac{(N+1)(N+2)}{2L_N} \left(f(x_N) - f_\star - \frac{1}{2L_N} \|\nabla f(x_N)\|^2 \right) + \sum_{k\in K} \frac{(k+1)(k+2)}{4} \left(\frac{1}{L_{k+1}^2} - \frac{1}{L_k^2} \right) \|\nabla f(x_k)\|^2 \\
&\geq \frac{(N+1)(N+2)}{2L_N} (f(y_{N+1}) - f_\star) + \sum_{k\in K} \frac{(k+1)(k+2)}{4} \left(\frac{1}{L_{k+1}^2} - \frac{1}{L_k^2} \right) \|\nabla f(x_k)\|^2
\end{aligned}$$

which indicates

$$f(y_{N+1}) - f_\star \leq \frac{L_N}{(N+1)(N+2)} \left(\|x_0 - x_\star\|^2 + \sum_{k\in K} \frac{(k+1)(k+2)}{2} \left(\frac{1}{L_k^2} - \frac{1}{L_{k+1}^2} \right) \|\nabla f(x_k)\|^2 \right).$$

Note that K would be a sparse set (informally) that is subset of $\{1, 2, \ldots, N\}$. The justification of $(*)$ is deferred to Appendix A.2. ■

Discussion. The rates of OGM, OGM-simple (Park et al., 2023), OBL-F$_\flat$, and OBL-F all have the same leading-term constants, i.e. the limit of convergence rate's ratio when $k \to \infty$ is 1. We clarify that although OBL-F$_\flat$ and OGM-simple (Park et al., 2023) are similar in their forms, the two algorithms are distinct.

5.2 $\mathcal{A}^\star$-optimality of FGM

A question that motivated this work was whether FGM is an exactly optimal algorithm in some sense. Here, we provide the answer that FGM is $\mathcal{A}^\star$-optimal conditioned on a set of inequalities that are handy for both randomized coordinate updates and backtracking linesearches. Indeed FGM does admit the variants FGM-RC$^\sharp$ and FGM-BL as discussed in Section 1.1.

Theorem 7 ($\mathcal{A}^\star$-optimality of FGM) *Nesterov's FGM is $\mathcal{A}^\star$-optimal in the sense that*

$$\text{FGM} = \mathcal{A}^\star_N(f(y_{N+1}) - f_\star, \|x_0 - x_\star\| \leq R, \mathcal{I}_{\text{FGM}})$$

and has the minimax optimal rate

$$\mathcal{R}^\star(\mathfrak{A}_N, f(y_{N+1}) - f_\star, \|x_0 - x_\star\| \leq R, \mathcal{I}_{\text{FGM}}) = \frac{LR^2}{2\theta_N^2}$$

with respect to the inequalities

$$\begin{aligned}\mathcal{I}_{\text{FGM}} =& \left\{ f(x_k) \geq f(y_{k+1}) + \frac{1}{2L}\|\nabla f(x_k)\|^2 \right\}_{k=0}^N \bigcup \left\{ f(y_k) \geq f(x_k) + \langle \nabla f(x_k), y_k - x_k \rangle \right\}_{k=1}^N \\ &\bigcup \left\{ f_\star \geq f(x_k) + \langle \nabla f(x_k), x_\star - x_k \rangle \right\}_{k=0}^N .\end{aligned}$$

We defer the proof of Theorem 7 to Section B.

5.3 OBL-G

Optimized backtracking linesearch - gradient norm$_\flat$ (**OBL-G$_\flat$**) is defined as

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ z_{k+1} &= z_k - \frac{1}{L}\frac{N-k+1}{2}\nabla f(x_k) \\ x_{k+1} &= \frac{N-k-2}{N-k+2}y_{k+1} + \frac{4}{N-k+2}z_{k+1} \end{aligned}$$

for $k = 1, 2, \dots, N-1$ where $y_0 = z_0 = x_0$, and

$$\begin{aligned} y_1 &= x_0 - \frac{1}{L}\nabla f(x_0) \\ z_1 &= z_0 - \frac{1}{L}\frac{1 + \sqrt{\frac{N(N+1)}{2}}}{2}\nabla f(x_k) \\ x_1 &= \frac{N-2}{N+2}y_{k+1} + \frac{4}{N+2}z_{k+1}. \end{aligned}$$

The PEP characterizing OBL-G$_\flat$ turns out to be bi-convex (hence non-convex) and this non-convexity prevents us from establishing $\mathcal{A}^\star$-optimality of OBL-G$_\flat$. This non-convexity was also present in the prior work of OGM-G by Kim and Fessler (2021), as we further discuss in Section 5.3.1. Nevertheless, numerical evidence indicates that OBL-G$_\flat$ is likely $A^\star$-optimal, so we state the following claim as a conjecture.

Conjecture 8 ($\mathcal{A}^\star$-optimality of OBL-G$_\flat$) *OBL-G$_\flat$ is $\mathcal{A}^\star$-optimal in the sense that*

$$\text{OBL-G}_\flat = \mathcal{A}^\star_N(\|\nabla f(x_N)\|^2, f(x_0) - f_\star \leq \frac{1}{2}LR^2, \mathcal{I}_{\text{OBL-G}_\flat})$$

and has the minimax optimal rate

$$\mathcal{R}^\star(\mathfrak{A}_N, \|\nabla f(x_N)\|^2, f(x_0) - f_\star \le \frac{1}{2}LR^2, \mathcal{I}_{\text{OBL-G}_\flat}) = 2L^2R^2 \frac{N^2 + N - \sqrt{2N(N+1)}}{N^2(N+1)^2 - 2\sqrt{2N(N+1)}}$$

with respect to the inequalities

$$\begin{aligned}\mathcal{I}_{\text{OBL-G}_\flat} = &\left\{ f(x_k) \ge f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle + \frac{1}{2L} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 \right\}_{k=0}^{N-1} \\ &\bigcup \{ f(x_N) \ge f(x_k) + \langle \nabla f(x_k), x_k - x_N \rangle \}_{k=0}^{N} \\ &\bigcup \left\{ f(x_N) \ge f_\star + \frac{1}{2L} \|\nabla f(x_N)\|^2 \right\},\end{aligned}$$

Note that the inequalities in $\mathcal{I}_{\text{OBL-G}_\flat}$ are handy for backtracking linesearches.

Since Conjecture 8 is just a conjecture, the following convergence rate of OBL-G$_\flat$ must be established as a standalone result.[3] Again, we will modify this proof later for the proof of Theorem 10.

Theorem 9 *Assume (A1) (A2), and (A4). OBL-G$_\flat$'s x_k-sequence exhibits the rate*

$$\|\nabla f(x_N)\|^2 \le 4L \frac{N^2 + N - \sqrt{2N(N+1)}}{N^2(N+1)^2 - 2\sqrt{2N(N+1)}} (f(x_0) - f_\star) \le \frac{4L}{N^2}(f(x_0) - f_\star).$$

Proof For $k = 1, 2 \dots, N-1$, define

$$\begin{aligned} U_k = &\frac{1}{(N-k+1)(N-k+2)} \left(\frac{1}{2L} \|\nabla f(x_k)\|^2 + f(x_k) - f(x_N) - \langle \nabla f(x_k), x_k - y_k \rangle \right) \\ &+ \frac{4L}{(N-k)(N-k+1)(N-k+2)(N-k+3)} \langle z_k - y_k, z_k - x_N \rangle. \end{aligned}$$

and

$$U_N = \frac{1}{4L} \|\nabla f(x_N)\|^2, \qquad U_0 = \frac{N(N+1) - \sqrt{2N(N+1)}}{(N-1)N(N+1)(N+2)} (f(x_0) - f(x_N)).$$

Then, we have $U_k \overset{(*)}{\ge} U_{k+1}$, which implies

$$\begin{aligned} \frac{1}{4L} \|\nabla f(x_N)\|^2 = U_N \le \dots \le U_0 &= \frac{N(N+1) - \sqrt{2N(N+1)}}{(N-1)N(N+1)(N+2)} (f(x_0) - f(x_N)) \\ &\le \frac{N(N+1) - \sqrt{2N(N+1)}}{(N-1)N(N+1)(N+2)} \left(f(x_0) - f(x_\star) - \frac{1}{2L} \|\nabla f(x_N)\|^2 \right). \end{aligned}$$

To complete the proof, it remains to justify the (*) part. We defer the calculations to Appendix A.3. ■

Note that the proof only utilized inequalities in $\mathcal{I}_{\text{OBL-G}_\flat}$. The Lyapunov function in this proof was inspired by the Lyapunov function used in the analysis of OGM-G in (Lee et al., 2021).

3. This conjecture was recently resolved on page 7 of (Kim et al., 2023), employing the concept of H-duality.

Backtracking linesearch version. Define optimized backtracking linesearch - gradient norm (**OBL-G**), a line backtracking version of OBL-G$_\flat$, as follows. Initialize L_0 and $\eta > 1$. For $k = 1, 2, \dots, 0$ we define $x_{k+1}, y_{k+1}, z_{k+1}$ as

$$\begin{aligned}
y_{k+1} &= x_k - \frac{1}{L_{k+1}} \nabla f(x_k) \\
z_{k+1} &= z_k - \frac{1}{L_{k+1}} \frac{N-k+1}{2} \nabla f(x_k) \\
x_{k+1} &= \frac{N-k-2}{N-k+2} y_{k+1} + \frac{4}{N-k+2} z_{k+1}
\end{aligned}$$

and

$$\begin{aligned}
y_1 &= x_0 - \frac{1}{L_1} \nabla f(x_0) \\
z_1 &= z_0 - \frac{1}{L_1} \frac{1+\sqrt{\frac{N(N+1)}{2}}}{2} \nabla f(x_k) \\
x_1 &= \frac{N-2}{N+2} y_1 + \frac{4}{N+2} z_1.
\end{aligned}$$

with $L_{k+1} = \eta^{i_{k+1}} L_k$ where $y_0 = z_0 = x_0$. The backtracking linesearch finds the smallest i_{k+1} such that

$$\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L_{k+1}} \left\| \nabla f(x_k) - \nabla f(x_{k+1}) \right\|^2 + \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle \right) \geq 0.$$

Theorem 10 *Assume (A1), (A2), and (A4). OBL-G exhibits the rate as*

$$\begin{aligned}
&\frac{1}{4L_N^2} \left\| \nabla f(x_N) \right\|^2 \\
&\quad \leq -\sum_{k \in I} \frac{1}{(N-k)(N-k+1)} \left(\frac{1}{L_k} - \frac{1}{L_{k+1}} \right) \left(f(x_k) - \frac{1}{2} \left(\frac{1}{L_k} + \frac{1}{L_{k+1}} \right) \left\| \nabla f(x_k) \right\|^2 - f(x_N) \right) \\
&\qquad + \frac{1}{(N+1)^2} \left(f(x_0) - f(x_N) \right).
\end{aligned}$$

Proof Let $x_{-1} = x_0$. For $k = 1, 2 \dots, N-1$, define

$$\begin{aligned}
U_k = &\frac{1}{(N-k+1)(N-k+2)L_k} \left(\frac{1}{2L_k} \|\nabla f(x_k)\|^2 + f(x_k) - f(x_N) - \langle \nabla f(x_k), x_k - y_k \rangle \right) \\
&+ \frac{4}{(N-k)(N-k+1)(N-k+2)(N-k+3)} \langle z_k - y_k, z_k - x_N \rangle
\end{aligned}$$

and

$$U_N = \frac{1}{4L_N^2} \left\| \nabla f(x_N) \right\|^2, \qquad U_0 = \frac{1}{L_0} \frac{N(N+1) - \sqrt{2N(N+1)}}{(N-1)N(N+1)(N+2)} \left(f(x_0) - f(x_N) \right).$$

Then, we have

$$U_k - U_{k+1} \overset{(*)}{\geq} \frac{1}{(N-k)(N-k+1)} \left(\frac{1}{L_k} - \frac{1}{L_{k+1}} \right) \left(f(x_k) - \frac{1}{2} \left(\frac{1}{L_k} + \frac{1}{L_{k+1}} \right) \|\nabla f(x_k)\|^2 - f(x_N) \right),$$

which indicates

$$\frac{1}{4L_N^2} \|\nabla f(x_N)\|^2 + \sum_{k \in K} \frac{1}{(N-k)(N-k+1)} \left(\frac{1}{L_k} - \frac{1}{L_{k+1}} \right) \left(f(x_k) - \frac{1}{2} \left(\frac{1}{L_k} + \frac{1}{L_{k+1}} \right) \|\nabla f(x_k)\|^2 - f(x_N) \right)$$
$$\leq \cdots \leq \frac{1}{L_0} \frac{N(N+1) - \sqrt{2N(N+1)}}{(N-1)N(N+1)(N+2)} \left(f(x_0) - f(x_N) \right).$$

where K is defined as the set of having smooth factor-jump. Note that K would be a sparse set (informally) that is a subset of $\{1, 2, \dots, N\}$. The justification of $(*)$ is deferred to Appendix A.4. ■

5.3.1 Discussion

The prior PEP formulations of OGM-G by Kim and Fessler (2021) and of APPM by Kim (2021) share the bi-convex structure we encounter with OBL-G$_\flat$. APPM is an accelerated algorithm for reducing the magnitude of the output of a maximal monotone operator, and the bi-convexity seems to arise from using the squared gradient magnitude, rather than the function-value suboptimality, as the performance measure. Both OGM-G and APPM were obtained by numerically solving the bi-convex PEP.

More specifically, Kim and Fessler obtained OGM-G by solving a PEP formulation using the inequalities

$$\mathcal{I}_{\text{OGM-G}} = \left\{ f(x_k) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle + \frac{1}{2L} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 \right\}_{k=0}^{N-1}$$
$$\bigcup \left\{ f(x_N) \geq f(x_k) + \langle \nabla f(x_k), x_k - x_N \rangle + \frac{1}{2L} \|\nabla f(x_k) - \nabla f(x_N)\|^2 \right\}_{k=0}^{N}$$
$$\bigcup \left\{ f(x_N) \geq f_\star + \frac{1}{2L} \|\nabla f(x_N)\|^2 \right\}.$$

When the bi-convex optimization problem was solved through alternating minimization, the iterates would converge to OGM-, from many different starting points. Based on this numerical evidence, we presume OGM-G is $\mathcal{A}^\star$-optimal, i.e.,

$$\text{OGM-G} \overset{?}{=} \mathcal{A}^\star_N(\|\nabla f(x_N)\|^2, f(x_0) - f_\star \leq \frac{1}{2} L R^2, \mathcal{I}_{\text{OGM-G}}).$$

Kim and Fessler (2021) did prove

$$\mathcal{R}^\star(\|\nabla f(x_N)\|^2, f(x_0) - f_\star \leq \frac{1}{2} L R^2, \mathcal{I}_{\text{OGM-G}})$$
$$\leq \mathcal{R}(\text{OGM-G}, \|\nabla f(x_N)\|^2, f(x_0) - f_\star \leq \frac{1}{2} L R^2, \mathcal{I}_{\text{OGM-G}}) = \frac{L^2 R^2}{\tilde{\theta}_N^2},$$

so the conjecture is that the inequality holds with equality. Our numerical experiments for finding OBL-$G_\flat$ exhibit this same behavior, so we conjecture that OBL-$G_\flat$ is also $\mathcal{A}^\star$-optimal.

The rates of OGM-G and OBL-$G_\flat$ have the same leading-term constants, i.e. the limit of convergence rate's ratio when $k \to \infty$ is 1. Moreover, OBL-$G_\flat$ turns out to be a "memory-saving algorithm" in the sense of (Zhou et al., 2022), i.e., the coefficients of the algorithm have a non-inductive form and therefore do not need to be pre-computed. We clarify that although OBL-$G_\flat$ and M-OGM-G (Zhou et al., 2022) are similar in their forms, the two algorithms are distinct. In fact, the rate OBL-$G_\flat$ has a leading-term constant that is smaller (better) by a factor of 2 compared to that of M-OGM-G (Zhou et al., 2022).

6. Conclusion

In this work, we presented an algorithm design methodology based on the notion of $\mathcal{A}^\star$-optimality and handy inequalities. We demonstrated the effectiveness of this methodology by finding new algorithms utilizing randomized coordinate updates and backtracking linesearches that improve upon the prior state-of-the-art rates.

By making the dependence on inequalities explicit, the notion of $\mathcal{A}^\star$-optimality provides a more fine-grained understanding of the optimality algorithms, and we expect this idea to be broadly applicable to the analysis and design of optimization algorithms. Investigating $\mathcal{A}^\star$-optimal algorithms for setups with stochastic gradients (Taylor and Bach, 2019) and monotone operators and splitting methods (Bauschke and Combettes, 2011; Ryu and Boyd, 2016; Ryu et al., 2020; Ryu and Yin, 2022) are interesting directions of future work.

Acknowledgements

CP was supported by an undergraduate research internship in the second half of the 2020 Seoul National University College of Natural Sciences and the 2021 Student-Directed Education Regular Program. EKR was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)] and the Samsung Science and Technology Foundation (Project Number SSTF-BA2101-02). We thank Gyumin Roh and Shuvomoy Das Gupta for providing valuable feedback.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. *ICML*, 2016.

Heinz H Bauschke and Patrick L Combettes. Convex analysis and monotone operator theory in Hilbert spaces. *New York: Springer-Verlag*, 2011.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Etienne De Klerk, Francois Glineur, and Adrien B Taylor. Worst-case convergence analysis of inexact gradient and newton methods through semidefinite programming performance estimation. *SIAM Journal on Optimization*, 30(3):2053–2082, 2020.

Jelena Diakonikolas and Puqian Wang. Potential function-based framework for making the gradients small in convex and min-max optimization. *SIAM Journal on Optimization*, 32 (3):1668–1697, 2022.

Yoel Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.

Yoel Drori and Adrien Taylor. On the oracle complexity of smooth strongly convex minimization. *Journal of Complexity*, 68:101590, 2022.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014.

Mahyar Fazlyab, Alejandro Ribeiro, Manfred Morari, and Victor M Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018.

Guoyong Gu and Junfeng Yang. Tight sublinear convergence rate of the proximal point algorithm for maximal monotone inclusion problems. *SIAM Journal on Optimization*, 30 (3):1905–1921, 2020.

Bin Hu and Laurent Lessard. Dissipativity theory for Nesterov's accelerated method. *ICML*, 2017.

Donghwan Kim. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190(1-2):57–87, 2021.

Donghwan Kim and Jeffrey A Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1):81–107, 2016.

Donghwan Kim and Jeffrey A Fessler. Another look at the fast iterative shrinkage/thresholding algorithm (FISTA). *SIAM Journal on Optimization*, 28(1):223–250, 2018a.

Donghwan Kim and Jeffrey A Fessler. Generalizing the optimized gradient method for smooth convex minimization. *SIAM Journal on Optimization*, 28(2):1920–1950, 2018b.

Donghwan Kim and Jeffrey A Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, 188(1):192–219, 2021.

Jaeyeon Kim, Asuman Ozdaglar, Chanwoo Park, and Ernest K Ryu. Time-reversed dissipation induces duality between minimizing gradient norm and function value. *NeurIPS*, 2023.

Jongmin Lee, Chanwoo Park, and Ernest K Ryu. A geometric structure of acceleration and its role in making gradients small fast. *NeurIPS*, 2021.

Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. *FOCS*, 2013.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Felix Lieder. On the convergence rate of the Halpern-iteration. *Optimization Letters*, 15(2): 405–418, 2021.

Arkadi Semenovich Nemirovsky and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization.* 1983.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Yurii Nesterov and Sebastian U Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017.

Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, pages 1–38, 2020.

Chanwoo Park, Jisun Park, and Ernest K Ryu. Factor-2 acceleration of accelerated gradient methods. *Applied Mathematics & Optimization*, 88(3):77, 2023.

Ernest K Ryu and Stephen Boyd. Primer on monotone operator methods. *Applied and Computational Mathematics*, 15(1):3–43, 2016.

Ernest K Ryu and Wotao Yin. *Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators.* Cambridge University Press, 2022.

Ernest K Ryu, Adrien B Taylor, Carolina Bergeling, and Pontus Giselsson. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *SIAM Journal on Optimization*, 30(3):2251–2271, 2020.

Jacob H Seidman, Mahyar Fazlyab, Victor M Preciado, and George J Pappas. A control-theoretic approach to analysis and parameter selection of Douglas–Rachford splitting. *IEEE Control Systems Letters*, 4(1):199–204, 2019.

Adrien Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, 199(1-2):557–594, 2023.

Adrien B Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. *COLT*, 2019.

Adrien B Taylor, Julien M Hendrickx, and François Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3):1283–1313, 2017a.

Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017b.

Adrien B Taylor, Julien M Hendrickx, and François Glineur. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *Journal of Optimization Theory and Applications*, 178(2):455–476, 2018.

Bryan Van Scoy, Randy A Freeman, and Kevin M Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2017.

TaeHo Yoon and Ernest K Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. *ICML*, 2021.

Guodong Zhang, Xuchan Bao, Laurent Lessard, and Roger Grosse. A unified analysis of first-order methods for smooth games via integral quadratic constraints. *Journal of Machine Learning Research*, 22(103):1–39, 2021.

Kaiwen Zhou, Lai Tian, Anthony Man-Cho So, and James Cheng. Practical schemes for finding near-stationary points of convex finite-sums. *AISTATS*, 2022.

Appendix A. Deferred calculations

A.1 Missing part of Corollary 5

For U_k, we have

$$
\begin{aligned}
&U_k - U_{k+1} \\
&\quad= \frac{(k+1)(k+2)}{2}\left(f(x_k) - f_\star - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2\right) - \frac{(k+2)(k+3)}{2}\left(f(x_{k+1}) - f_\star - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2\right) \\
&\qquad + \frac{L}{2}\left\|z_{k+1} - x_\star\right\|^2 - \frac{L}{2}\left\|z_{k+2} - x_\star\right\|^2 \\
&\quad= \frac{(k+1)(k+2)}{2}\left(f(x_k) - f_\star - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2\right) - \frac{(k+2)(k+3)}{2}\left(f(x_{k+1}) - f_\star - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2\right) \\
&\qquad - \langle (k+2)\nabla f(x_{k+1}), x_\star - z_{k+1}\rangle - \frac{(k+2)^2}{2L}\left\|\nabla f(x_{k+1})\right\|^2 \\
&\quad= \frac{(k+1)(k+2)}{2}\left(f(x_k) - f_\star - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2\right) - \frac{(k+2)(k+3)}{2}\left(f(x_{k+1}) - f_\star + \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2\right) \\
&\qquad - \langle (k+2)\nabla f(x_{k+1}), x_\star - z_{k+1}\rangle + \frac{k+2}{2L}\left\|\nabla f(x_{k+1})\right\|^2 \\
&\quad= \frac{(k+1)(k+2)}{2}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2\right) \\
&\qquad - (k+2)\left(f(x_{k+1}) - f_\star + \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2\right) \\
&\qquad\qquad - \langle (k+2)\nabla f(x_{k+1}), x_\star - z_{k+1}\rangle + \frac{k+2}{2L}\left\|\nabla f(x_{k+1})\right\|^2 \\
&\quad= \frac{(k+1)(k+2)}{2}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 - \frac{1}{2L}\left\|\nabla f(x_{k+1})\right\|^2\right) \\
&\qquad - (k+2)\left(f(x_{k+1}) - f_\star\right) \\
&\qquad\qquad - \langle (k+2)\nabla f(x_{k+1}), x_\star - x_{k+1}\rangle - \langle (k+2)\nabla f(x_{k+1}), x_{k+1} - z_{k+1}\rangle \\
&\quad= \frac{(k+1)(k+2)}{2}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\left\|\nabla f(x_k) - \nabla f(x_{k+1})\right\|^2 + \langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle\right) \\
&\qquad + (k+2)\left(f_\star - f(x_{k+1}) - \langle \nabla f(x_{k+1}), x_\star - x_{k+1}\rangle\right) \\
&\quad\geq 0
\end{aligned}
$$

which completes the proof of Corollary 5.

A.2 Missing part of Theorem 6

For U_k, we have

$$
\begin{aligned}
&U_k - U_{k+1} \\
&= \frac{(k+1)(k+2)}{2L_k}\left(f(x_k) - f_\star - \frac{1}{2L_k}\|\nabla f(x_k)\|^2\right) - \frac{(k+2)(k+3)}{2L_{k+1}}\left(f(x_{k+1}) - f_\star - \frac{1}{2L_{k+1}}\|\nabla f(x_{k+1})\|^2\right) \\
&\quad + \frac{1}{2}\|z_{k+1} - x_\star\|^2 - \frac{1}{2}\|z_{k+2} - x_\star\|^2 \\
&= \frac{(k+1)(k+2)}{2L_k}\left(f(x_k) - f_\star - \frac{1}{2L_k}\|\nabla f(x_k)\|^2\right) - \frac{(k+2)(k+3)}{2L_{k+1}}\left(f(x_{k+1}) - f_\star\right) \\
&\quad - \frac{1}{L_{k+1}}\langle (k+2)\nabla f(x_{k+1}), x_\star - z_{k+1}\rangle - \frac{(k+2)^2}{2L_{k+1}^2}\|\nabla f(x_{k+1})\|^2 + \frac{(k+2)(k+3)}{4L_{k+1}^2}\|\nabla f(x_{k+1})\|^2 \\
&= \frac{(k+1)(k+2)}{2L_{k+1}}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L_k}\|\nabla f(x_k)\|^2\right) \\
&\quad + \left(\frac{(k+1)(k+2)}{2L_k} - \frac{(k+1)(k+2)}{2L_{k+1}}\right)\left(f(x_k) - f_\star - \frac{1}{2L_k}\|\nabla f(x_k)\|^2\right) \\
&\qquad - \frac{1}{L_{k+1}}\langle (k+2)\nabla f(x_{k+1}), x_\star - z_{k+1}\rangle - \frac{(k+2)^2}{2L_{k+1}^2}\|\nabla f(x_{k+1})\|^2 + \frac{(k+2)(k+3)}{4L_{k+1}^2}\|\nabla f(x_{k+1})\|^2 \\
&\qquad\quad - \frac{k+2}{L_{k+1}}\left(f(x_{k+1}) - f_\star\right) \\
&= \frac{(k+1)(k+2)}{2L_{k+1}}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L_k}\|\nabla f(x_k)\|^2\right) \\
&\quad + \frac{(k+2)}{L_{k+1}}\left(f_\star - f(x_{k+1})\right) - \frac{(k+2)^2}{2L_{k+1}^2}\|\nabla f(x_{k+1})\|^2 + \frac{(k+2)(k+3)}{4L_{k+1}^2}\|\nabla f(x_{k+1})\|^2 \\
&\qquad - \frac{1}{L_{k+1}}\langle (k+2)\nabla f(x_{k+1}), x_\star - x_{k+1}\rangle - \frac{1}{L_{k+1}}\langle (k+2)\nabla f(x_{k+1}), x_{k+1} - z_{k+1}\rangle \\
&\qquad\quad + \left(\frac{(k+1)(k+2)}{2L_k} - \frac{(k+1)(k+2)}{2L_{k+1}}\right)\left(f(x_k) - f_\star - \frac{1}{2L_k}\|\nabla f(x_k)\|^2\right) \\
&= \frac{(k+1)(k+2)}{2L_{k+1}}\left(f(x_k) - f(x_{k+1}) - \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle - \frac{1}{2L_{k+1}}\|\nabla f(x_k) - \nabla f(x_{k+1})\|^2\right) \\
&\quad + \frac{(k+1)(k+2)}{2L_{k+1}}\left(-(\frac{1}{2L_k} - \frac{1}{2L_{k+1}})\|\nabla f(x_k)\|^2 + \frac{1}{2L_{k+1}}\|\nabla f(x_{k+1})\|^2\right) + \frac{(k+2)(k+3)}{4L_{k+1}^2}\|\nabla f(x_{k+1})\|^2 \\
&\qquad + \frac{(k+2)}{L_{k+1}}\left(f_\star - f(x_{k+1}) - \langle \nabla f(x_{k+1}), x_\star - x_{k+1}\rangle\right) - \frac{(k+2)^2}{2L_{k+1}^2}\|\nabla f(x_{k+1})\|^2 \\
&\qquad\quad + \left(\frac{(k+1)(k+2)}{2L_k} - \frac{(k+1)(k+2)}{2L_{k+1}}\right)\left(f(x_k) - f_\star - \frac{1}{2L_k}\|\nabla f(x_k)\|^2\right) \\
&\geq \frac{(k+1)(k+2)}{2L_{k+1}}\left(-(\frac{1}{2L_k} - \frac{1}{2L_{k+1}})\|\nabla f(x_k)\|^2 + \frac{1}{2L_{k+1}}\|\nabla f(x_{k+1})\|^2\right) \\
&\quad - \frac{(k+2)^2}{2L_{k+1}^2}\|\nabla f(x_{k+1})\|^2 + \frac{(k+2)(k+3)}{4L_{k+1}^2}\|\nabla f(x_{k+1})\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\left(\frac{(k+1)(k+2)}{2L_k}-\frac{(k+1)(k+2)}{2L_{k+1}}\right)\left(-\frac{1}{2L_k}\left\|\nabla f(x_k)\right\|^2\right)\\
&=\frac{(k+1)(k+2)}{4}\left(\frac{1}{L_{k+1}^2}-\frac{1}{L_k^2}\right)\left\|\nabla f(x_k)\right\|^2
\end{aligned}
$$

which completes the proof of Theorem 6.

A.3 Missing part of Theorem 9

For $k = 1, 2, \ldots, N-2$, we have

$$
\begin{aligned}
&\frac{4L}{(N-k)(N-k+1)(N-k+2)(N-k+3)} \langle z_k - y_k, z_k - x_N \rangle \\
&\quad - \frac{4L}{(N-k-1)(N-k)(N-k+1)(N-k+2)} \langle z_{k+1} - y_{k+1}, z_{k+1} - x_N \rangle \\
&= \frac{4L}{(N-k)(N-k+1)(N-k+2)} \left(\left\langle \frac{1}{N-k+3} (z_k - y_k), z_k - x_N \right\rangle \right. \\
&\quad \left. - \left\langle \frac{1}{N-k-1} (z_{k+1} - y_{k+1}), z_{k+1} - x_N \right\rangle \right) \\
&= \frac{4L}{(N-k)(N-k+1)(N-k+2)} \left(\left\langle \frac{1}{N-k+3} (z_k - y_k), z_k - x_N \right\rangle \right. \\
&\quad \left. - \left\langle \frac{1}{N-k-1} (z_{k+1} - y_{k+1}), z_k - \frac{1}{L} \frac{N-k+1}{2} \nabla f(x_k) - x_N \right\rangle \right) \\
&= \frac{4L}{(N-k)(N-k+1)(N-k+2)} \left(\left\langle \frac{1}{N-k-1} (z_k - x_k), z_k - x_N \right\rangle \right. \\
&\quad \left. - \left\langle \frac{1}{N-k-1} (z_{k+1} - y_{k+1}), z_k - \frac{1}{L} \frac{N-k+1}{2} \nabla f(x_k) - x_N \right\rangle \right) \\
&= \frac{4L}{(N-k-1)(N-k)(N-k+1)(N-k+2)} \left(\langle z_k - x_k, z_k - x_N \rangle \right. \\
&\quad \left. - \left\langle z_k - x_k - \frac{1}{L} \frac{N-k-1}{2} \nabla f(x_k), z_k - x_N - \frac{1}{L} \frac{N-k+1}{2} \nabla f(x_k) \right\rangle \right) \\
&= \frac{4}{(N-k-1)(N-k)(N-k+1)(N-k+2)} \times \\
&\quad \left\langle \nabla f(x_k), (N-k) z_k - \frac{N-k-1}{2} x_N - \frac{N-k+1}{2} x_k - \frac{1}{L} \frac{N-k-1}{2} \frac{N-k+1}{2} \nabla f(x_k) \right\rangle \\
&= \frac{4}{(N-k-1)(N-k)(N-k+1)(N-k+2)} \times \\
&\quad \left\langle \nabla f(x_k), (N-k)(z_k - x_k) + \frac{N-k-1}{2} (x_k - x_N) - \frac{1}{L} \frac{N-k-1}{2} \frac{N-k+1}{2} \nabla f(x_k) \right\rangle \\
&= \frac{4}{(N-k-1)(N-k)(N-k+1)(N-k+2)} \times \\
&\quad \left\langle \nabla f(x_k), \frac{(N-k)(N-k-1)}{4} (x_k - y_k) + \frac{N-k-1}{2} (x_k - x_N) - \frac{1}{L} \frac{N-k-1}{2} \frac{N-k+1}{2} \nabla f(x_k) \right\rangle .
\end{aligned}
$$

Therefore, for U_k, we have

$$
\begin{aligned}
&U_k - U_{k+1}\\
&= \frac{1}{(N-k+1)(N-k+2)}\left(\frac{1}{2L}\|\nabla f(x_k)\|^2 + f(x_k) - f(x_N) - \langle \nabla f(x_k), x_k - y_k\rangle\right)\\
&- \frac{1}{(N-k)(N-k+1)}\left(\frac{1}{2L}\|\nabla f(x_{k+1})\|^2 + f(x_{k+1}) - f(x_N) - \langle \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle\right)\\
&+ \frac{4}{(N-k-1)(N-k)(N-k+1)(N-k+2)}\times\\
&\left\langle \nabla f(x_k), \frac{(N-k)(N-k-1)}{4}(x_k - y_k) + \frac{N-k-1}{2}(x_k - x_N) - \frac{1}{L}\frac{N-k-1}{2}\frac{N-k+1}{2}\nabla f(x_k)\right\rangle\\
&= \frac{1}{(N-k)(N-k+1)}\left(f(x_k) - f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle - \frac{1}{2L}\|\nabla f(x_k)\|^2 - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2\right)\\
&+ \frac{2}{(N-k)(N-k+1)(N-k+2)}\left(f(x_N) - f(x_k) + \langle \nabla f(x_k), x_k - x_N\rangle\right),
\end{aligned}
$$

which completes the proof of Theorem 9.

A.4 Missing part of Theorem 10

For U_k, we have

$$
\begin{aligned}
&U_k - U_{k+1}\\
&= \frac{1}{(N-k+1)(N-k+2)L_k}\left(\frac{1}{2L_k}\|\nabla f(x_k)\|^2 + f(x_k) - f(x_N) - \langle \nabla f(x_k), x_k - y_k\rangle\right)\\
&- \frac{1}{(N-k)(N-k+1)L_{k+1}}\left(\frac{1}{2L_{k+1}}\|\nabla f(x_{k+1})\|^2 + f(x_{k+1}) - f(x_N) - \langle \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle\right)\\
&+ \frac{4}{(N-k-1)(N-k)(N-k+1)(N-k+2)L_k}\times\\
&\quad\left\langle \nabla f(x_k), \frac{(N-k)(N-k-1)}{4}(x_k - y_k) + \frac{N-k-1}{2}(x_k - x_N) - \frac{1}{L_k}\frac{N-k-1}{2}\frac{N-k+1}{2}\nabla f(x_k)\right\rangle.\\
&= \frac{1}{(N-k)(N-k+1)L_{k+1}}\Bigg(f(x_k) - f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle\\
&\quad - \frac{1}{2L_{k+1}}\|\nabla f(x_k)\|^2 - \frac{1}{2L_{k+1}}\|\nabla f(x_{k+1})\|^2\Bigg)\\
&+ \frac{2}{(N-k)(N-k+1)(N-k+2)L_k}\left(f(x_N) - f(x_k) + \langle \nabla f(x_k), x_k - x_N\rangle\right)\\
&+ \frac{1}{(N-k)(N-k+1)}\left(\frac{1}{L_k} - \frac{1}{L_{k+1}}\right)\left(f(x_k) - \frac{1}{2}\left(\frac{1}{L_k} + \frac{1}{L_{k+1}}\right)\|\nabla f(x_k)\|^2 - f(x_N)\right)\\
&\geq \frac{1}{(N-k)(N-k+1)}\left(\frac{1}{L_k} - \frac{1}{L_{k+1}}\right)\left(f(x_k) - \frac{1}{2}\left(\frac{1}{L_k} + \frac{1}{L_{k+1}}\right)\|\nabla f(x_k)\|^2 - f(x_N)\right),
\end{aligned}
$$

which completes the proof of Theorem 10.

Appendix B. Proofs of $\mathcal{A}^\star$-optimality

In this section, we prove Theorems 4 and 7, and discuss Conjecture 8 using the PEP machinery. Again, to verify the lengthy calculations, we provide Matlab scripts verifying the analytical solutions of the SDPs: `https://github.com/chanwoo-park-official/A-star-map/`.

B.1 Proof of $\mathcal{A}^\star$-optimality of OBL-F$_\flat$

To obtain OBL-F as an $\mathcal{A}^\star$-optimal algorithm, set $f(x_N) - f_\star$ to be the performance measure and $\|x_0 - x_\star\| \le R$ to be the initial condition. Since the constraints and the objective of the problem are homogenous, we assume $R = 1$ without loss of generality. For the argument of homogeneous, we refer to (Drori and Teboulle, 2014; Kim and Fessler, 2016; Taylor et al., 2017b). We use the set of inequalities that are handy for backtracking linesearches:

$$\mathcal{I}_{\text{OBL-F}_\flat} = \left\{ f_{k-1,0} \ge f_{k,0} + \langle g_k, x_{k-1} - x_k \rangle + \frac{1}{2L} \|g_{k-1} - g_k\|^2 \right\}_{k=1}^{N} \bigcup \left\{ f_\star \ge f_{k,0} + \langle g_k, x_\star - x_k \rangle \right\}_{k=0}^{N}.$$

For calculating $\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{OBL-F}_\flat})$ with fixed $\mathcal{A}_N$, define the PEP with

$$\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{OBL-F}_\flat}) = \left(\begin{array}{lllll} \text{maximize} & f_{N,0} - f_\star & & & \\ \text{subject to} & 1 & \ge & \|x_0 - x_\star\|^2 & \\ & f_{k-1,0} & \ge & f_{k,0} + \langle g_k, x_{k-1} - x_k \rangle + \frac{1}{2L} \|g_{k-1} - g_k\|^2, & k \in \{1, 2, \dots, N\} \\ & f_\star & \ge & f_{k,0} + \langle g_k, x_\star - x_k \rangle, & k \in \{0, 1, \dots, N\} \\ & x_k & & \text{is following the algorithm } \mathcal{A}_N. & \end{array} \right)$$

Using the notation of Section 4.2, we reformulate the problem of computing the risk $\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{OBL-F}_\flat})$ as the following SDP:

$$\begin{array}{lll} \underset{\mathbf{G}, \mathbf{F}_0}{\text{maximize}} & \boldsymbol{f}_N^\intercal \mathbf{F}_0 & \\ \text{subject to} & 1 \ge \boldsymbol{x}_0^\intercal \mathbf{G} \boldsymbol{x}_0 & \\ & 0 \ge (\boldsymbol{f}_k - \boldsymbol{f}_{k-1})^\intercal \mathbf{F}_0 + \boldsymbol{g}_k^\intercal \mathbf{G} (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) + \dfrac{1}{2L} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal \mathbf{G} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k), & k \in \{1, 2, \dots, N\} \\ & 0 \ge \boldsymbol{f}_k^\intercal \mathbf{F}_0 - \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{x}_k, & k \in \{0, 1, \dots, N\}. \\ & \mathbf{G} \succeq 0, \mathbf{F}_0 \ge 0 & \end{array}$$

For above transformation, $d \ge N + 2$ is used (Taylor et al., 2017b). The Lagrangian of the optimization problem becomes

$$\begin{aligned} &\Lambda(\mathbf{F}_0, \mathbf{G}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \tau) \\ &\quad = -\boldsymbol{f}_N^\intercal \mathbf{F}_0 + \tau (\boldsymbol{x}_0^\intercal \mathbf{G} \boldsymbol{x}_0 - 1) + \sum_{k=0}^{N} \beta_k \left(\boldsymbol{f}_k^\intercal \mathbf{F}_0 - \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{x}_k \right) \\ &\quad + \sum_{k=1}^{N} \lambda_k \left((\boldsymbol{f}_k - \boldsymbol{f}_{k-1})^\intercal \mathbf{F}_0 + \boldsymbol{g}_k^\intercal \mathbf{G} (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) + \frac{1}{2L} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal \mathbf{G} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k) \right) \end{aligned}$$

with dual variables $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_N) \in \mathbb{R}^N_+$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_N) \in \mathbb{R}^{N+1}_+$, and $\tau \geq 0$.

Then the dual formulation of PEP problem is

$$\begin{array}{ll} \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},\tau)\geq\mathbf{0}}{\text{maximize}} & -\tau \\ \text{subject to} & \mathbf{0} = -\boldsymbol{f}_N + \sum_{k=1}^{N} \lambda_k(\boldsymbol{f}_k - \boldsymbol{f}_{k-1}) + \sum_{k=0}^{N} \beta_k \boldsymbol{f}_k \\ & 0 \preceq S(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau), \end{array} \tag{25}$$

where S is defined as

$$\begin{aligned} S(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau) &= \tau \boldsymbol{x}_0 \boldsymbol{x}_0^\intercal + \sum_{k=0}^{N} \frac{\beta_k}{2} \left(-\boldsymbol{g}_k \boldsymbol{x}_k^\intercal - \boldsymbol{x}_k \boldsymbol{g}_k^\intercal \right) \\ &\quad + \sum_{k=1}^{N} \frac{\lambda_k}{2} \left(\boldsymbol{g}_k (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k)^\intercal + (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) \boldsymbol{g}_k^\intercal + \frac{1}{L} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal \right). \end{aligned}$$

We have a strong duality argument

$$\underset{s_{i,j}}{\arg\min}\ \underset{\mathbf{G},\mathbf{F}_0}{\text{maximize}}\ \ \boldsymbol{f}_{N+1}^\intercal \mathbf{F}_0 = \underset{h_{i,j}}{\arg\min}\ \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},\tau)\geq\mathbf{0}}{\text{minimize}}\ \ \tau,$$

as ORC-F's optimality proof. Remind that (25) finds the "best" proof for the algorithm. Now we investigate the optimization step for algorithm. The last part is minimizing (25) with stepsize, i.e.

$$\underset{h_{i,j}}{\text{minimize}}\ \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},,\tau)\geq\mathbf{0}}{\text{maximize}} \quad \tau \tag{26}$$

$$\text{subject to} \quad \mathbf{0} = -\boldsymbol{f}_N + \sum_{k=1}^{N} \lambda_k(\boldsymbol{f}_k - \boldsymbol{f}_{k-1}) + \sum_{k=0}^{N} \beta_k \boldsymbol{f}_k \tag{27}$$

$$0 \preceq S(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau). \tag{28}$$

We note that $\boldsymbol{f}_i$ is a standard unit vector mentioned in (7) (not a variable), we can write (27) as

$$\left(\begin{array}{ll} \beta_k = \lambda_{k+1} - \lambda_k, & k \in \{1, \dots, N-1\} \\ \beta_0 = \lambda_1 & \\ \beta_N = 1 - \lambda_N. & \end{array} \right) \tag{29}$$

We consider (28) with (29) and FSFO's $h_{i,j}$. To be specific, we substitute $\boldsymbol{\beta}$ to $\boldsymbol{\lambda}$ in $S(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau)$. To show the dependency of S to $(h_{i,j})$ since $\boldsymbol{x}_k$ are represented with $(h_{i,j})$, we will explicitly write S as $S(\boldsymbol{\lambda}, \tau; (h_{i,j}))$. Then, we get

$$\begin{aligned} S(\boldsymbol{\lambda}, \tau; (h_{i,j})) &= \tau \boldsymbol{x}_0 \boldsymbol{x}_0^\intercal + \sum_{k=1}^{N} \frac{\lambda_k}{2L} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal \\ &+ \sum_{k=1}^{N-1} \sum_{t=0}^{k-1} \left(\frac{\lambda_k}{2} \frac{h_{k,t}}{L} + \frac{\lambda_{k+1} - \lambda_k}{2} \sum_{j=t+1}^{k} \frac{h_{j,t}}{L} \right) (\boldsymbol{g}_k \boldsymbol{g}_t^\intercal + \boldsymbol{g}_t \boldsymbol{g}_k^\intercal) \\ &+ \sum_{t=0}^{N-1} \left(\frac{\lambda_N}{2} \frac{h_{N,t}}{L} + \frac{1 - \lambda_N}{2} \sum_{j=t+1}^{N} \frac{h_{j,t}}{L} \right) (\boldsymbol{g}_N \boldsymbol{g}_t^\intercal + \boldsymbol{g}_t \boldsymbol{g}_N^\intercal) \\ &- \sum_{k=1}^{N-1} \frac{\lambda_{k+1} - \lambda_k}{2} \left(\boldsymbol{x}_0 \boldsymbol{g}_k^\intercal + \boldsymbol{g}_k \boldsymbol{x}_0^\intercal \right) - \frac{\lambda_1}{2} \left(\boldsymbol{x}_0 \boldsymbol{g}_0^\intercal + \boldsymbol{g}_0 \boldsymbol{x}_0^\intercal \right) - \frac{1 - \lambda_N}{2} \left(\boldsymbol{x}_0 \boldsymbol{g}_N^\intercal + \boldsymbol{g}_N \boldsymbol{x}_0^\intercal \right). \end{aligned}$$

Using the fact that $\boldsymbol{x}_0, \boldsymbol{g}_i, \boldsymbol{f}_i$ are unit vectors, we can represent $S(\boldsymbol{\lambda}, \tau; (h_{i,j}))$ with $\boldsymbol{\gamma}(\boldsymbol{\lambda}) = -L\boldsymbol{\beta} = -L(\lambda_1, \lambda_2 - \lambda_1, \ldots, 1 - \lambda_N) = (\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}), \gamma_N(\boldsymbol{\lambda}))$ and $\tau' = 2L\tau$ as

$$S(\boldsymbol{\lambda}, \tau'; (h_{i,j})) = \frac{1}{L} \begin{pmatrix} \frac{1}{2}\tau' & \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})^\intercal & \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) \\ \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) & Q(\boldsymbol{\lambda}; (h_{i,j})) & q(\boldsymbol{\lambda}; (h_{i,j})) \\ \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) & q(\boldsymbol{\lambda}; (h_{i,j}))^\intercal & \frac{\lambda_N}{2} \end{pmatrix} \succeq 0.$$

Here, Q and $\mathbf{q}$ are defined as

$$\begin{aligned} Q(\boldsymbol{\lambda}; (h_{i,j})) &= \sum_{k=1}^{N-1} \frac{\lambda_k}{2} (\boldsymbol{g}'_{k-1} - \boldsymbol{g}'_k)(\boldsymbol{g}'_{k-1} - \boldsymbol{g}'_k)^\intercal + \frac{\lambda_N}{2} \boldsymbol{g}'_{N-1} \boldsymbol{g}'^\intercal_{N-1} \\ &+ \sum_{k=1}^{N-1} \sum_{t=0}^{k-1} \left(\frac{\lambda_k}{2} h_{k,t} + \frac{\lambda_{k+1} - \lambda_k}{2} \sum_{j=t+1}^{k} h_{j,t} \right) \left(\boldsymbol{g}'_k \boldsymbol{g}'^\intercal_t + \boldsymbol{g}'_t \boldsymbol{g}'^\intercal_k \right) \end{aligned}$$

and

$$\begin{aligned} \mathbf{q}(\boldsymbol{\lambda}; (h_{i,j})) &= -\frac{\lambda_N}{2} \boldsymbol{g}'_{N-1} + \sum_{t=0}^{N-1} \left(\frac{\lambda_N}{2} h_{N,t} + \frac{1 - \lambda_N}{2} \sum_{j=t+1}^{N} h_{j,t} \right) \boldsymbol{g}'_t \\ &= \sum_{t=0}^{N-2} \left(\frac{\lambda_N}{2} h_{N,t} + \frac{1 - \lambda_N}{2} \sum_{j=t+1}^{N} h_{j,t} \right) \boldsymbol{g}'_t + \left(\frac{1}{2} h_{N,N-1} - \frac{\lambda_N}{2} \right) \boldsymbol{g}'_{N-1}. \end{aligned}$$

where $\boldsymbol{g}'_k = e_{k+1} \in \mathbb{R}^{N+1}$. Note that (26) is equivalent to

$$\begin{aligned} \underset{h_{i,j}}{\text{minimize}} \ \underset{(\boldsymbol{\lambda}, \tau') \geq \mathbf{0}}{\text{minimize}} \quad & \tau' \\ \text{subject to} \quad & \begin{pmatrix} \frac{1}{2}\tau' & \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})^\intercal & \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) \\ \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) & Q(\boldsymbol{\lambda}; (h_{i,j})) & q(\boldsymbol{\lambda}; (h_{i,j})) \\ \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) & q(\boldsymbol{\lambda}; (h_{i,j}))^\intercal & \frac{\lambda_N}{2} \end{pmatrix} \succeq 0 \end{aligned} \tag{30}$$

and dividing this optimized value with $2L$ gives the optimized value of (26). Using Schur complement (?), (30) can be converted to the problem as

$$\underset{h_{i,j}}{\text{minimize}}\ \underset{(\boldsymbol{\lambda},\tau')\geq\mathbf{0}}{\text{minimize}} \quad \tau' \tag{31}$$

$$\text{subject to} \quad \begin{pmatrix} Q - \frac{2\mathbf{q}\mathbf{q}^\intercal}{\lambda_N} & \frac{1}{2}(\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) - \frac{2\mathbf{q}\gamma_N(\boldsymbol{\lambda})}{\lambda_N}) \\ \frac{1}{2}(\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) - \frac{2\mathbf{q}\gamma_N(\boldsymbol{\lambda})}{\lambda_N})^\intercal & \frac{1}{2}(\tau' - \frac{\gamma_N(\boldsymbol{\lambda})^2}{\lambda_N}) \end{pmatrix} \succeq 0. \tag{32}$$

So far we simplified SDP. We will have three steps: finding variables that make (32)'s left hand side zero, showing that the solution from the first step satisfies KKT condition, and finally showing that obtained algorithm is equivalent to OBL-F$_\flat$.

Claim 5 *There is a point that makes* (32)*'s left-hand side zero.*

Proof Defining $\{r_{k,t}\}_{k=1,2,\dots,N,t=0,\dots,k-1}$ as

$$r_{k,t} = \lambda_k h_{k,t} - \frac{\gamma_k}{L}\sum_{j=t+1}^{k} h_{j,t}.$$

Then, if $r_{i,j}$ is determined, (Drori and Teboulle, 2014, Theorem 5.1) indicates this uniquely determine $h_{i,j}$. We set $s_N = \frac{N(N+1)}{2}$, $T = \frac{1}{s_N+\sqrt{s_N}}$, and set $(\lambda_k)_{k=0}^N$ and $(r_{N,k})_{k=0}^{N-1}$ as

$$\begin{aligned} \lambda_k &= \frac{k(k+1)}{2}T, \qquad k \in \{1,2,\dots,N\} \\ r_{N,k} &= \frac{k+1}{\sqrt{s_N}+1}, \qquad k \in \{0,2,\dots,N-2\} \\ r_{N,N-1} - \lambda_N &= \frac{N}{\sqrt{s_N}+1}. \end{aligned} \tag{33}$$

Moreover, we set

$$\begin{aligned} r_{k,t} &= \frac{(k+1)(t+1)}{s_N+\sqrt{s_N}}, \qquad k \in \{1,2,\dots,N-1\}, \quad t \in \{0,1,\dots,k-2\} \\ r_{k,k-1} &= \frac{k(k+1)+s_N}{s_N+\sqrt{s_N}}, \qquad k \in \{1,2\dots,N-1\}. \end{aligned} \tag{34}$$

In addition, we set $\hat{\boldsymbol{\gamma}}$ as

$$\begin{aligned} \gamma_t &= \frac{\gamma_N}{\lambda_N} r_{N,t}, \qquad t \in \{0,1,\dots,N-2\} \\ \gamma_{N-1} &= \frac{\gamma_N(r_{N,N-1}-\lambda_N)}{\lambda_N}. \end{aligned}$$

Lastly, we set τ' as

$$\tau' = \frac{L^2}{s_N+\sqrt{s_N}}, \tag{35}$$

and $\tau = \frac{L}{2(s_N+\sqrt{s_N})} = \frac{L}{N(N+1)+\sqrt{2N(N+1)}}$. These variables make (32)'s left-hand side zero. ■

Claim 6 (33), (34) *and* (35) *are an optimal solution of* (31).

Proof Let we represent S with the variable $(r_{i,j})$. We will denote this as $\mathbf{A}$. To be specific,

$$\mathbf{A}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau', (r_{i,j})) = S(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \tau'; (h_{i,j})) = \begin{pmatrix} \frac{1}{2}\tau' & -\frac{L}{2}\hat{\boldsymbol{\beta}}^\intercal & -\frac{L}{2}\beta_N \\ -\frac{L}{2}\hat{\boldsymbol{\beta}}^\intercal & Q(\boldsymbol{\lambda}; (r_{i,j})) & \mathbf{q}((r_{i,j})) \\ -\frac{L}{2}\beta_N & \mathbf{q}((r_{i,j}))^\intercal & \frac{\lambda_N}{2} \end{pmatrix} \succeq 0.$$

Here, $\boldsymbol{\beta} = (\hat{\boldsymbol{\beta}}^\intercal, \beta_N)^\intercal$,

$$Q(\boldsymbol{\lambda}; (r_{i,j})) = \sum_{k=1}^{N-1} \frac{\lambda_k}{2}(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal + \frac{\lambda_N}{2}\boldsymbol{g}_{N-1}\boldsymbol{g}_{N-1}^\intercal + \sum_{k=1}^{N-1}\sum_{t=0}^{k-1} \frac{r_{k,t}}{2}\left(\boldsymbol{g}_k\boldsymbol{g}_t^\intercal + \boldsymbol{g}_t\boldsymbol{g}_k^\intercal\right)$$

and

$$\mathbf{q}((r_{i,j})) = \sum_{t=0}^{N-1} \frac{r_{N,t}}{2}\boldsymbol{g}_t - \frac{\lambda_N}{2}\boldsymbol{g}_{N-1}.$$

Define a linear SDP relaxation of (30) as

$$\begin{aligned} \underset{r_{i,j}}{\text{minimize}}\ \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},\tau')\geq \mathbf{0}}{\text{minimize}}\quad & \tau' \\ \text{subject to}\quad & \mathbf{A}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau', (r_{i,j})) \succeq 0. \\ & \mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau') = \left(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau'\right) \geq 0 \\ & \mathbf{C}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = (-\lambda_1 + \beta_0, \lambda_1 - \lambda_2 + \beta_1, \dots, \lambda_{N-1} - \lambda_N + \beta_{N-1}, -1 + \lambda_N + \beta_N) = 0. \end{aligned} \tag{36}$$

(Drori and Teboulle, 2014, Theorem 3) indicates that if we prove the choice in the previous claim satisfies KKT condition of (36), then this is also an optimal solution for the original problem since $(r_{i,j})$ uniquely determines $h_{i,j}$. The Lagrangian of the minimization problem is

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau', (r_{i,j}), \mathbf{K}, \mathbf{b}, \mathbf{c}) = \frac{1}{2}\tau' - \operatorname{tr}\left\{\mathbf{A}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau', (r_{i,j}))\mathbf{K}\right\} - \mathbf{b}^\intercal\mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau') - \mathbf{c}^\intercal\mathbf{C}(\boldsymbol{\lambda}, \boldsymbol{\beta})$$

and the KKT conditions of the minimization problems are

$$\begin{aligned} & \mathbf{A}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau'; (r_{i,j})) \succeq 0, \mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau') \geq 0, \mathbf{C}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = 0, \\ & \nabla_{(\boldsymbol{\lambda},\boldsymbol{\beta},\tau',(r_{i,j}))}\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau', (r_{i,j}), \mathbf{K}, \mathbf{b}, \mathbf{c}) = 0, \\ & \mathbf{K} \succeq 0, \mathbf{b} \geq 0, \\ & \operatorname{tr}\left\{\mathbf{A}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau', (r_{i,j}))\mathbf{K}\right\} = 0, \mathbf{b}^\intercal\mathbf{B}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau') = 0, \end{aligned}$$

where $\mathbf{K}$ is a symmetric matrix. Here, $\mathbf{b} = (\mathbf{u}, \mathbf{v}, s)$. We re-index K's column and row starting from -1 (so K's rows and columns index are $\{-1, 0, 1. \dots, N\}$). Now, we will show that there exist a dual optimal solution $(\mathbf{K}, \mathbf{b}, \mathbf{c})$ that $(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau', (r_{i,j}), \mathbf{K}, \mathbf{b}, \mathbf{c})$ satisfies KKT condition, which proves a pair $(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau', (r_{i,j}))$ is an optimal solution for primal problem. The stationary condition $\nabla_{(\boldsymbol{\lambda},\boldsymbol{\beta},\tau',(r_{i,j}))}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\beta},\tau',(r_{i,j}),\mathbf{K},\mathbf{b},\mathbf{c})=0$ can be rewritten as

$$\begin{aligned}
&\frac{\partial\mathcal{L}}{\partial\lambda_k}=-\frac{1}{2}\left(K_{k-1,k-1}-K_{k-1,k}-K_{k,k-1}+K_{k,k}\right)-u_k+c_{k-1}-c_k=0, \qquad k\in\{1,2,\dots,N\}\\
&\frac{\partial\mathcal{L}}{\partial\beta_k}=\frac{L}{2}\left(K_{-1,k}+K_{k,-1}\right)-v_k-c_k=0, \qquad k\in\{0,1,\dots,N\}\\
&\frac{\partial\mathcal{L}}{\partial\tau'}=\frac{1}{2}-\frac{1}{2}K_{-1,-1}-s=0\\
&\frac{\partial\mathcal{L}}{\partial r_{k,t}}=-\frac{1}{2}(K_{k,t}+K_{t,k})=0, \qquad k\in\{1,2,\dots,N\}, \quad t\in\{0,1,\dots,k-1\}.
\end{aligned} \tag{37}$$

We already know that $\mathbf{B}(\boldsymbol{\lambda},\boldsymbol{\beta},\tau')\neq 0$, we can set $\mathbf{b}=0$. Then, (37) reduces to

$$\begin{aligned}
&K_{k,t}=0, \qquad k\in\{1,2,\dots,N,t=0,1,\dots,k-1\}\\
&-\frac{1}{2}(K_{k-1,k-1}+K_{k,k})+c_{k-1}-c_k=0, \qquad k\in\{1,2,\dots,N\}\\
&LK_{-1,k}-c_k=0, \qquad k\in\{0,1,\dots,N\}\\
&K_{-1,-1}=1.
\end{aligned}$$

Then, we have

$$\mathbf{K}=\begin{pmatrix}
1 & \frac{c_0}{L} & \frac{c_1}{L} & \dots & \frac{c_{N-1}}{L} & \frac{c_N}{L}\\
\frac{c_0}{L} & K_{0,0} & 0 & \dots & 0 & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
\frac{c_{N-1}}{L} & 0 & 0 & \dots & K_{N-1,N-1} & 0\\
\frac{c_N}{L} & 0 & 0 & \dots & 0 & K_{N,N}
\end{pmatrix}\succeq 0$$

and since $\operatorname{tr}\{\mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\tau',(r_{i,j}))\mathbf{K}\}=0$ with $\mathbf{A}\succeq 0$, we can replace this condition by $\mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\tau',(r_{i,j}))\mathbf{K}=0$. Then the KKT condition for the given $(\boldsymbol{\lambda},\boldsymbol{\beta},\tau',(r_{i,j}))$ reduces to

$$\begin{aligned}
&\frac{1}{2}\tau'-\frac{1}{2}\boldsymbol{\beta}^\intercal\mathbf{c}=0\\
&\frac{1}{2L}\tau'\mathbf{c}-\frac{L}{2}\operatorname{diag}(K_{0,0},\dots,K_{N-1,N-1},K_{N,N})\boldsymbol{\beta}=0\\
&-\boldsymbol{\beta}\mathbf{c}^\intercal+\begin{pmatrix} Q & \mathbf{q}\\ \mathbf{q}^\intercal & \frac{1}{2}\lambda_N\end{pmatrix}\operatorname{diag}(K_{0,0},\dots,K_{N-1,N-1},K_{N,N})=0.
\end{aligned}$$

This clearly indicates that $K_{N,N}>0$ and c_i are all determined as positive, and $K\succeq 0$. ■

Claim 7 *The obtained algorithm is OBL-F$_\flat$.*

Proof By calculating $(h_{i,j})$ of OBL-F$_\flat$, we can prove the equivalence of the obtained solution and OBL-F$_\flat$. Indeed, OBL-F$_\flat$ is obtained by using (Lee et al., 2021)'s auxiliary sequences.

Except the last step of OBL-F$_\flat$, we will show that obtained $(\hat{h}_{i,j})$ satisfies

$$x_0 - \sum_{i=1}^{k+1}\sum_{j=0}^{i-1} \frac{\hat{h}_{i.j}}{L}\nabla f(x_j) = \left(1 - \frac{2}{k+3}\right)\left(x_0 - \sum_{i=1}^{k}\sum_{j=0}^{i-1}\frac{\hat{h}_{i,j}}{L}\nabla f(x_j) - \frac{1}{L}\nabla f(x_k)\right) + \frac{2}{k+3}\left(x_0 - \sum_{j=0}^{k}\frac{j+1}{L}\nabla f(x_j)\right),$$

which is re-written form of OBL-F$_\flat$. Comparing $\nabla f(x_j)$'s each coefficient, we should prove

$$\sum_{i=j+1}^{k+1}\hat{h}_{i,j} = \left(1-\frac{2}{k+3}\right)\sum_{i=j+1}^{k}\hat{h}_{i,j} + \frac{2}{k+3}(j+1) \qquad j \in \{0,1,\dots,k-1\}$$

$$\hat{h}_{k+1,k} = \left(1-\frac{2}{k+3}\right) + \frac{2}{k+3}(k+1),$$

which is exactly equal to the recursive rule of (Drori and Teboulle, 2014, Theorem 3). The last step of OBL-F$_\flat$ can be analyzed similarly. ■

In sum, the algorithm's performance criterion $f(x_N) - f_\star$ is bounded as

$$f(x_N) - f_\star \le \frac{L}{N(N+1)+\sqrt{2N(N+1)}}\left\|x_0 - x_\star\right\|^2.$$

Overall, we showed that OBL-F$_\flat$ is the "best" algorithm under $\mathcal{I}_{\text{OBL-F}_\flat}$.

$$\text{OBL-F}_\flat = \mathcal{A}^\star_N(f(x_N) - f_\star, \|x_0 - x_\star\| \le R, \mathcal{I}_{\text{OBL-F}_\flat})$$

and

$$\begin{aligned}\mathcal{R}(\text{OBL-F}_\flat, f(x_N) - f_\star, \|x_0 - x_\star\| \le R, \mathcal{I}_{\text{OBL-F}_\flat}) \\ &= \mathcal{R}^*(\mathfrak{A}_N, f(x_N) - f_\star, \|x_0 - x_\star\| \le R, \mathcal{I}_{\text{OBL-F}_\flat}) \\ &= \frac{LR^2}{N(N+1)+\sqrt{2N(N+1)}}\end{aligned}$$

hold.

B.2 Proof of $\mathcal{A}^\star$-optimality of FGM

To obtain FGM as an $\mathcal{A}^\star$-optimal algorithm, set $f(y_{N+1}) - f_\star$ to be the performance measure and $\|x_0 - x_\star\| \le R$ to be the initial condition. Since the constraints and the objective of the problem are homogenous, we assume $R = 1$ without loss of generality. For the argument of homogeneous, we refer to (Drori and Teboulle, 2014; Kim and Fessler, 2016; Taylor et al., 2017b). We use the set of inequalities that are handy for randomized coordinate updates and backtracking linesearches:

$$\mathcal{I}_{\text{FGM}} = \left\{f_{k,0} \ge f_{k+1,1} + \frac{1}{2L}\|g_k\|^2\right\}_{k=0}^{N} \bigcup \left\{f_{k,1} \ge f_{k,0} + \langle g_k, y_k - x_k\rangle\right\}_{k=1}^{N} \bigcup \left\{f_\star \ge f_{k,0} + \langle g_k, x_\star - x_k\rangle\right\}_{k=0}^{N}.$$

For calculating $\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{FGM}})$ with fixed $\mathcal{A}_N$, define the PEP with $\mathcal{I}_{\text{FGM}}$ as

$$\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{FGM}}) = \left(\begin{array}{llll} \text{maximize} & f_{N+1,1} & -f_\star & \\ \text{subject to} & 1 \quad \geq & \|x_0 - x_\star\|^2 & \\ & f_{k,0} \geq & f_{k+1,1} + \frac{1}{2L}\|g_k\|^2, & k \in \{0,1,\ldots,N\} \\ & f_{k,1} \geq & f_{k,0} + \langle g_k, y_k - x_k \rangle, & k \in \{1,\ldots,N\} \\ & f_\star \quad \geq & f_{k,0} + \langle g_k, x_\star - x_k \rangle, & k \in \{0,1,\ldots,N\} \\ & x_k, y_k & \text{are following the algorithm } \mathcal{A}_N. & \end{array} \right)$$

Using the notation of Section 4.2, we reformulate the problem of computing the risk $\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{FGM}})$ as the following SDP:

$$\begin{array}{lll} \underset{\mathbf{G},\mathbf{F}_0,\mathbf{F}_1}{\text{maximize}} & \boldsymbol{f}_{N+1}^\intercal \mathbf{F}_1 & \\ \text{subject to} & 1 \geq \boldsymbol{x}_0^\intercal \mathbf{G} \boldsymbol{x}_0 & \\ & 0 \geq \boldsymbol{f}_{k+1}^\intercal \mathbf{F}_1 - \boldsymbol{f}_k^\intercal \mathbf{F}_0 + \dfrac{1}{2L} \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{g}_k, & k \in \{0,1,\ldots,N\} \\ & 0 \geq \boldsymbol{f}_k^\intercal (\mathbf{F}_0 - \mathbf{F}_1) + \boldsymbol{g}_k^\intercal \mathbf{G}(\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) - \dfrac{1}{L} \boldsymbol{g}_{k-1}^\intercal \mathbf{G} \boldsymbol{g}_k, & k \in \{1,2,\ldots,N\} \\ & 0 \geq \boldsymbol{f}_k^\intercal \mathbf{F}_0 - \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{x}_k, & k \in \{0,1,\ldots,N\}. \\ & \mathbf{G} \succeq 0, \mathbf{F}_0 \geq 0, \mathbf{F}_1 \geq 0. & \end{array}$$

For above transformation, $d \geq N+2$ is used (Taylor et al., 2017b). The Lagrangian of the optimization problem becomes

$$\begin{aligned} &\Lambda(\mathbf{F}_0, \mathbf{F}_1, \mathbf{G}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau) \\ &\quad = -\boldsymbol{f}_{N+1}^\intercal \mathbf{F}_1 + \tau(\boldsymbol{x}_0^\intercal \mathbf{G} \boldsymbol{x}_0 - 1) + \sum_{k=0}^{N} \alpha_k \left(\boldsymbol{f}_{k+1}^\intercal \mathbf{F}_1 - \boldsymbol{f}_k^\intercal \mathbf{F}_0 + \frac{1}{2L} \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{g}_k \right) \\ &\quad + \sum_{k=1}^{N} \lambda_k \left(\boldsymbol{f}_k^\intercal (\mathbf{F}_0 - \mathbf{F}_1) + \boldsymbol{g}_k^\intercal \mathbf{G}(\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) - \frac{1}{L} \boldsymbol{g}_{k-1}^\intercal \mathbf{G} \boldsymbol{g}_k \right) \\ &\quad + \sum_{k=0}^{N} \beta_k \left(\boldsymbol{f}_k^\intercal \mathbf{F}_0 - \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{x}_k \right) \end{aligned}$$

with dual variables $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N) \in \mathbb{R}_+^N$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_N) \in \mathbb{R}_+^{N+1}$, $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_N) \in \mathbb{R}_+^{N+1}$, and $\tau \geq 0$. Then the dual formulation of PEP problem is

$$\begin{array}{ll} \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau) \geq \mathbf{0}}{\text{maximize}} & -\tau \\ \text{subject to} & \mathbf{0} = -\displaystyle\sum_{k=0}^{N} \alpha_k \boldsymbol{f}_k + \sum_{k=1}^{N} \lambda_k \boldsymbol{f}_k + \sum_{k=0}^{N} \beta_k \boldsymbol{f}_k \\ & \mathbf{0} = -\boldsymbol{f}_{N+1} - \displaystyle\sum_{k=1}^{N} \lambda_k \boldsymbol{f}_k + \sum_{k=0}^{N} \alpha_k \boldsymbol{f}_{k+1} \\ & 0 \preceq S(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau), \end{array} \tag{38}$$

where S is defined as

$$S(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau) = \tau\boldsymbol{x}_0\boldsymbol{x}_0^\intercal + \sum_{k=0}^{N}\alpha_k\left(\frac{1}{2L}\boldsymbol{g}_k\boldsymbol{g}_k^\intercal\right) + \sum_{k=0}^{N}\frac{\beta_k}{2}\left(-\boldsymbol{g}_k\boldsymbol{x}_k^\intercal - \boldsymbol{x}_k\boldsymbol{g}_k^\intercal\right)$$
$$+\sum_{k=1}^{N}\frac{\lambda_k}{2}\left(\boldsymbol{g}_k(\boldsymbol{x}_{k-1}-\boldsymbol{x}_k)^\intercal + (\boldsymbol{x}_{k-1}-\boldsymbol{x}_k)\boldsymbol{g}_k^\intercal - \frac{1}{L}\boldsymbol{g}_{k-1}\boldsymbol{g}_k^\intercal - \frac{1}{L}\boldsymbol{g}_k\boldsymbol{g}_{k-1}^\intercal\right).$$

We have a strong duality argument

$$\operatorname*{arg\,min}_{h_{i,j}}\ \operatorname*{maximize}_{\mathbf{G},\mathbf{F}_0,\mathbf{F}_1}\ \boldsymbol{f}_{N+1}^\intercal\mathbf{F}_1 = \operatorname*{arg\,min}_{h_{i,j}}\ \operatorname*{minimize}_{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau)\geq\mathbf{0}}\ \tau,$$

as ORC-F's optimality proof. Remind that (38) finds the "best" proof for the algorithm. Now we investigate the optimization step for algorithm. The last part is minimizing (38) with stepsize, i.e.

$$\operatorname*{minimize}_{h_{i,j}}\ \operatorname*{minimize}_{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau)\geq\mathbf{0}}\quad \tau \tag{39}$$
$$\text{subject to}\quad \mathbf{0} = -\sum_{k=0}^{N}\alpha_k\boldsymbol{f}_k + \sum_{k=1}^{N}\lambda_k\boldsymbol{f}_k + \sum_{k=0}^{N}\beta_k\boldsymbol{f}_k \tag{40}$$
$$\mathbf{0} = -\boldsymbol{f}_{N+1} - \sum_{k=1}^{N}\lambda_k\boldsymbol{f}_k + \sum_{k=0}^{N}\alpha_k\boldsymbol{f}_{k+1} \tag{41}$$
$$0 \preceq S(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau). \tag{42}$$

We note that $\boldsymbol{f}_i$ is a standard unit vector mentioned in (7) (not a variable), we can write (40) and (41) as

$$\begin{array}{c}\left(\begin{array}{ll}\beta_k = \alpha_k - \lambda_k = \lambda_{k+1} - \lambda_k, & k\in\{1,\dots,N-1\}\\ \beta_0 = \alpha_0 = \lambda_1 & \\ \beta_N = \alpha_N - \lambda_N = 1 - \lambda_N. & \end{array}\right)\\ \left(\begin{array}{ll}\alpha_N = 1 & \\ \alpha_k = \lambda_{k+1}, & k\in\{0,1,\dots,N-1\}\end{array}\right)\end{array} \tag{43}$$

We consider (42) with (43) and FSFO's $h_{i,j}$. To be specific, we substitute $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to $\boldsymbol{\lambda}$ in $S(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau)$. To show the dependency of S to $(h_{i,j})$ since $\boldsymbol{x}_k$ are represented with $(h_{i,j})$, we will explicitly write S as $S(\boldsymbol{\lambda}, \tau; (h_{i,j}))$. Then, we get

$$\begin{aligned}
S(\boldsymbol{\lambda}, \tau; (h_{i,j})) = \tau \boldsymbol{x}_0 \boldsymbol{x}_0^\intercal - \sum_{k=1}^{N} \frac{\lambda_k}{2L} \boldsymbol{g}_k \boldsymbol{g}_k^\intercal + \frac{1}{2L} \boldsymbol{g}_N \boldsymbol{g}_N^\intercal + \sum_{k=1}^{N} \frac{\lambda_k}{2L} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal \\
+ \sum_{k=1}^{N-1} \sum_{t=0}^{k-1} \left(\frac{\lambda_k}{2} \frac{h_{k,t}}{L} + \frac{\lambda_{k+1} - \lambda_k}{2} \sum_{j=t+1}^{k} \frac{h_{j,t}}{L} \right) \left(\boldsymbol{g}_k \boldsymbol{g}_t^\intercal + \boldsymbol{g}_t \boldsymbol{g}_k^\intercal \right) \\
+ \sum_{t=0}^{N-1} \left(\frac{\lambda_N}{2} \frac{h_{N,t}}{L} + \frac{1 - \lambda_N}{2} \sum_{j=t+1}^{N} \frac{h_{j,t}}{L} \right) \left(\boldsymbol{g}_N \boldsymbol{g}_t^\intercal + \boldsymbol{g}_t \boldsymbol{g}_N^\intercal \right) \\
- \sum_{k=1}^{N-1} \frac{\lambda_{k+1} - \lambda_k}{2} \left(\boldsymbol{x}_0 \boldsymbol{g}_k^\intercal + \boldsymbol{g}_k \boldsymbol{x}_0^\intercal \right) - \frac{\lambda_1}{2} \left(\boldsymbol{x}_0 \boldsymbol{g}_0^\intercal + \boldsymbol{g}_0 \boldsymbol{x}_0^\intercal \right) - \frac{1 - \lambda_N}{2} \left(\boldsymbol{x}_0 \boldsymbol{g}_N^\intercal + \boldsymbol{g}_N \boldsymbol{x}_0^\intercal \right).
\end{aligned}$$

Using the fact that $\boldsymbol{x}_0, \boldsymbol{g}_i, \boldsymbol{f}_i$ are unit vectors, we can represent $S(\boldsymbol{\lambda}, \tau; (h_{i,j}))$ with $\boldsymbol{\gamma}(\boldsymbol{\lambda}) = -L\boldsymbol{\beta} = -L(\lambda_1, \lambda_2 - \lambda_1, \dots, 1 - \lambda_N) = (\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}), \gamma_N(\boldsymbol{\lambda}))$ and $\tau' = 2L\tau$ as

$$S(\boldsymbol{\lambda}, \tau'; (h_{i,j})) = \frac{1}{L} \begin{pmatrix} \frac{1}{2}\tau' & \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})^\intercal & \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) \\ \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) & Q(\boldsymbol{\lambda}; (h_{i,j})) & q(\boldsymbol{\lambda}; (h_{i,j})) \\ \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) & q(\boldsymbol{\lambda}; (h_{i,j}))^\intercal & \frac{1}{2} \end{pmatrix} \succeq 0.$$

Here, Q and $\mathbf{q}$ are defined as

$$\begin{aligned}
Q(\boldsymbol{\lambda}; (h_{i,j})) = - \sum_{k=1}^{N-1} \frac{\lambda_k}{2} \boldsymbol{g}'_k \boldsymbol{g}_k^{'\intercal} + \sum_{k=1}^{N-1} \frac{\lambda_k}{2} (\boldsymbol{g}'_{k-1} - \boldsymbol{g}'_k)(\boldsymbol{g}'_{k-1} - \boldsymbol{g}'_k)^\intercal + \frac{\lambda_N}{2} \boldsymbol{g}'_{N-1} \boldsymbol{g}_{N-1}^{'\intercal} \\
+ \sum_{k=1}^{N-1} \sum_{t=0}^{k-1} \left(\frac{\lambda_k}{2} h_{k,t} + \frac{\lambda_{k+1} - \lambda_k}{2} \sum_{j=t+1}^{k} h_{j,t} \right) \left(\boldsymbol{g}'_k \boldsymbol{g}_t^{'\intercal} + \boldsymbol{g}'_t \boldsymbol{g}_k^{'\intercal} \right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{q}(\boldsymbol{\lambda}; (h_{i,j})) &= -\frac{\lambda_N}{2} \boldsymbol{g}'_{N-1} + \sum_{t=0}^{N-1} \left(\frac{\lambda_N}{2} h_{N,t} + \frac{1 - \lambda_N}{2} \sum_{j=t+1}^{N} h_{j,t} \right) \boldsymbol{g}'_t \\
&= \sum_{t=0}^{N-2} \left(\frac{\lambda_N}{2} h_{N,t} + \frac{1 - \lambda_N}{2} \sum_{j=t+1}^{N} h_{j,t} \right) \boldsymbol{g}'_t + \left(\frac{1}{2} h_{N,N-1} - \frac{\lambda_N}{2} \right) \boldsymbol{g}'_{N-1}.
\end{aligned}$$

where $\boldsymbol{g}'_k = e_{k+1} \in \mathbb{R}^{N+1}$. Note that (39) is equivalent to

$$\begin{aligned}
&\underset{h_{i,j}}{\text{minimize}} \ \underset{(\boldsymbol{\lambda}, \tau') \geq \mathbf{0}}{\text{minimize}} \quad \tau' \\
&\text{subject to} \quad \begin{pmatrix} \frac{1}{2}\tau' & \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})^\intercal & \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) \\ \frac{1}{2}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) & Q(\boldsymbol{\lambda}; (h_{i,j})) & q(\boldsymbol{\lambda}; (h_{i,j})) \\ \frac{1}{2}\gamma_N(\boldsymbol{\lambda}) & q(\boldsymbol{\lambda}; (h_{i,j}))^\intercal & \frac{1}{2} \end{pmatrix} \succeq 0
\end{aligned} \tag{44}$$

and dividing this optimized value with $2L$ gives the optimized value of (39). Using Schur complement (**?**), (44) can be converted to the problem as

$$\underset{h_{i,j}}{\text{minimize}}\ \underset{(\boldsymbol{\lambda},\tau')\geq \mathbf{0}}{\text{minimize}} \quad \tau' \tag{45}$$

$$\text{subject to} \quad \begin{pmatrix} Q-2\mathbf{q}\mathbf{q}^\intercal & \frac{1}{2}(\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})-2\mathbf{q}\gamma_N(\boldsymbol{\lambda})) \\ \frac{1}{2}(\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})-2\mathbf{q}\gamma_N(\boldsymbol{\lambda}))^\intercal & \frac{1}{2}(\tau'-\gamma_N(\boldsymbol{\lambda})^2) \end{pmatrix} \succeq 0. \tag{46}$$

So far we simplified SDP. We will have three steps: finding variables that make (46)'s left hand side zero, showing that the solution from the first step satisfies KKT condition, and finally showing that obtained algorithm is equivalent to FGM.

Claim 8 *There is a point that makes* (46)*'s left-hand side zero.*

Proof Defining $\{r_{k,t}\}_{k=1,2,\ldots,N,t=0,\ldots,k-1}$ as

$$r_{k,t} = \lambda_k h_{k,t} - \frac{\gamma_k}{L}\sum_{j=t+1}^{k} h_{j,t}.$$

Then, if $r_{i,j}$ is determined, (Drori and Teboulle, 2014, Theorem 5.1) indicates this uniquely determine $h_{i,j}$. We set $(\lambda_k)_{k=0}^N$ and $(r_{N,k})_{k=0}^{N-1}$ as

$$\begin{aligned} \lambda_k &= \frac{\theta_{k-1}^2}{\theta_N^2}, && k\in\{1,2,\ldots,N\} \\ r_{N,k} &= \frac{\theta_k}{\theta_N}, && k\in\{0,1,\ldots,N-2\} \\ r_{N,N-1}-\lambda_N &= \frac{\theta_{N-1}}{\theta_N}. \end{aligned} \tag{47}$$

Moreover, we set

$$\begin{aligned} r_{k,t} &= \frac{\theta_{k-1}\theta_{t-1}}{\theta_N^2}, && k\in\{1,2,\ldots,N-1\}, \quad t\in\{0,1,\ldots,k-2\} \\ r_{k,k-1} &= \frac{\theta_{k-1}\theta_{k-2}}{\theta_N^2}+\frac{\theta_{k-1}^2}{\theta_N^2}, && k\in\{1,2\ldots,N-1\}. \end{aligned} \tag{48}$$

In addition, we set $\hat{\boldsymbol{\gamma}}$ as

$$\begin{aligned} \gamma_t &= \gamma_N r_{N,t}, && t\in\{0,1,\ldots,N-2\} \\ \gamma_{N-1} &= \gamma_N(r_{N,N-1}-\lambda_N) \\ \gamma_N &= L(1-\lambda_N). \end{aligned}$$

Lastly, we set τ' as

$$\tau' = \frac{L^2}{\theta_N^2}, \tag{49}$$

and $\tau = \frac{L}{2\theta_N^2}$. These variables make (46)'s left-hand side zero. ■

Claim 9 (47), (48) *and* (49) *are an optimal solution of* (45).

Proof Let we represent S with the variable $(r_{i,j})$. We will denote this as $\mathbf{A}$. To be specific,

$$\mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))=S(\boldsymbol{\lambda},\boldsymbol{\gamma},\tau';(h_{i,j}))=\begin{pmatrix}\frac{1}{2}\tau' & -\frac{L}{2}\hat{\boldsymbol{\beta}}^\intercal & -\frac{L}{2}\beta_N\\ -\frac{L}{2}\hat{\boldsymbol{\beta}}^\intercal & Q(\boldsymbol{\lambda};(r_{i,j})) & \mathbf{q}((r_{i,j}))\\ -\frac{L}{2}\beta_N & \mathbf{q}((r_{i,j}))^\intercal & \frac{1}{2}\end{pmatrix}\succeq 0.$$

Here, $\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}^\intercal,\beta_N)^\intercal$,

$$Q(\boldsymbol{\lambda};(r_{i,j}))-\sum_{k=1}^{N-1}\frac{\lambda_k}{2}\boldsymbol{g}_k\boldsymbol{g}_k^\intercal+\sum_{k=1}^{N-1}\frac{\lambda_k}{2}(\boldsymbol{g}_{k-1}-\boldsymbol{g}_k)(\boldsymbol{g}_{k-1}-\boldsymbol{g}_k)^\intercal+\frac{\lambda_N}{2}\boldsymbol{g}_{N-1}\boldsymbol{g}_{N-1}^\intercal+\sum_{k=1}^{N-1}\sum_{t=0}^{k-1}\frac{r_{k,t}}{2}\left(\boldsymbol{g}_k\boldsymbol{g}_t^\intercal+\boldsymbol{g}_t\boldsymbol{g}_k^\intercal\right)$$

and

$$\mathbf{q}((r_{i,j}))=\sum_{t=0}^{N-1}\frac{r_{N,t}}{2}\boldsymbol{g}_t-\frac{\lambda_N}{2}\boldsymbol{g}_{N-1}.$$

Define a linear SDP relaxation of (44) as

$$\begin{aligned}\underset{r_{i,j}}{\text{minimize}}\ \underset{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau')\geq\mathbf{0}}{\text{minimize}}\quad & \tau'\\ \text{subject to}\quad & \mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))\succeq 0.\\ & \mathbf{B}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau')=\big(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau'\big)\geq 0\\ & \mathbf{C}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha})=(-\alpha_0+\beta_0,-\alpha_1+\lambda_1+\beta_1,\dots,-\alpha_N+\lambda_N+\beta_N)=0\\ & \mathbf{D}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha})=(-\lambda_1+\alpha_0,-\lambda_2+\alpha_1,\dots,-\lambda_N+\alpha_{N-1},\alpha_N-1)=0.\end{aligned}\tag{50}$$

(Drori and Teboulle, 2014, Theorem 3) indicates that if we prove the choice in the previous claim satisfies KKT condition of (50), then this is also an optimal solution for the original problem. The Lagrangian of the minimization problem is

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}),\mathbf{K},\mathbf{b},\mathbf{c},\mathbf{d})\\ &\quad=\frac{1}{2}\tau'-\operatorname{tr}\left\{\mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))\mathbf{K}\right\}-\mathbf{b}^\intercal\mathbf{B}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau')-\mathbf{c}^\intercal\mathbf{C}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha})-\mathbf{d}^\intercal\mathbf{D}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha})\end{aligned}$$

and the KKT conditions of the minimization problems are

$$\begin{aligned}&\mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau';(r_{i,j}))\succeq 0,\mathbf{B}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau')\geq 0,\mathbf{C}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha})=0,\mathbf{D}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha})=0,\\ &\nabla_{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}),\mathbf{K},\mathbf{b},\mathbf{c},\mathbf{d})=0,\\ &\mathbf{K}\succeq 0,\mathbf{b}\geq 0,\\ &\operatorname{tr}\left\{\mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))\mathbf{K}\right\}=0,\mathbf{b}^\intercal\mathbf{B}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau')=0,\end{aligned}$$

where $\mathbf{K}$ is a symmetric matrix. Here, $\mathbf{b}=(\mathbf{u},\mathbf{v},\mathbf{w},s)$. We re-index K's column and row starting from -1 (so K's rows and columns index are $\{-1,0,1.\dots,N\}$). Now, we will show that there exist a dual optimal solution $(\mathbf{K},\mathbf{b},\mathbf{c},\mathbf{d})$ that $(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}),\mathbf{K},\mathbf{b},\mathbf{c},\mathbf{d})$ satisfies KKT condition, which proves a pair $(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))$ is an optimal solution for primal problem. The stationary condition $\nabla_{(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}),\mathbf{K},\mathbf{b},\mathbf{c},\mathbf{d})=0$ can be rewritten as

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\lambda_k} &= -\frac{1}{2}\left(K_{k-1,k-1}-K_{k-1,k}-K_{k,k-1}\right)-u_k-c_k+d_{k-1}=0, \qquad k\in\{1,2,\ldots,N\}\\
\frac{\partial\mathcal{L}}{\partial\beta_k} &= \frac{L}{2}\left(K_{-1,k}+K_{k,-1}\right)-v_k-c_k=0, \qquad k\in\{0,1,\ldots,N\}\\
\frac{\partial\mathcal{L}}{\partial\alpha_k} &= -w_k+c_k-d_k=0, \qquad k\in\{0,1,\ldots,N\}\\
\frac{\partial\mathcal{L}}{\partial\tau'} &= \frac{1}{2}-\frac{1}{2}K_{-1,-1}-s=0\\
\frac{\partial\mathcal{L}}{\partial r_{k,t}} &= -\frac{1}{2}(K_{k,t}+K_{t,k})=0, \qquad k\in\{1,2,\ldots,N\},\quad t\in\{0,1,\ldots,k-1\}.
\end{aligned} \tag{51}$$

We already know that $\mathbf{B}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau')\neq 0$, we can set $\mathbf{b}=0$. Then, (51) reduces to

$$\begin{aligned}
&K_{k,t}=0, \qquad k\in\{1,2,\ldots,N\},\quad t\in\{0,1,\ldots,k-1\}\\
&-\frac{1}{2}K_{k-1,k-1}-c_k+d_{k-1}=0, \qquad k\in\{1,2,\ldots,N\}\\
&LK_{-1,k}-c_k=0, \qquad k\in\{0,1,\ldots,N\}\\
&c_k-d_k=0, \qquad k\in\{0,1,\ldots,N\}\\
&K_{-1,-1}=1.
\end{aligned}$$

Then, we have

$$\mathbf{K}=\begin{pmatrix}
1 & \frac{c_0}{L} & \frac{c_1}{L} & \ldots & \frac{c_{N-1}}{L} & \frac{c_N}{L}\\
\frac{c_0}{L} & 2c_0-2c_1 & 0 & \ldots & 0 & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
\frac{c_{N-1}}{L} & 0 & 0 & \ldots & 2c_{N-1}-2c_N & 0\\
\frac{c_N}{L} & 0 & 0 & \ldots & 0 & K_{N,N}
\end{pmatrix}\succeq 0$$

and since $\mathrm{tr}\left\{\mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))\mathbf{K}\right\}=0$ with $\mathbf{A}\succeq 0$, we can replace this condition by $\mathbf{A}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))\mathbf{K}=0$. Then the KKT condition for the given $(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha},\tau',(r_{i,j}))$ reduces to

$$\begin{aligned}
&\frac{1}{2}\tau'-\frac{1}{2}\boldsymbol{\beta}^\intercal\mathbf{c}=0\\
&\frac{1}{2L}\tau'\mathbf{c}-\frac{L}{2}\,\mathrm{diag}(2c_0-2c_1,\ldots,2c_{N-1}-2c_N,K_{N,N})\boldsymbol{\beta}=0\\
&-\boldsymbol{\beta}\mathbf{c}^\intercal+\begin{pmatrix} Q & \mathbf{q}\\ \mathbf{q}^\intercal & \frac{1}{2}\end{pmatrix}\mathrm{diag}(2c_0-2c_1,\ldots,2c_{N-1}-2c_N,K_{N,N})=0
\end{aligned}$$

and this is equivalent to

$$\begin{aligned}
&\sum_{i=0}^{N}\theta_i c_i=L^2\\
&c_i=(2c_i-2c_{i+1})\theta_i \qquad \text{for } i=0,1,\ldots,N-1\\
&c_N=K_{N,N}\theta_N.
\end{aligned}$$

This clearly indicates that $K_{N,N} > 0$ and c_i are all determined as positive, and

$$\mathbf{K} = \begin{pmatrix} 1 & \frac{c_0}{L} & \frac{c_1}{L} & \dots & \frac{c_{N-1}}{L} & \frac{c_N}{L} \\ \frac{c_0}{L} & \frac{c_0}{\theta_0} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{c_{N-1}}{L} & 0 & 0 & \dots & \frac{c_{N-1}}{\theta_{N-1}} & 0 \\ \frac{c_N}{L} & 0 & 0 & \dots & 0 & \frac{c_N}{\theta_N} \end{pmatrix} \succeq 0$$

since $\sum_{i=0}^{N} \theta_i c_i = L^2$. ■

Claim 10 *The obtained algorithm is FGM.*

Proof By calculating $(h_{i,j})$ of FGM, we can prove the equivalence of the obtained solution and FGM. Indeed, FGM is obtained by using (Lee et al., 2021)'s auxiliary sequences. We will show that obtained $(\hat{h}_{i,j})$ satisfies

$$x_0 - \sum_{i=1}^{k+1} \sum_{j=0}^{i-1} \frac{\hat{h}_{i.j}}{L} \nabla f(x_j) = \left(1 - \frac{1}{\theta_{k+1}}\right) \left(x_0 - \sum_{i=1}^{k} \sum_{j=0}^{i-1} \frac{\hat{h}_{i,j}}{L} \nabla f(x_j) - \frac{1}{L} \nabla f(x_k)\right) + \frac{1}{\theta_{k+1}} \left(x_0 - \sum_{j=0}^{k} \frac{\theta_j}{L} \nabla f(x_j)\right),$$

which is re-written form of FGM. Comparing $\nabla f(x_j)$'s each coefficient, we should prove

$$\sum_{i=j+1}^{k+1} \hat{h}_{i,j} = \left(1 - \frac{1}{\theta_{k+1}}\right) \sum_{i=j+1}^{k} \hat{h}_{i,j} + \frac{1}{\theta_{k+1}} \theta_j \qquad j \in \{0, 1, \dots, k-1\}$$

$$\hat{h}_{k+1,k} = \left(1 - \frac{1}{\theta_{k+1}}\right) + \frac{1}{\theta_{k+1}} \theta_k,$$

which is exactly equal to the recursive rule of (Drori and Teboulle, 2014, Theorem 3). ■

In sum, the algorithm's performance criterion $f(y_{N+1}) - f_\star$ is bounded as

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

Overall, we showed that FGM is the "best" algorithm under $\mathcal{I}_{\text{FGM}}$.

$$\text{FGM} = \mathcal{A}_N^\star(f(y_{N+1}) - f_\star, \|x_0 - x_\star\| \leq R, \mathcal{I}_{\text{FGM}})$$

and

$$\begin{aligned} &\mathcal{R}(\text{FGM}, f(y_{N+1}) - f_\star, \|x_0 - x_\star\| \leq R, \mathcal{I}_{\text{FGM}}) \\ &\qquad = \mathcal{R}^*(f(y_{N+1}) - f_\star, \|x_0 - x_\star\| \leq R, \mathcal{I}_{\text{FGM}}) \\ &\qquad = \frac{LR^2}{2\theta_N^2} \end{aligned}$$

hold.

B.3 Conjecture of $\mathcal{A}^\star$-optimality of OBL-G$_\flat$

We give a conjecture for $\mathcal{A}^\star$-optimality of OBL-G$_\flat$. Set $\|\nabla f(x_N)\|^2$ to be the performance measure and $f(x_0) - f_\star \le \frac{1}{2}LR^2$ to be the initial condition. We use the set of inequalities that are handy for backtracking linesearches. Since the constraints and the objective of the problem are homogenous, we assume $R = 1$ without loss of generality. For the argument of homogeneous, we refer to(Drori and Teboulle, 2014; Kim and Fessler, 2016; Taylor et al., 2017b). We use the set of inequalities that are handy for backtracking linesearches:

$$
\begin{aligned}
\mathcal{I}_{\text{OBL-G}_\flat} = &\left\{ f_{k-1,0} \ge f_{k,0} + \langle g_k, x_{k-1} - x_k\rangle + \frac{1}{2L}\|g_{k-1,0} - g_k\|^2 \right\}_{k=1}^{N} \\
&\bigcup \left\{ f_{N,0} \ge f_{k,0} + \langle g_k, x_N - x_k\rangle \right\}_{k=0}^{N-1} \bigcup \left\{ f_{N,0} \ge f_\star + \frac{1}{2}\|g_{N,0}\|^2 \right\}.
\end{aligned}
$$

For calculating $\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{OBL-G}_\flat})$ with fixed $\mathcal{A}_N$, define the PEP with $\mathcal{I}_{\text{OBL-G}_\flat}$ as

$$
\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{OBL-G}_\flat}) = \left(\begin{array}{lll}
\text{maximize} & \|g_N\|^2 & \\
\text{subject to} & 0 & \ge f_{0,0} - \frac{1}{2}L \\
 & f_{k-1,0} & \ge f_{k,0} + \langle g_k, x_{k-1} - x_k\rangle + \frac{1}{2L}\|g_{k-1} - g_k\|^2, k \in \{1, 2, \dots, N\} \\
 & f_{N,0} & \ge f_{k,0} + \langle g_k, x_N - x_k\rangle, k \in \{0, 1, \dots, N-1\} \\
 & f_{N,0} & \ge f_\star + \frac{1}{2L}\|g_N\|^2 \\
 & x_k & \text{is following the algorithm } \mathcal{A}_N.
\end{array} \right)
$$

Using the notation of Section 4.2, we reformulate the problem of computing the risk $\mathcal{R}(\mathcal{A}_N, \mathcal{P}, \mathcal{C}, \mathcal{I}_{\text{OBL-G}_\flat})$ as the following SDP:

$$
\begin{aligned}
\underset{\mathbf{G}, \mathbf{F}_0}{\text{maximize}} \quad & \boldsymbol{g}_N^\intercal \mathbf{G} \boldsymbol{g}_N \\
\text{subject to} \quad & 0 \ge \boldsymbol{f}_0^\intercal F_0 - \frac{1}{2}L \\
& 0 \ge (\boldsymbol{f}_k - \boldsymbol{f}_{k-1})^\intercal \mathbf{F}_0 + \boldsymbol{g}_k^\intercal \mathbf{G}(\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) + \frac{1}{2L}(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal \mathbf{G}(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k), \qquad k \in \{1, 2, \dots, N\} \\
& 0 \ge (\boldsymbol{f}_k^\intercal - \boldsymbol{f}_N)^\intercal \mathbf{F}_0 + \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{x}_N - \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{x}_k, \qquad k \in \{0, 1, \dots, N-1\} \\
& 0 \ge -\boldsymbol{f}_N^\intercal \mathbf{F}_0 + \frac{1}{2L}\boldsymbol{g}_N^\intercal \mathbf{G} \boldsymbol{g}_N \\
& \mathbf{G} \succeq 0, \mathbf{F}_0 \ge 0.
\end{aligned}
$$

For above transformation, $d \geq N+2$ is used (Taylor et al., 2017b). The Lagrangian of the optimization problem becomes

$$
\begin{aligned}
&\Lambda(\mathbf{F}_0, \mathbf{G}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \tau, c) \\
&\quad = -\boldsymbol{g}_N^\intercal \mathbf{G} \boldsymbol{g}_N + \tau \left(\boldsymbol{f}_0^\intercal \mathbf{F}_0 - \frac{1}{2} L \right) + c \left(-\boldsymbol{f}_N^\intercal \mathbf{F}_0 + \frac{1}{2L} \boldsymbol{g}_N^\intercal \mathbf{G} \boldsymbol{g}_N \right) \\
&\quad + \sum_{k=1}^{N} \lambda_k \left((\boldsymbol{f}_k - \boldsymbol{f}_{k-1})^\intercal \mathbf{F}_0 + \boldsymbol{g}_k^\intercal \mathbf{G} (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) + \frac{1}{2L} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal \mathbf{G} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k) \right) \\
&\quad + \sum_{k=0}^{N-1} \beta_k \left((\boldsymbol{f}_k - \boldsymbol{f}_N)^\intercal \mathbf{F}_0 + \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{x}_N - \boldsymbol{g}_k^\intercal \mathbf{G} \boldsymbol{x}_k \right).
\end{aligned}
$$

with dual variables $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_N) \in \mathbb{R}_+^N$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{N-1}) \in \mathbb{R}_+^N$, and $\tau, c \geq 0$.

Then the dual formulation of PEP problem is

$$
\begin{aligned}
\underset{(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau, c) \geq \mathbf{0}}{\text{maximize}} \quad & -\frac{1}{2} L \tau \\
\text{subject to} \quad & \mathbf{0} = \tau \boldsymbol{f}_0 - c \boldsymbol{f}_N + \sum_{k=1}^{N} \lambda_k (\boldsymbol{f}_k - \boldsymbol{f}_{k-1}) + \sum_{k=0}^{N-1} \beta_k (\boldsymbol{f}_k - \boldsymbol{f}_N) \\
& 0 \preceq S(\boldsymbol{\lambda}, \boldsymbol{\beta}, c),
\end{aligned}
\tag{52}
$$

where S is defined as

$$
\begin{aligned}
S(\boldsymbol{\lambda}, \boldsymbol{\beta}, c) = & -\boldsymbol{g}_N \boldsymbol{g}_N^\intercal + \frac{c}{2L} \boldsymbol{g}_N \boldsymbol{g}_N^\intercal + \sum_{k=1}^{N} \frac{\lambda_k}{2} \left(\boldsymbol{g}_k (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k)^\intercal + (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) \boldsymbol{g}_k^\intercal + \frac{1}{L} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal \right) \\
& + \sum_{k=0}^{N} \frac{\beta_k}{2} \left(\boldsymbol{g}_k \boldsymbol{x}_N^\intercal + \boldsymbol{x}_N \boldsymbol{g}_k^\intercal - \boldsymbol{g}_k \boldsymbol{x}_k^\intercal - \boldsymbol{x}_k \boldsymbol{g}_k^\intercal \right).
\end{aligned}
$$

We have a strong duality argument

$$
\underset{h_{i,j}}{\arg\min} \underset{\mathbf{G}, \mathbf{F}_0}{\text{maximize}} \;\; \boldsymbol{g}_N^\intercal \mathbf{G} \boldsymbol{g}_N = \underset{h_{i,j}}{\arg\min} \underset{(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau, c) \geq \mathbf{0}}{\text{minimize}} \;\; \frac{1}{2} L \tau,
$$

as ORC-F's optimality proof. Remind that (52) finds the "best" proof for the algorithm. Now we investigate the optimization step for algorithm. The last part is minimizing (52) with stepsize, i.e.

$$
\underset{h_{i,j}}{\text{minimize}} \underset{(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau, c) \geq \mathbf{0}}{\text{minimize}} \quad \tau \tag{53}
$$

$$
\text{subject to} \quad \mathbf{0} = \tau \boldsymbol{f}_0 - c \boldsymbol{f}_N + \sum_{k=1}^{N} \lambda_k (\boldsymbol{f}_k - \boldsymbol{f}_{k-1}) + \sum_{k=0}^{N-1} \beta_k (\boldsymbol{f}_k - \boldsymbol{f}_N) \tag{54}
$$

$$
0 \preceq S(\boldsymbol{\lambda}, \boldsymbol{\beta}, c). \tag{55}
$$

We note that $\boldsymbol{f}_i$ is a standard unit vector mentioned in (7) (not a variable), we can write (54) as

$$\begin{pmatrix} \tau - \lambda_1 + \beta_0 = 0 & \\ \lambda_k - \lambda_{k+1} + \beta_k = 0, & k \in \{1, \dots, N-1\} \\ -c + \lambda_N - \sum_{k=0}^{N-1} \beta_k = 0. & \end{pmatrix} \tag{56}$$

We consider (55) with (56) and FSFO's $h_{i,j}$. To be specific, we represent the dependency of S to $(h_{i,j})$ since $\boldsymbol{x}_k$ are represented with $(h_{i,j})$. Then, we get

$$\begin{aligned} S(\boldsymbol{\lambda}, \boldsymbol{\beta}, c; (h_{i,j})) = & -\boldsymbol{g}_N \boldsymbol{g}_N^\intercal + \frac{c}{2L} \boldsymbol{g}_N \boldsymbol{g}_N^\intercal \\ & + \sum_{k=1}^{N} \frac{\lambda_k}{2} \left(\boldsymbol{g}_k \left(\sum_{t=0}^{k-1} \frac{h_{k,t}}{L} \boldsymbol{g}_t \right)^\intercal + \left(\sum_{t=0}^{k-1} \frac{h_{k,t}}{L} \boldsymbol{g}_t \right) \boldsymbol{g}_k^\intercal + \frac{1}{L} (\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)(\boldsymbol{g}_{k-1} - \boldsymbol{g}_k)^\intercal \right) \\ & + \sum_{k=0}^{N-1} \frac{\beta_k}{2} \left(- \left(\sum_{j=k+1}^{N} \sum_{t=0}^{j-1} \frac{h_{j,t}}{L} \boldsymbol{g}_t \right) \boldsymbol{g}_k^\intercal - \boldsymbol{g}_k \left(\sum_{j=k+1}^{N} \sum_{t=0}^{j-1} \frac{h_{j,t}}{L} \boldsymbol{g}_t \right)^\intercal \right). \end{aligned} \tag{57}$$

Claim 11 *There is a point that makes* (57)*'s right-hand side as zero.*

Proof By calculating $2LS(\boldsymbol{\lambda}, \boldsymbol{\beta}, c; (h_{i,j}))$'s $\boldsymbol{g}_i \boldsymbol{g}_j^\intercal$ coefficients, we have

$$\begin{aligned} & \lambda_1 - 2\beta_0 \sum_{l=1}^{N} h_{l,0} \\ & \lambda_i + \lambda_{i+1} - 2\beta_i \sum_{l=i+1}^{N} h_{l,i}, \qquad i \in \{1, 2, \dots, N-1\} \\ & \lambda_N + c - 2L \\ & \lambda_i (h_{i,i-1} - 1) - \beta_i \sum_{l=i+1}^{N} h_{l,i-1} - \beta_{i-1} \sum_{l=i+1}^{N} h_{l,i}, \qquad i \in \{1, 2, \dots, N-1\} \\ & \lambda_N (h_{N,N-1} - 1) \\ & \lambda_i h_{i,j} - \beta_i \sum_{l=i+1}^{N} h_{l,j} - \beta_j \sum_{l=i+1}^{N} h_{l,i}, i \in \{2, 3, \dots, N-1\}, \qquad j \in \{0, 1, \dots, i-2\} \\ & \lambda_N h_{N,j}, \qquad j \in \{0, 1, \dots, N-2\}. \end{aligned} \tag{58}$$

For finding a solution of $S = 0$, we will set λ_i as

$$\lambda_{N-k+1} = \frac{1}{k(k+1)} \lambda.$$

(56) can be written as

$$\begin{aligned}
&\tau - \frac{1}{N(N+1)}\lambda + \beta_0 = 0 \\
&\frac{1}{(N-k+1)(N-k+2)}\lambda - \frac{1}{(N-k)(N-k+1)}\lambda + \beta_k = 0, \qquad k \in \{1, \dots, N-1\}, \\
&-c + \frac{1}{2}\lambda - \sum_{k=0}^{N-1} \beta_k = 0.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\beta_k &= \frac{2}{(N-k)(N-k+1)(N-k+2)}\lambda, \qquad k \in \{1, 2, \dots, N-1\} \\
\beta_0 &= \frac{1}{N(N+1)}\lambda - \tau = \hat{\beta}_0 \lambda \\
c &= \tau.
\end{aligned} \tag{59}$$

Then, with (58) and (59), we get

$$
\begin{aligned}
0 &= \frac{1}{N(N+1)}\lambda - 2\hat{\beta}_0\lambda\sum_{l=1}^{N} h_{l,0} \\
0 &= \frac{2}{(N-i)(N-i+2)}\lambda - \frac{4}{(N-i)(N-i+1)(N-i+2)}\lambda\sum_{l=i+1}^{N} h_{l,i}, \qquad i\in\{1,2,\dots,N-1\} \\
0 &= \frac{1}{2}\lambda + \tau - 2L \\
0 &= \frac{1}{(N-i+1)(N-i+2)}\lambda(h_{i,i-1}-1) - \frac{2}{(N-i)(N-i+1)(N-i+2)}\lambda\sum_{l=i+1}^{N} h_{l,i-1} \\
&\qquad - \frac{2}{(N-i+1)(N-i+2)(N-i+3)}\lambda\sum_{l=i+1}^{N} h_{l,i}, \qquad i\in\{2,\dots,N-1\} \\
0 &= \frac{1}{N(N+1)}\lambda(h_{1,0}-1) - \frac{2}{(N-1)N(N+1)}\lambda\sum_{l=2}^{N} h_{l,0} - \hat{\beta}_0\lambda\sum_{l=2}^{N} h_{l,1} \\
0 &= \frac{1}{2}\lambda(h_{N,N-1}-1) \\
0 &= \frac{1}{(N-i+1)(N-i+2)}\lambda h_{i,j} - \frac{2}{(N-i)(N-i+1)(N-i+2)}\lambda\sum_{l=i+1}^{N} h_{l,j} \\
&\qquad - \frac{2}{(N-j)(N-j+1)(N-j+2)}\lambda\sum_{l=i+1}^{N} h_{l,i}, \qquad i\in\{2,3,\dots,N-1\}, \qquad j\in\{1,\dots,i-2\} \\
0 &= \frac{1}{(N-i+1)(N-i+2)}\lambda h_{i,0} - \frac{2}{(N-i)(N-i+1)(N-i+2)}\lambda\sum_{l=i+1}^{N} h_{l,0} \\
&\qquad - \hat{\beta}_0\lambda\sum_{l=i+1}^{N} h_{l,i}, \qquad i\in\{2,3,\dots,N-1\} \\
0 &= \frac{1}{2}\lambda h_{N,j}.
\end{aligned} \tag{60}
$$

Last equation indicates $h_{N,j}=0$ for all $j=0,1,\dots,N-2$, and by fifth equation, $h_{N,N-1}=1$ also holds. By first and second equation, we have

$$
\begin{aligned}
\sum_{l=1}^{N} h_{l,0} &= \frac{1}{2\hat{\beta}_0 N(N+1)} \\
\sum_{l=i+1}^{N} h_{l,i} &= \frac{N-i+1}{2}, \qquad i\in\{1,2,\dots,N-1\}.
\end{aligned} \tag{61}
$$

Since, (60)'s forth equation is equivalent to

$$\begin{aligned} 0 &= (h_{i,i-1} - 1) - \frac{2}{N-i} \sum_{l=i+1}^{N} h_{l,i-1} - \frac{2}{N-i+3} \sum_{l=i+1}^{N} h_{l,i}, \qquad i \in \{2, \dots, N-1\} \\ 0 &= \frac{1}{N(N+1)} \lambda (h_{1,0} - 1) - \frac{2}{(N-1)N(N+1)} \lambda \sum_{l=2}^{N} h_{l,0} - \hat{\beta}_0 \lambda \sum_{l=2}^{N} h_{l,1}. \end{aligned} \tag{62}$$

Combining with (61) and , we have

$$0 = (h_{i,i-1} - 1) - \frac{2}{N-i} \left(\frac{N-i+2}{2} - h_{i,i-1} \right) - \frac{2}{N-i+3} \frac{N-i+1}{2}$$

for $i = 2, 3, \dots, N-1$ and

$$0 = (h_{1,0} - 1) - \frac{2}{N-1} \left(\frac{1}{2\hat{\beta}_0 N(N+1)} - h_{1,0} \right) - \hat{\beta}_0 N(N+1) \frac{N}{2}.$$

Therefore, we have

$$\begin{aligned} h_{1,0} &= \frac{N-1}{N+1} + \frac{1}{N(N+1)^2 \hat{\beta}_0} + \frac{N^2(N-1)}{2} \hat{\beta}_0 \\ h_{i,i-1} &= \frac{3(N-i+1)}{N-i+3}. \end{aligned} \tag{63}$$

With (63), (60)'s seventh equation is equivalent to

$$0 = \frac{1}{(N-i+1)(N-i+2)} h_{i,j} - \frac{2}{(N-i)(N-i+1)(N-i+2)} \sum_{l=i+1}^{N} h_{l,j} - \frac{N-i+1}{(N-j)(N-j+1)(N-j+2)} \tag{64}$$

for $i = 2, 3, \dots, N-1$ and $j = 1, \dots, i-2$. For $i = 1, 2, \dots N-2$, and $j = 1, \dots, i-1$, we have

$$0 = \frac{1}{(N-i)(N-i+1)} h_{i+1,j} - \frac{2}{(N-i-1)(N-i)(N-i+1)} \sum_{l=i+2}^{N} h_{l,j} - \frac{N-i}{(N-j)(N-j+1)(N-j+2)}. \tag{65}$$

With $(N-i+2)(64) - (N-i-1)(65)$, we have

$$0 = \frac{1}{N-i+1} h_{i,j} - \frac{1}{N-i} h_{i+1,j} - \frac{4(N-i)+2}{(N-j)(N-j+1)(N-j+2)} \tag{66}$$

for $i = 2, 3, \dots, N-2$ and $j = 1, \dots, i-2$, and putting $i = N-1$ in (64),

$$h_{N-1,j} = \frac{12}{(N-j)(N-j+1)(N-j+2)} \tag{67}$$

for $j = 1, \dots, N-2$. (60)'s eighth equation is equivalent to

$$\hat{\beta}_0 \frac{N-i+1}{2} = \frac{1}{(N-i+1)(N-i+2)} h_{i,0} - \frac{2}{(N-i)(N-i+1)(N-i+2)} \sum_{l=i+1}^{N} h_{l,0} \tag{68}$$

for $i = 2, 3, \dots, N-1$. For $i = 1, 2, \dots, N-2$, we have

$$\hat{\beta}_0 \frac{N-i}{2} = \frac{1}{(N-i)(N-i+1)} h_{i+1,0} - \frac{2}{(N-i-1)(N-i)(N-i+1)} \sum_{l=i+2}^{N} h_{l,0} \tag{69}$$

With $(N-i+2)(68) - (N-i-1)(69)$, we have

$$\hat{\beta}_0 (2(N-i)+1) = \frac{1}{N-i+1} h_{i,0} - \frac{1}{N-i} h_{i+1,0}$$

for $i = 2, 3, \dots, N-2$, which indicates

$$h_{i,0} = \left(\frac{1}{N-1} h_{2,0} - \hat{\beta}_0 (i-2)(2N-i) \right) (N-i+1) \tag{70}$$

for $i = 3, \dots, N-1$. Putting $i = 2$ in (68) and using (70), we have

$$\begin{aligned} h_{2,0} &= \frac{2}{N-2} \sum_{l=3}^{N} h_{l,0} + \hat{\beta}_0 \frac{N(N-1)^2}{2} \\ &= \frac{(N-3)N}{(N-2)(N-1)} h_{2,0} - \hat{\beta}_0 \frac{(N-3)N(N+1)}{2} + \hat{\beta}_0 \frac{N(N-1)^2}{2} \end{aligned}$$

which indicates $h_{2,0} = N(N-1)(N-2)\hat{\beta}_0$. With (70),

$$h_{i,0} = \left(N(N-2) - (i-2)(2N-i) \right) (N-i+1) \hat{\beta}_0$$

for $i = 2, \dots, N-1$. Moreover, by (61), we have

$$\begin{aligned} \sum_{i=1}^{N} h_{i,0} &= h_{1,0} + \sum_{i=2}^{N} h_{i,0} \\ &= \frac{N-1}{N+1} + \frac{1}{N(N+1)^2 \hat{\beta}_0} + \frac{N^2(N-1)}{2} \hat{\beta}_0 + \frac{(N-2)(N-1)N(N+1)}{4} \hat{\beta}_0 \\ &= \frac{1}{2\hat{\beta}_0 N(N+1)} \end{aligned}$$

which indicates $\hat{\beta}_0 = \frac{2\left(\sqrt{\frac{N(N+1)}{2}} - 1\right)}{(N-1)N(N+1)(N+2)}$, and all $h_{i,0}$ is determined. Using (66) and (67), we can also derive

$$h_{i,j} = \frac{2(N-i)(N-i+1)(N-i+2)}{(N-j)(N-j+1)(N-j+2)}$$

for $i = 2, \dots, N$ and $j = 1, \dots, i-2$. In sum,

$$H = \begin{pmatrix} \frac{N+\sqrt{2N(N+1)}}{N+2} & & & & & \\ (N-2)(N-1)N\hat{\beta}_0 & \frac{3(N-1)}{N+1} & & & & \\ (N-3)(N-2)(N-1)\hat{\beta}_0 & \frac{2(N-3)(N-2)(N-1)}{(N-1)(N)(N+1)} & \frac{3(N-2)}{N} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ 1*2*3\hat{\beta}_0 & \frac{2*1*2*3}{(N-1)(N)(N+1)} & \frac{2*1*2*3}{(N-2)(N-1)(N)} & \dots & \frac{3*2}{4} & \\ 0 & 0 & 0 & \dots & 0 & \frac{3*1}{3} \end{pmatrix}$$

satisfies $S = 0$ with $\tau, \hat{\beta}_0, \lambda$ obtained above. ■

Since there cannot find the relationship similar to (Drori and Teboulle, 2014, Theorem 3), we cannot find the optimality of OBL-G$_\flat$. However, (53) is bi-convex over $(h_{i,j})$ and $(\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau, c)$, so far each given N we numerically solved and suspect that obtained solution in the previous claim is the solution for (53).

Claim 12 *The obtained algorithm has the same $(h_{i,j})$ with OBL-G$_\flat$.*

Proof By calculating $(h_{i,j})$ of OBL-G$_\flat$, we can prove the equivalence of the obtained solution and OBL-G$_\flat$. Indeed, OBL-G$_\flat$ is obtained by using (Lee et al., 2021)'s auxiliary sequences. We will show that obtained $(\hat{h}_{i,j})$ satisfies

$$\begin{aligned} x_0 - \sum_{i=1}^{k+1}\sum_{j=0}^{i-1} \frac{\hat{h}_{i.j}}{L}\nabla f(x_j) = \frac{N-k-2}{N-k+2}&\left(x_0 - \sum_{i=1}^{k}\sum_{j=0}^{i-1}\frac{\hat{h}_{i,j}}{L}\nabla f(x_j) - \frac{1}{L}\nabla f(x_k)\right) \\ &+ \frac{4}{N-k+2}\left(x_0 - \sum_{j=0}^{k}\frac{N-k+1}{2L}\nabla f(x_j)\right), \end{aligned}$$

which is re-written form of OBL-G$_\flat$. Comparing $\nabla f(x_j)$'s each coefficient, we should prove

$$\begin{aligned} \sum_{i=j+1}^{k+1}\hat{h}_{i,j} &= \frac{N-k-2}{N-k+2}\sum_{i=j+1}^{k}\hat{h}_{i,j} + \frac{4}{N-k+2}\frac{N-j+1}{2} \qquad j \in \{0, 1, \dots, k-1\} \\ \hat{h}_{k+1,k} &= \frac{N-k-2}{N-k+2} + \frac{4}{N-k+2}\frac{N-k+1}{2}, \end{aligned}$$

which can be easily checked with the matrix H. ■