Optimal Acceleration for Minimax and Fixed-Point Problems is Not Unique

TaeHo Yoon [1] **Jaeyeon Kim** [1] **Jaewook J. Suh** [1] **Ernest K. Ryu** [2]

Abstract

Recently, accelerated algorithms using the anchoring mechanism for minimax optimization and fixed-point problems have been proposed, and matching complexity lower bounds establish their optimality. In this work, we present the surprising observation that the optimal acceleration mechanism in minimax optimization and fixed-point problems is not unique. Our new algorithms achieve exactly the same worst-case convergence rates as existing anchor-based methods while using materially different acceleration mechanisms. Specifically, these new algorithms are dual to the prior anchor-based accelerated methods in the sense of H-duality. This finding opens a new avenue of research on accelerated algorithms since we now have a family of methods that empirically exhibit varied characteristics while having the same optimal worst-case guarantee.

1. Introduction

Accelerated algorithms using the so-called anchoring mechanism have been recently proposed for solving minimax optimization and fixed-point problems. Furthermore, these algorithms are optimal: for minimax problems, the gap between the upper and lower bounds is a constant factor of 16, and for fixed-point problems, there is no gap, not even a constant factor. Therefore, anchoring was thought to be "the" correct acceleration mechanism for these setups.

In this work, however, we present the surprising observation that the optimal acceleration mechanism in minimax optimization and fixed-point problems is not unique. For minimax optimization, we introduce a new algorithm with the same worst-case rate as the best-known algorithm. For fixed-point problems, we introduce a continuous family of exact optimal algorithms, all achieving the same worst-case rate that exactly matches the known lower bound. The representative cases of our new accelerated algorithms are dual algorithms of the prior anchor-based accelerated algorithms in the sense of H-duality. The resulting new acceleration mechanisms are materially different from the existing anchoring mechanism.

These findings show that anchor-based acceleration is not unique and sufficient as the mechanism of achieving the exact optimal complexity and enable us to correctly reframe the study of optimal acceleration as a study of a *family* of acceleration mechanisms rather than a singular one. This shift in perspective will likely be critical in the future research toward a more complete and fundamental understanding of accelerations in fixed-point and minimax problems.

1.1. Preliminaries

We use standard notation for set-valued operators (Bauschke & Combettes, 2017; Ryu & Yin, 2022). An operator $\mathbb{A}\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a set-valued function (so $\mathbb{A}(x) \subseteq \mathbb{R}^d$ for $x \in \mathbb{R}^d$). For simplicity, we write $\mathbb{A}x = \mathbb{A}(x)$. The graph of $\mathbb{A}$ is defined and denoted as $\operatorname{Gra}\mathbb{A} = \{(x,y) \,|\, x \in \mathbb{R}^d, y \in \mathbb{A}x\}$. The inverse of $\mathbb{A}$ is defined by $\mathbb{A}^{-1}y = \{x \in \mathbb{R}^d \,|\, y \in \mathbb{A}x\}$. Scalar multiples and sums of operators are defined in the Minkowski sense. If $\mathbb{T}x$ is a singleton for all $x \in \mathbb{R}^d$, we write $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$ and treat it as a function. An operator $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz ($L > 0$) if $\|\mathbb{T}x - \mathbb{T}y\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$. We say $\mathbb{T}$ is nonexpansive if it is 1-Lipschitz.

A function $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is convex-concave if $\mathbf{L}(u,v)$ is convex in u for all fixed $v \in \mathbb{R}^m$ and concave in v for all fixed $u \in \mathbb{R}^n$. If $\mathbf{L}(u_\star, v) \leq \mathbf{L}(u_\star, v_\star) \leq \mathbf{L}(u, v_\star)$ for all $(u,v) \in \mathbb{R}^n \times \mathbb{R}^m$, then $(u_\star, v_\star)$ is a saddle point of $\mathbf{L}$. For $L > 0$, if $\mathbf{L}$ is differentiable and $\nabla\mathbf{L}$ is L-Lipschitz on $\mathbb{R}^n \times \mathbb{R}^m$, we say $\mathbf{L}$ is L-smooth. In this case, we define the saddle operator of $\mathbf{L}$ by $\nabla_\pm\mathbf{L}(u,v) = (\nabla_u\mathbf{L}(u,v), -\nabla_v\mathbf{L}(u,v))$. In most of the cases, we use the joint variable notation $x = (u,v)$ and concisely write $\nabla_\pm\mathbf{L}(x)$ in place of $\nabla_\pm\mathbf{L}(u,v)$.

1.2. Related work

Here, we quickly review the most closely related prior work while deferring the more comprehensive literature survey to Appendix A.

[1]Department of Mathematical Sciences, Seoul National University [2]Department of Mathematics, University of California, Los Angeles. Correspondence to: Ernest Ryu <eryu@math.ucla.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Fixed-point algorithms. A fixed-point problem solves

$$\underset{y\in\mathbb{R}^d}{\text{find}} \quad y = \mathbb{T}y \tag{1}$$

for $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$. The magnitude of the fixed-point residual $y_k - \mathbb{T}y_k$ is one natural performance measure. Sabach & Shtern (2017) first achieved the rate $\|y_k - \mathbb{T}y_k\|^2 = \mathcal{O}(1/k^2)$ through the Sequential Averaging Method, and Lieder (2021) showed that Halpern iteration with specific parameters, which we call OHM in Section 2.1, improves upon the rate of Sabach & Shtern (2017) by a factor of 16. Furthermore, Park & Ryu (2022) provided a matching complexity lower bound showing that the rate of Lieder (2021) is exactly optimal.

Minimax algorithms. Minimax optimization solves

$$\underset{u\in\mathbb{R}^n}{\text{minimize}}\ \underset{v\in\mathbb{R}^m}{\text{maximize}} \quad \mathbf{L}(u,v) \tag{2}$$

for $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. Under the assumption of convex-concavity, $\|\nabla \mathbf{L}(u_k, v_k)\|^2$ is one natural performance measure. Yoon & Ryu (2021) first provided the (order-optimal) accelerated $\|\nabla \mathbf{L}(u_k, v_k)\|^2 = \mathcal{O}(1/k^2)$ rate via the Extra Anchored Gradient (EAG) algorithm together with $\Omega(1/k^2)$ complexity lower bound. The Fast Extragradient (FEG) algorithm of (Lee & Kim, 2021) then improved this rate by a constant factor, achieving the currently best-known constant.

Duality of algorithms. H-duality (Kim et al., 2023a;b) is a duality correspondence between algorithms. The H-duality theory of Kim et al. (2023a) shows that in smooth convex minimization, an algorithm's rate with respect to function value translates to the rate of its H-dual algorithm with respect to gradient norm and vice versa. Our paper establishes a different H-duality theory for fixed-point algorithms.

1.3. Contribution and organization

Contributions. This work is presenting a new class of accelerations in fixed-point and minimax problems. Our findings provide the perspective that the study of optimal acceleration must be viewed as a study of a *family* of acceleration mechanisms rather than a singular one.

Organization. Section 2 provides an overview of the novel algorithms Dual-OHM and Dual-FEG and their continuous-time model. Section 3 presents the analysis of Dual-OHM. Section 4 presents an infinite family of fixed-point algorithms achieving the same exact optimal rates. Section 5 presents the H-duality for fixed-point problems, which explains the symmetry and connection underlying OHM and Dual-OHM. Section 6 provides the analysis of Dual-FEG and its H-dual relationship with FEG. Section 7 explores a continuous-time perspective of the new algorithms. Section 8 provides numerical simulations.

2. Summary of new acceleration results

In this section, we provide an overview of novel accelerated algorithms for several setups. For each setup, we first review the existing (primal) algorithm using the anchor acceleration mechanism and then show its dual counterpart with identical rates but using a materially different acceleration mechanism. (The meaning "dual" is clarified later.) Throughout the paper, we write $N \geq 1$ to denote the pre-specified iteration count of the algorithm.

2.1. Fixed-point problems

Consider the fixed-point problem (1), where $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$ is nonexpansive. We denote $\operatorname{Fix}\mathbb{T} = \{y \in \mathbb{R}^d \,|\, y = \mathbb{T}y\}$ and assume $\operatorname{Fix}\mathbb{T} \neq \emptyset$.

The (primal) Optimal Halpern Method (OHM)[1] (Halpern, 1967; Lieder, 2021) is

$$y_{k+1} = \frac{k+1}{k+2}\mathbb{T}y_k + \frac{1}{k+2}y_0 \tag{OHM}$$

for $k = 0, 1, \ldots$. Equivalently, we can write

$$y_{k+1} = y_k - \frac{1}{k+2}(y_k - \mathbb{T}y_k) + \frac{k}{k+2}(\mathbb{T}y_k - \mathbb{T}y_{k-1})$$

where we define $\mathbb{T}y_{-1} = y_0$. OHM exhibits the rate

$$\|y_{k-1} - \mathbb{T}y_{k-1}\|^2 \leq \frac{4\|y_0 - y_\star\|^2}{k^2}$$

for $k = 1, 2, \ldots$ and $y_\star \in \operatorname{Fix}\mathbb{T}$ (Lieder, 2021).

We present the new method, Dual Optimal Halpern Method (Dual-OHM):

$$y_{k+1} = y_k + \frac{N-k-1}{N-k}(\mathbb{T}y_k - \mathbb{T}y_{k-1}) \tag{Dual-OHM}$$

for $k = 0, 1, \ldots, N-2$, where we define $\mathbb{T}y_{-1} = y_0$. Equivalently,

$$\begin{aligned} z_{k+1} &= \frac{N-k-1}{N-k}z_k - \frac{1}{N-k}(y_k - \mathbb{T}y_k) \\ y_{k+1} &= \mathbb{T}y_k - z_{k+1} \end{aligned} \tag{3}$$

for $k = 0, 1, \ldots, N-2$, where $z_0 = 0$. Dual-OHM exhibits the rate

$$\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 \leq \frac{4\|y_0 - y_\star\|^2}{N^2}$$

for $y_\star \in \operatorname{Fix}\mathbb{T}$. This rate exactly coincides with the rate of OHM for $k = N$. We discuss the detailed analysis in Section 3.

[1] Some prior work referred to this method as the "Optimized" Halpern Method, but we now know the method is (exactly) optimal as (Park & Ryu, 2022) provided a matching lower bound.

As shown in (Park & Ryu, 2022), there exists a nonexpansive operator $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$ with $d \geq 2N-2$ such that

$$\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 \geq \frac{4\|y_0 - y_\star\|^2}{N^2}$$

for any deterministic algorithm using $N-1$ evaluations of $\mathbb{T}$. Therefore, OHM is exactly optimal; it cannot be improved, not even by a constant factor, in terms of worst-case performance. The discovery of Dual-OHM is surprising as it shows that the exact optimal algorithm is not unique.

2.2. Smooth convex-concave minimax optimization

Consider the minimax optimization problem (2), where $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is convex-concave and L-smooth. Convex-concave minimax problems are closely related to fixed-point problems, and the anchoring mechanism of OHM for accelerating fixed-point algorithms has been used to accelerate algorithms for minimax problems (Yoon & Ryu, 2021; Lee & Kim, 2021). We show that Dual-OHM also has its minimax counterpart. In the following, write $\mathbb{A} = \nabla_{\pm}\mathbf{L}$ for notational conciseness.

The (primal) Fast Extragradient[2] (FEG) (Lee & Kim, 2021) is

$$\begin{aligned} x_{k+1/2} &= x_k + \frac{1}{k+1}(x_0 - x_k) - \frac{k}{k+1}\alpha\mathbb{A}x_k \\ x_{k+1} &= x_k + \frac{1}{k+1}(x_0 - x_k) - \alpha\mathbb{A}x_{k+1/2} \end{aligned} \tag{FEG}$$

for $k = 0, 1, \ldots$. If $0 < \alpha \leq \frac{1}{L}$, FEG exhibits the rate

$$\|\nabla\mathbf{L}(x_k)\|^2 = \|\mathbb{A}x_k\|^2 \leq \frac{4\|x_0 - x_\star\|^2}{\alpha^2 k^2}$$

for $k = 1, 2, \ldots$ and a saddle point (solution) $x_\star = (u_\star, v_\star)$. To the best of our knowledge, this result with $\alpha = \frac{1}{L}$ is the fastest known rate.

We present the new method, Dual Fast Extragradient (Dual-FEG):

$$\begin{aligned} x_{k+1/2} &= x_k - \alpha z_k - \alpha\mathbb{A}x_k \\ x_{k+1} &= x_{k+1/2} - \frac{N-k-1}{N-k}\alpha\left(\mathbb{A}x_{k+1/2} - \mathbb{A}x_k\right) \\ z_{k+1} &= \frac{N-k-1}{N-k}z_k - \frac{1}{N-k}\mathbb{A}x_{k+1/2} \end{aligned} \tag{Dual-FEG}$$

for $k = 0, 1, \ldots, N-1$, where $z_0 = 0$. For $0 < \alpha \leq \frac{1}{L}$, Dual-FEG exhibits the rate

$$\|\nabla\mathbf{L}(x_N)\|^2 = \|\mathbb{A}x_N\|^2 \leq \frac{4\|x_0 - x_\star\|^2}{\alpha^2 N^2}.$$

[2]FEG was designed primarily for weakly nonconvex-nonconcave problems, but we consider its application to the special case of convex-concave problems.

This rate exactly coincides with the rate of FEG for $k = N$. We discuss the detailed analysis in Section 6.

As shown in (Yoon & Ryu, 2021), there exists L-smooth convex-concave $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with $m = n \geq 3N+2$ such that

$$\|\nabla\mathbf{L}(x_N)\|^2 \geq \frac{L^2\|x_0 - x_\star\|^2}{(2\lfloor N/2\rfloor + 1)^2}$$

for any deterministic N-step first-order algorithm. Therefore, FEG and Dual-FEG are optimal up to a constant factor.

2.3. Continuous-time analysis

This section introduces continuous-time analyses corresponding to the algorithms of Sections 2.1 and 2.2. In continuous-time limits, the algorithms for fixed-point problems and minimax optimization reduce to the same continuous-time ODE. Let $\mathbb{A} = 2(\mathbb{I} + \mathbb{T})^{-1} - \mathbb{I}$ for fixed-point problem (1) and $\mathbb{A} = \nabla_{\pm}\mathbf{L}$ for minimax problem (2).

The (primal) Anchor ODE (Ryu et al., 2019; Suh et al., 2023) is

$$\dot{X}(t) = -\mathbb{A}(X(t)) + \frac{1}{t}(X_0 - X(t)) \tag{4}$$

which has an equivalent 2nd-order form

$$\ddot{X}(t) + \frac{2}{t}\dot{X} + \frac{1}{t}\mathbb{A}(X(t)) + \frac{d}{dt}\mathbb{A}(X(t)) = 0$$

where $X(0) = X_0$ (and $\dot{X}(0) = -\frac{1}{2}\mathbb{A}(X_0)$ for 2nd-order form) is the initial condition. Anchor ODE exhibits the rate

$$\|\mathbb{A}(X(t))\|^2 \leq \frac{4\|X_0 - X_\star\|^2}{t^2}$$

for $t > 0$, where $X_\star$ is a solution (zero of $\mathbb{A}$).

We present the new Dual-Anchor ODE:

$$\begin{aligned} \dot{X}(t) &= -Z(t) - \mathbb{A}(X(t)) \\ \dot{Z}(t) &= -\frac{1}{T-t}Z(t) - \frac{1}{T-t}\mathbb{A}(X(t)), \end{aligned} \tag{5}$$

which has an equivalent 2nd-order form

$$\ddot{X}(t) + \frac{1}{T-t}\dot{X}(t) + \frac{d}{dt}\mathbb{A}(X(t)) = 0$$

for $t \in (0, T)$, where $T > 0$ is a pre-specified terminal time, and $X(0) = X_0$ and $Z(0) = 0$ (or $\dot{X}(0) = -\mathbb{A}(X_0)$ for the 2nd-order form) are initial conditions. Dual-Anchor ODE exhibits the rate

$$\|\mathbb{A}(X(T))\|^2 \leq \frac{4\|X_0 - X_\star\|^2}{T^2}.$$

This rate exactly coincides with the rate of Anchor ODE for $t = T$. We discuss the detailed analysis in Section 7.

"Dual" in the sense of H-duality. Our new algorithms are "dual" to the known primal algorithms in the sense of H-duality, a recently developed notion of duality between convex minimization algorithms (Kim et al., 2023a). We provide the detailed discussion of H-duality for fixed-point algorithms in Section 5.

3. Analysis of Dual-OHM

In this section, we present the convergence analysis of Dual-OHM, showing that it is another exact optimal algorithm for solving nonexpansive fixed-point problems.

3.0. Preliminaries: Monotone operators

In the following, we express our analysis using the language of monotone operators. We quickly set up the notation and review the connections between fixed-point problems and monotone operators.

Monotone operators. A set-valued (non-linear) operator $\mathbb{A}\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is monotone if $\langle g-g', x-x'\rangle \geq 0$ for all $x, x' \in \mathbb{R}^d$, $g \in \mathbb{A}x$, and $g' \in \mathbb{A}x'$. If $\mathbb{A}$ is monotone and there is no monotone operator $\mathbb{A}'$ for which $\operatorname{Gra}\mathbb{A} \subset \operatorname{Gra}\mathbb{A}'$ properly, then $\mathbb{A}$ is maximally monotone. If $\mathbb{A}$ is maximally monotone, then its resolvent $\mathbb{J}_{\mathbb{A}} := (\mathbb{I}+\mathbb{A})^{-1}\colon \mathbb{R}^d \to \mathbb{R}^d$ is a well-defined single-valued operator.

Fixed-point problems are monotone inclusion problems. There exists a natural correspondence between the classes of nonexpansive operators and maximally monotone operators in the following sense.

Proposition 3.1. *(Eckstein & Bertsekas, 1992, Theorem 2) If $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$ is a nonexpansive operator, then $\mathbb{A} = 2(\mathbb{I}+\mathbb{T})^{-1} - \mathbb{I}$ is maximally monotone. Conversely, if $\mathbb{A}\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally monotone, then $\mathbb{T} = 2\mathbb{J}_{\mathbb{A}} - \mathbb{I}$ is nonexpansive.*

When $\mathbb{T} = 2\mathbb{J}_{\mathbb{A}} - \mathbb{I}$, we have $x = \mathbb{T}x \iff 0 \in \mathbb{A}x$, i.e., $\operatorname{Fix}\mathbb{T} = \operatorname{Zer}\mathbb{A} := \{x \in \mathbb{R}^d \,|\, 0 \in \mathbb{A}x\}$. Therefore, Proposition 3.1 induces a one-to-one correspondence between nonexpansive fixed-point problems and monotone inclusion problems.

Fixed-point residual norm and operator norm. Given $y \in \mathbb{R}^d$, its accuracy as an approximate fixed-point solution is often measured by $\|y - \mathbb{T}y\|$, the norm of fixed-point residual. Let $\mathbb{A}$ be the maximal monotone operator satisfying $\mathbb{T} = 2\mathbb{J}_{\mathbb{A}} - \mathbb{I}$, and let $x = \mathbb{J}_{\mathbb{A}}y$. Then we see that

$$y \in (\mathbb{I}+\mathbb{A})(x) = x + \mathbb{A}x \iff y - x \in \mathbb{A}x.$$

Denote $\tilde{\mathbb{A}}x = y - x \in \mathbb{A}x$. Then

$$y - \mathbb{T}y = y - (2\mathbb{J}_{\mathbb{A}} - \mathbb{I})(y) = 2(y - \mathbb{J}_{\mathbb{A}}y) = 2\tilde{\mathbb{A}}x.$$

Therefore, $\|y - \mathbb{T}y\| = 2\|\tilde{\mathbb{A}}x\|$.

Minimax optimization and monotone operators. The minimax problem (2) can also be recast as a monotone inclusion problem. Precisely, for L-smooth convex-concave $\mathbf{L}$, its saddle operator $\nabla_{\pm}\mathbf{L}$ is monotone and L-Lipschitz. In this case, $x_\star = (u_\star, v_\star)$ is a minimax solution for $\mathbf{L}$ if and only if $\nabla_{\pm}\mathbf{L}(x_\star) = 0$.

Finally, we quickly state a handy lemma used in the convergence analyses throughout the paper.

Lemma 3.2. *Let $\mathbb{A}\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be monotone and let $x, y \in \mathbb{R}^d$, $\tilde{\mathbb{A}}x \in \mathbb{A}x$. Suppose, for some $\rho > 0$,*

$$\rho\left\|\tilde{\mathbb{A}}x\right\|^2 + \left\langle \tilde{\mathbb{A}}x, x - y\right\rangle \leq 0$$

holds. Then, for $x_\star \in \operatorname{Zer}\mathbb{A}$, $\left\|\tilde{\mathbb{A}}x\right\|^2 \leq \frac{\|y-x_\star\|^2}{\rho^2}$.

Proof. By monotonicity of $\mathbb{A}$ and Young's inequality,

$$\begin{aligned} 0 &\geq \rho\left\|\tilde{\mathbb{A}}x\right\|^2 + \left\langle \tilde{\mathbb{A}}x, x - y\right\rangle \\ &\geq \rho\left\|\tilde{\mathbb{A}}x\right\|^2 + \left\langle \tilde{\mathbb{A}}x, x_\star - y\right\rangle \\ &\geq \frac{\rho}{2}\left\|\tilde{\mathbb{A}}x\right\|^2 - \frac{1}{2\rho}\|x_\star - y\|^2. \end{aligned}$$

□

3.1. Convergence analyses of OHM and Dual-OHM

We formally state the convergence result of Dual-OHM and outline its proof.

Theorem 3.3. *Let $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$ be nonexpansive and $y_\star \in \operatorname{Fix}\mathbb{T}$. For $N \geq 1$, Dual-OHM exhibits the rate*

$$\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 \leq \frac{4\|y_0 - y_\star\|^2}{N^2}.$$

Proof outline. Let $\mathbb{A}$ be the unique maximally monotone operator such that $\mathbb{T} = 2\mathbb{J}_{\mathbb{A}} - \mathbb{I}$ (defined as in Proposition 3.1). Let $x_{k+1} = \mathbb{J}_{\mathbb{A}}(y_k)$ for $k = 0, 1, \ldots$, so that $\tilde{\mathbb{A}}x_{k+1} = y_k - x_{k+1} \in \mathbb{A}x_{k+1}$. Recall the alternative form (3) of Dual-OHM. Define

$$\begin{aligned} V_k = &-\frac{N-k-1}{N-k}\left\|z_k + 2\tilde{\mathbb{A}}x_N\right\|^2 \\ &+ \frac{2}{N-k}\left\langle z_k + 2\tilde{\mathbb{A}}x_N, y_k - y_{N-1}\right\rangle \end{aligned}$$

for $k = 0, 1, \ldots, N-1$. We show in Appendix C that

$$\begin{aligned} &V_k - V_{k+1} \\ &= \frac{4}{(N-k)(N-k-1)}\left\langle x_N - x_{k+1}, \tilde{\mathbb{A}}x_N - \tilde{\mathbb{A}}x_{k+1}\right\rangle, \end{aligned}$$

i.e., $V_k \geq V_{k+1}$ for $k = 0, 1, \ldots, N-2$. Observe that $V_{N-1} = 0$ and because $z_0 = 0$,

$$\begin{aligned} V_0 &= -\frac{4(N-1)}{N}\left\|\tilde{\mathbb{A}}x_N\right\|^2 + \frac{4}{N}\left\langle \tilde{\mathbb{A}}x_N, y_0 - y_{N-1}\right\rangle \\ &= -4\left\|\tilde{\mathbb{A}}x_N\right\|^2 + \frac{4}{N}\left\langle \tilde{\mathbb{A}}x_N, y_0 - x_N\right\rangle \end{aligned}$$

where the second line uses $x_N = y_{N-1} - \tilde{\mathbb{A}}x_N$. Finally, divide both sides of $V_0 \geq \cdots \geq V_{N-1} = 0$ by $\frac{4}{N}$, apply Lemma 3.2 and the identity $y_{N-1} - \mathbb{T}y_{N-1} = 2\tilde{\mathbb{A}}x_N$:

$$\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 = 4\left\|\tilde{\mathbb{A}}x_N\right\|^2 \leq \frac{4\|y_0 - y_\star\|^2}{N^2}.$$

□

We point out that the convergence analysis for OHM can be done in a similar style (Lieder, 2021). Define $\mathbb{A}$, x_{k+1} and $\tilde{\mathbb{A}}x_{k+1}$ as above. Define $U_0 = 0$ and

$$U_k = k^2\left\|\tilde{\mathbb{A}}x_k\right\|^2 + k\left\langle \tilde{\mathbb{A}}x_k, x_k - y_0\right\rangle$$

for $k = 1, 2, \ldots$. It can be shown that (Ryu & Yin, 2022)

$$U_j - U_{j+1} = j(j+1)\left\langle x_{j+1} - x_j, \tilde{\mathbb{A}}x_{j+1} - \tilde{\mathbb{A}}x_j\right\rangle \geq 0$$

for $j = 0, 1, \ldots$. Then $0 = U_0 \geq \cdots \geq U_k$, and dividing both sides by k and applying Lemma 3.2 gives the rate $\|\tilde{\mathbb{A}}x_k\|^2 \leq \frac{\|y_0 - x_\star\|^2}{k^2}$.

3.2. Proximal forms of OHM and Dual-OHM

It is known that OHM can be equivalently written as

$$\begin{aligned} x_{k+1} &= \mathbb{J}_{\mathbb{A}}(y_k) \\ y_{k+1} &= x_{k+1} + \frac{k}{k+2}(x_{k+1} - x_k) - \frac{k}{k+2}(x_k - y_{k-1}) \end{aligned}$$

for $k = 0, 1, \ldots$, where $x_0 = y_0$. This proximal form is called Accelerated Proximal Point Method (APPM) (Kim, 2021). Likewise, Dual-OHM can be equivalently written as

$$\begin{aligned} x_{k+1} &= \mathbb{J}_{\mathbb{A}}(y_k) \\ y_{k+1} &= x_{k+1} + \frac{N-k-1}{N-k}(x_{k+1} - x_k) \\ &\quad - \frac{N-k-1}{N-k}(x_k - y_{k-1}) - \frac{1}{N-k}(x_{k+1} - y_k) \end{aligned}$$

for $k = 0, 1, \ldots, N-2$, where $x_{-1} = y_0 = x_0$. We prove the equivalence in Appendix B.

4. Continuous family of exact optimal fixed-point algorithms

Upon seeing the two algorithms OHM and Dual-OHM exhibiting the same exact optimal rate, it is natural to ask whether there are other exact optimal algorithms. In this section, we show that there is, in fact, an $(N-2)$-dimensional continuous family of exact optimal algorithms.

4.1. H-matrix representation

Fixed-point algorithm of $N-1$ iterations with fixed (non-adaptive) step-sizes can be written in the form

$$y_{k+1} = y_k - \sum_{j=0}^{k} h_{k+1,j+1}\underbrace{(y_j - \mathbb{T}y_j)}_{2\tilde{\mathbb{A}}x_{j+1}} \tag{6}$$

for $k = 0, 1, \ldots, N-2$, where the $\tilde{\mathbb{A}}$ notation uses the convention of Section 3. With this representation, the lower-triangular matrix $H \in \mathbb{R}^{(N-1)\times(N-1)}$, defined by $(H)_{k,j} = h_{k,j}$ if $j \leq k$ and $(H)_{k,j} = 0$ otherwise, fully specifies the algorithm.

4.2. Optimal algorithm family via H-matrices

We now state our result characterizing a family of exact optimal algorithms.

Theorem 4.1 (Optimal family)**.** *There exist a nonempty open convex set $C \subset \mathbb{R}^{N-2}$ and a continuous injective mapping $\Phi\colon C \to \mathbb{R}^{(N-1)\times(N-1)}$ such that for all $w \in C$ $H = \Phi(w)$ is lower-triangular, and algorithm (6) defined via the H-matrix $H = \Phi(w)$ exhibits the exact optimal rate (matching the lower bound)*

$$\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 = 4\left\|\tilde{\mathbb{A}}x_N\right\|^2 \leq \frac{4\|y_0 - y_\star\|^2}{N^2}$$

for nonexpansive $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$ and $y_\star \in \operatorname{Fix}\mathbb{T}$.

We briefly outline the high-level idea of the proof while deferring the details to Appendix E. From Section 3.1, we observe that the convergence proofs for OHM and Dual-OHM both work by establishing the identity

$$\begin{aligned} 0 = \left\langle \tilde{\mathbb{A}}x_N, x_N - y_0\right\rangle + N\|\tilde{\mathbb{A}}x_N\|^2 \\ + \sum_{(i,j)\in I}\lambda_{i,j}\left\langle \tilde{\mathbb{A}}x_i - \tilde{\mathbb{A}}x_j, x_i - x_j\right\rangle, \end{aligned} \tag{7}$$

where I is some set of tuple of indices (i, j) with $i > j$ (for both algorithms, the consecutive differences of Lyapunov functions are of the form $\lambda_{i,j}\left\langle \tilde{\mathbb{A}}x_i - \tilde{\mathbb{A}}x_j, x_i - x_j\right\rangle$; sum them up to obtain (7)) and $\lambda_{i,j} \geq 0$. For OHM, I consists of $(k+1, k)$ for $k = 1, \ldots, N-1$ while for Dual-OHM, (N, k) for $k = 1, \ldots, N-1$ are used. In the proof of Theorem 4.1, we identify algorithms (in terms of H-matrices) whose convergence can be proved via (7) where I is the union of the two sets of tuples, i.e.,

$$\begin{aligned} I = \{(k+1, k) \mid k = 1, \ldots, N-1\} \\ \cup \{(N, k) \mid k = 1, \ldots, N-1\}. \end{aligned}$$

Once (7) is established, we obtain the desired convergence rate from monotonicity of $\mathbb{A}$ and Lemma 3.2.

To clarify, the algorithm family as defined in Theorem 4.1 does not include OHM and Dual-OHM as C is an open set and OHM and Dual-OHM correspond to points on the boundary $\partial C = \overline{C}\backslash C$. We choose not to incorporate ∂C in Theorem 4.1 because doing so leads to cumbersome divisions-by-zero. However, with a specialized analysis, Proposition 4.2 shows that OHM and Dual-OHM are indeed a part of the parametrization Φ and therefore that the optimal family "connects" OHM and Dual-OHM. The proof is deferred to Appendix F.1.

Proposition 4.2. *Consider C and Φ defined in Theorem 4.1. One can continuously extend Φ to some $u, v \in \partial C$ such that $\Phi(u) = H_{OHM}$ and $\Phi(v) = H_{Dual\text{-}OHM}$, where H_{OHM} and $H_{Dual\text{-}OHM}$ are respectively the H-matrix representations of OHM and Dual-OHM.*

Our proof of Theorem 4.1 comes without an explicit characterization of Φ or the H-matrices (we only show existence[3]). Fortunately, for $N = 3$, we do have a simple explicit characterization, which directly illustrates the family interpolating between the OHM and Dual-OHM. For any 2×2 lower-triangular $\begin{bmatrix} h_{1,1} & 0 \\ h_{2,1} & h_{2,2} \end{bmatrix}$ satisfying $h_{2,1} = 1 - h_{1,1} - h_{2,2}$, $h_{1,1}h_{2,2} = \frac{1}{3}$ and $h_{1,1}, h_{2,2} \in \left[\frac{1}{2}, \frac{2}{3}\right]$, the iterate y_2 defined by (6) satisfies $\|y_2 - \mathbb{T}y_2\|^2 = 4\left\|\tilde{\mathbb{A}}x_3\right\|^2 \leq \frac{4\|y_0 - y_\star\|^2}{9}$. In particular, OHM and Dual-OHM each correspond to $(h_{1,1}, h_{2,2}) = \left(\frac{1}{2}, \frac{2}{3}\right)$ and $(h_{1,1}, h_{2,2}) = \left(\frac{2}{3}, \frac{1}{2}\right)$.

As a final comment, we clarify that Theorem 4.1 is not the exhaustive parametrization of optimal algorithms; there are other exact optimal algorithms that Theorem 4.1 does not cover. We provide such examples in Appendix F.3. The complete characterization of the set of exact optimal algorithms seems to be challenging, and we leave it to future work.

5. H-duality for fixed-point algorithms

In this section, we present an H-duality theory for fixed-point algorithms. At a high level, H-duality states that an algorithm and its convergence proof can be *dualized* in a certain sense. This result provides a formal connection between OHM and Dual-OHM.

5.1. H-dual operation

Before providing the formal definition, we observe the following. The H-matrix of OHM is

$$(H_{\text{OHM}})_{k,j} = \begin{cases} -\frac{j}{k(k+1)} & \text{if } j < k \\ \frac{k}{k+1} & \text{if } j = k \end{cases}$$

while the H-matrix of Dual-OHM is

$$(H_{\text{Dual-OHM}})_{k,j} = \begin{cases} -\frac{N-k}{(N-j)(N-j+1)} & \text{if } j < k \\ \frac{N-k}{N-k+1} & \text{if } j = k. \end{cases}$$

We derive the H-matrix representations in Appendix B. For now, note the relationship $H_{\text{Dual-OHM}} = H_{\text{OHM}}^{\text{A}}$, where superscript A denotes anti-diagonal transpose $\left(H^{\text{A}}\right)_{k,j} = (H)_{N-j,N-k}$. The anti-diagonal transpose operation reflects a square matrix along its anti-diagonal direction and preserves lower triangularity.

[3]The H-matrices are, however, "computable" in the sense that they can be obtained by solving an explicit system of linear equations, as outlined in Appendix E.

Given an algorithm represented by H, we define its *H-dual* as the algorithm represented by H^{A}. In other words, given a general iterative algorithm defined as (6), its H-dual is the algorithm with the update rule

$$y_{k+1} = y_k - \sum_{j=0}^{k} h_{N-j-1,N-k-1}\left(y_j - \mathbb{T}y_j\right) \tag{8}$$

for $k = 0, 1, \ldots, N-2$.

Proposition 5.1. *Dual-OHM and OHM are H-duals of each other.*

5.2. H-duality theorem

OHM and Dual-OHM are H-duals of each other, and they share the identical rates on $\|y_{N-1} - \mathbb{T}y_{N-1}\|^2$. This symmetry is not a coincidence; the following H-duality theorem explains the connection between them.

Theorem 5.2 (Informal)**.** *Let $H \in \mathbb{R}^{(N-1)\times(N-1)}$ be lower triangular and $\nu > 0$. Then the following are equivalent.*

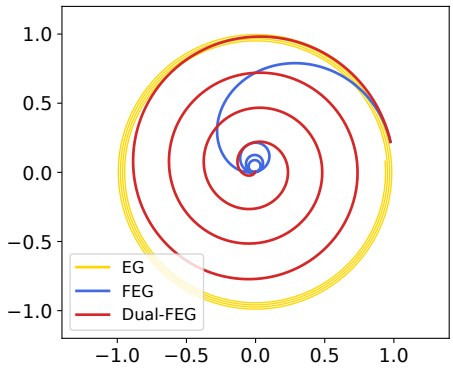
- $\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 \leq \nu \|y_0 - y_\star\|^2$ *for (6) with H can be proved with primal Lyapunov structure.*
- $\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 \leq \nu \|y_0 - y_\star\|^2$ *for (6) with H^A (H-dual) can be proved with dual Lyapunov structure.*

We defer the precise statement and the proof of Theorem 5.2 to Appendix D. The high-level takeaway of Theorem 5.2 is as follows: If an algorithm achieves a certain convergence rate with respect to $\|y_{N-1} - \mathbb{T}y_{N-1}\|^2$ (based on a certain proof structure), then the proof can be *H-dualized* to guarantee the exact same convergence rate for its H-dual. The relationship between OHM and Dual-OHM is an instance of this duality correspondence (with $\nu = 1/N^2$).

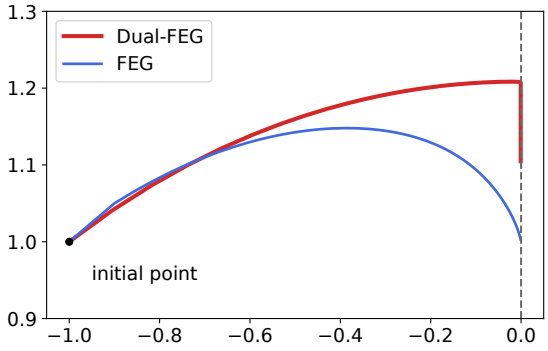
Different from the prior H-duality result for convex minimization (Kim et al., 2023a), we show that algorithms in H-dual relationship share the convergence rate with respect to the same performance measure $\|y_{N-1} - \mathbb{T}y_{N-1}\|^2$ and the same initial condition $\|y_0 - y_\star\|^2$. One could say that the performance measure $\|y_{N-1} - \mathbb{T}y_{N-1}\|^2$ is "self-dual". This contrasts with the prior H-duality (Kim et al., 2023a), which showed that function values are "dual" to gradient magnitude in convex minimization. While the H-duality theory of this work and that of (Kim et al., 2023a) share some superficial similarities, the unifying fundamental principle remains unknown, and finding one is an interesting subject of future research.

6. Analysis of Dual-FEG for minimax problems

In this section, we present the convergence analysis of Dual-FEG and additionally provide the observation that Dual-FEG is the H-dual of FEG.



(a) Trajectories from a 2D bilinear example



(b) Trajectories from a 2D non-bilinear example

Figure 1. Trajectories generated by minimax optimization algorithms on **(left)** $\mathbf{L}(u, v) = uv$ with a random initial point of norm 1 and **(right)** $\mathbf{L}(u, v) = u^2 v$ with initial point $(-1, 1)$, where the dashed vertical line indicates the set of optima (saddle points).

Algorithms such as OHM and Dual-OHM are sometimes referred to as "implicit methods" since they can be expressed using resolvents, as discussed in Section 3.2. The results of this section show that Dual-OHM has an "explicit" counterpart Dual-FEG, which uses direct gradient evaluations instead.

First, we formally state the convergence result.

Theorem 6.1. *Let* $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *be convex-concave and* L*-smooth. For* $N \geq 1$ *fixed, if* $0 < \alpha \leq \frac{1}{L}$*, Dual-FEG exhibits the rate*

$$\|\nabla \mathbf{L}(x_N)\|^2 = \|\nabla_{\pm} \mathbf{L}(x_N)\|^2 \leq \frac{4\|x_0 - x_\star\|^2}{\alpha^2 N^2},$$

where $x_\star \in \operatorname{Zer} \nabla_{\pm} \mathbf{L}$ *is a solution.*

Proof outline. Define $g_N = \nabla_{\pm} \mathbf{L}(x_N)$. In Appendix G.1, we show that

$$V_k = -\alpha \|z_k + g_N\|^2 + \frac{2}{N-k} \langle z_k + g_N, x_k - x_N \rangle$$

is nonincreasing in $k = 0, 1, \ldots, N-1$, and $V_{N-1} \geq 0$. This implies $0 \leq V_0 = -\alpha \|g_N\|^2 + \frac{2}{N} \langle g_N, x_0 - x_N \rangle$. Finally, divide both sides by $\frac{2}{N}$ and apply Lemma 3.2. □

Dual-FEG and FEG are H-duals. Dual-FEG has intermediate iterates $x_{k+1/2}$, serving a role similar to intermediate iterates of FEG. Consider the following H-matrix representation for Dual-FEG, analogous to (6):

$$x_{(\ell+1)/2} = x_{\ell/2} - \frac{1}{L} \sum_{i=0}^{\ell} h_{(\ell+1)/2, i/2} \nabla_{\pm} \mathbf{L}(x_{i/2})$$

for $\ell = 0, 1, \ldots, 2N-1$. For Dual-FEG, the H-matrix is the $2N \times 2N$ lower-triangular matrix

$$(H_{\text{Dual-FEG}})_{\ell, i} = \begin{cases} h_{\ell/2, (i-1)/2} & \text{if } i \leq \ell \\ 0 & \text{if } i > \ell. \end{cases}$$

We can analogously define the H-matrix H_{FEG} for FEG, which also has extragradient-type intermediate iterates. It turns out that H_{FEG} and $H_{\text{Dual-FEG}}$ are anti-diagonal transposes of each other.

Proposition 6.2. *Dual-FEG and FEG are H-duals of each other.*

We defer the proof to Appendix G.2. This intriguing H-dual relationship and identical convergence rates of FEG and Dual-FEG strongly indicate the possible existence of H-duality theory for smooth minimax optimization; we leave its formal treatment to future work.

7. Continuous-time analysis of dual-anchoring

In this section, we outline the analysis of the Dual-Anchor ODE (5), which is the common continuous-time model for both Dual-OHM and Dual-FEG, as derived in Appendix H.1. We then introduce the notion of H-dual for ODE models and show that Dual-Anchor ODE is the H-dual of Anchor ODE (4), the continuous-time model for OHM and FEG.

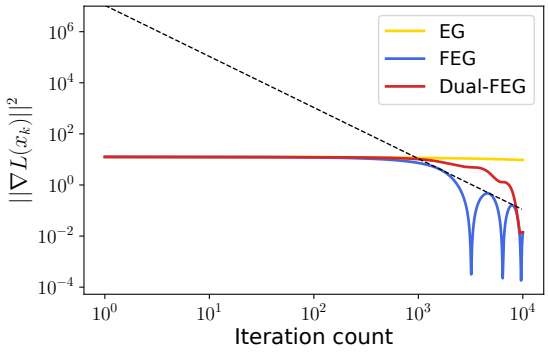
Theorem 7.1. *Let* $\mathbb{A}\colon \mathbb{R}^d \to \mathbb{R}^d$ *be Lipschitz continuous and monotone. For* $T > 0$*, the solution* $X\colon [0, T) \to \mathbb{R}^d$ *of the Dual-Anchor ODE (5) with initial conditions* $X(0) = X_0, Z(0) = 0$ *uniquely exists, and* $X(T) = \lim_{t \to T^-} X(t)$ *satisfies*

$$\|\mathbb{A}(X(T))\|^2 \leq \frac{4\|X_0 - X_\star\|^2}{T^2}$$

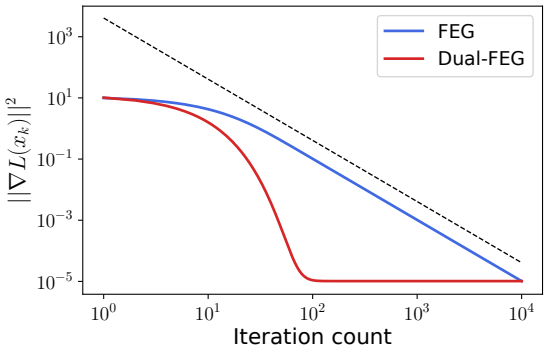
where $X_\star \in \operatorname{Zer} \mathbb{A}$.

Proof outline. Define $V\colon [0, T) \to \mathbb{R}$ by

$$V(t) = -\|Z(t) + \mathbb{A}(X(T))\|^2 + \frac{2}{T-t} \langle Z(t) + \mathbb{A}(X(T)), X(t) - X(T) \rangle.$$



(a) Worst case bilinear problem



(b) Strongly-convex-strongly-concave bilinear problem

Figure 2. Performance of minimax algorithms in reducing $\|\nabla \mathbf{L}(x_k)\|^2$ for bilinear problem instances. The dashed black line indicates the theoretical upper bounds for FEG.

In Appendix H.3.3, we show that

$$\dot{V}(t) = -\frac{2\langle X(t) - X(T), \mathbb{A}(X(t)) - \mathbb{A}(X(T))\rangle}{(T-t)^2} \leq 0.$$

Corollary H.4 shows $\lim_{t\to T^-} V(t) = 0$. From $Z(0) = 0$,

$$V(0) = -\|\mathbb{A}(X(T))\|^2 - \frac{2}{T}\langle \mathbb{A}(X(T)), X(T) - X_0\rangle.$$

Dividing both sides of $0 = \lim_{t\to T^-} V(t) \leq V(0)$ by $\frac{2}{T}$ and applying Lemma 3.2 we get the desired inequality. □

Dual-Anchor ODE and Anchor ODE are H-duals. The continuous-time analogue of the H-matrix representation (Kim & Yang, 2023b; Kim et al., 2023a) is

$$\dot{X}(t) = -\int_0^t H(t,s)\mathbb{A}(X(s))ds,$$

where we refer to $H(t,s)$ as the *H-kernel*. The *H-dual ODE* is defined by

$$\dot{X}(t) = -\int_0^t H^{\mathrm{A}}(t,s)\mathbb{A}(X(s))ds$$

where $H^{\mathrm{A}}(t,s) = H(T-s, T-t)$ is the analogue of the anti-diagonal transpose. We prove the following in Appendix H.2.

Proposition 7.2. *Dual-Anchor ODE* (5) *and Anchor ODE* (4) *are H-duals of each other.*

Generalization to differential inclusion. Theorem 7.1 assumes Lipschitz continuity of $\mathbb{A}$ for simplicity. However, even if we only assume that $\mathbb{A}$ is maximally monotone, the existence of a solution and the convergence result can be established for the differential inclusion

$$\dot{X}(t) \in -Z(t) - \mathbb{A}(X(t))$$
$$\dot{Z}(t) \in -\frac{1}{T-t}Z(t) - \frac{1}{T-t}\mathbb{A}(X(t)),$$

which is a generalized continuous-time model for possibly set-valued operators. We provide the details in Appendix H.4.

8. Experiments

In this section, we present some numerical simulations illustrating the dynamics of dual-anchor algorithm. In Figure 1, we compare the trajectories of FEG, Dual-FEG, and the Extragradient (EG) (Korpelevich, 1976) algorithms. Figure 1a uses a bilinear example $\mathbf{L}(u,v) = uv$ with $\alpha = 0.005$ and $N = 5000$ as an approximation of the algorithms' behavior in the limit $\alpha \to 0$, and Figure 1b uses a non-bilinear example $\mathbf{L}(u,v) = u^2 v$ (which is convex-concave and smooth on $[-1,1] \times \mathbb{R}_{\geq 0}$) with $\alpha = 0.05$ and $N = 10000$. In Figures 2 and 3, we plot $\|\nabla \mathbf{L}(x_k)\|^2$. Figure 2a uses a worst-case bilinear example due to Ouyang & Xu (2021):

$$\mathbf{L}(u,v) = \frac{1}{2}u^\intercal G u - g^\intercal u - \langle Au - b, v\rangle$$

where $u, v \in \mathbb{R}^n$ with $n = 200$. The precise terms of $\mathbf{L}$ are stated in Appendix I.2. We use initial points $u_0 = 0, v_0 = 0$ and use $\alpha = 1.0$ and $N = 10000$. Figure 2b uses the same example of Ouyang & Xu (2021) with additional μ-strongly convex term in u and μ-strongly-concave term in v (with $\mu = 0.1$):

$$\mathbf{L}_\mu(u,v) = \frac{1}{2}u^\intercal(G + \mu I)u - g^\intercal u - \langle Au - b, v\rangle - \frac{\mu}{2}\|v\|^2.$$

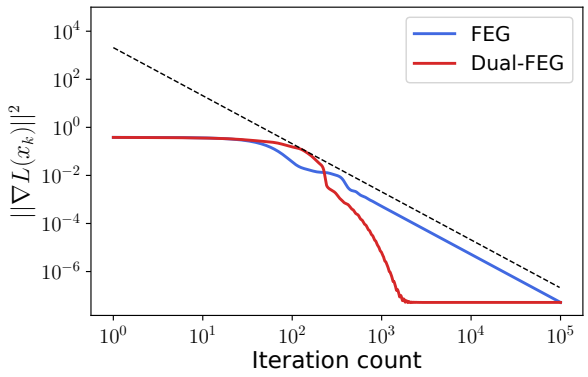


Figure 3. Performance of algorithms in reducing $\|\nabla \mathbf{L}(x_k)\|^2$ for the Lagrangian of a linearly constrained smooth convex minimization problem. The dashed black line indicates the theoretical upper bounds for FEG.

Figure 3 considers the linearly constrained convex minimization problem

$$\begin{array}{ll} \underset{u\in\mathbb{R}^n}{\text{minimize}} & h_\delta(u) \\ \text{subject to} & Au = b \end{array}$$

where $h_\delta(u) = \begin{cases} \frac{1}{2}\|u\|^2 & \text{if } \|u\| \le \delta \\ \delta\|u\| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$ is the Huber loss and $A \in \mathbb{R}^{m\times n}$ $(m < n)$. We run FEG and Dual-FEG on the Lagrangian $\mathbf{L}(u, v) = h_\delta(u) + \langle Au - b, v\rangle$. We choose $n = 100$, $m = 20$, randomly generate entries of A as i.i.d. $\mathcal{N}(0, 1/n^2)$, randomly choose $\frac{n}{10}$ coordinates where $\overline{u} \in \mathbb{R}^n$ has nonzero values (uniform random in $[0, 1]$) and set $b = A\overline{u}$. We use initial points u_0, v_0 with i.i.d. standard normal coordinates and choose $\delta = 0.1$, $\alpha = 0.5$ and $N = 10^5$.

The trajectories of Dual-FEG are qualitatively different from those of FEG, indicating that the two acceleration mechanisms are genuinely different. Interestingly, however, we find that the last iterates of Dual-FEG and FEG are identical when the operator is linear, as we show in Appendix I.3. Indeed, Figures 1a and 2 show the two algorithms arriving at the same point at the terminal iteration. However, this is not true for nonlinear operators; in Figure 1b, we observe that the terminal iterate of Dual-FEG does not coincide with the terminal iterate of FEG.

In Figures 2b and 3, we observe that Dual-FEG approaches the terminal accuracy much earlier compared to FEG, which progresses much more steadily. In Appendix I.4, we prove that in continuous-time, this phenomenon occurs when the objective function is strongly-convex-strongly-concave. We anticipate that the discrete dual algorithms also exhibit a similar phenomenon, and we leave this investigation to future work.

Comparison of primal and dual algorithms. We observe that the dual algorithms require N or T to be specified in advance, while the primal ones do not. In this respect, primal algorithms are advantageous over dual algorithms. On the other hand, we observe that for problems involving strongly monotone operators, dual algorithms exhibit much faster convergence than primal algorithms in the earlier iterations. As the primal and dual algorithms share the same worst-case guarantee but display distinct characteristics, determining the best choice of algorithm may require considering other criterion that depend on the particular application scenario.

9. Conclusion

This work presents a new class of accelerations in fixed-point and minimax problems and provides the correct perspective that optimal acceleration is a family of algorithms. Our findings open several new avenues of research on accelerated algorithms. Since the worst-case guarantee, as a criterion, does not uniquely identify a single best method, additional criteria should be introduced to break the tie. Moreover, the role of H-duality in discovering the new dual accelerated algorithms hints at a deeper yet unexplored significance of H-duality in the theory of first-order algorithms.

Acknowledgements

This work was supported by the Samsung Science and Technology Foundation (Project Number SSTF-BA2101-02). We thank Donghwan Kim and Felix Lieder for the discussion on their processes of using the performance estimation problem (PEP) methodology for finding the exact optimal algorithm presented as APPM and OHM. We thank the anonymous referees for inspiring the exploration of fast convergence of dual algorithms with strongly monotone operators.

Impact statement

This paper presents a theoretical study of optimization. Given its abstract nature, we do not expect our work to raise any significant ethical or societal concerns and consider it neutral in terms of immediate real-world impact.

References

Anderson, D. G. Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 12(4):547–560, 1965.

Aubin, J.-P. and Cellina, A. *Differential Inclusions*. Springer, 1984.

Baillon, J.-B. and Bruck, R. E. Optimal rates of asymptotic regularity for averaged nonexpansive mappings. *Fixed Point Theory and Applications*, 128:27–66, 1992.

Banach, S. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, second edition, 2017.

Borwein, J., Reich, S., and Shafrir, I. Krasnoselski-Mann iterations in normed spaces. *Canadian Mathematical Bulletin*, 35(1):21–28, 1992.

Boţ, R. I., Csetnek, E. R., and Nguyen, D.-K. Fast optimistic gradient descent ascent (OGDA) method in continuous and discrete time. *Foundations of Computational Mathematics*, 2023.

Bravo, M. and Cominetti, R. Sharp convergence rates for averaged nonexpansive maps. *Israel Journal of Mathematics*, 227:163–188, 2018.

Cai, Y. and Zheng, W. Accelerated single-call methods for constrained min-max optimization. *International Conference on Learning Representations*, 2023.

Cai, Y., Oikonomou, A., and Zheng, W. Finite-time last-iterate convergence for learning in multi-player games. *Neural Information Processing Systems*, 2022.

Cominetti, R., Soto, J. A., and Vaisman, J. On the rate of convergence of Krasnosel'skiĭ-Mann iterations and their connection with sums of Bernoullis. *Israel Journal of Mathematics*, 199(2):757–772, 2014.

Csetnek, E. R., Malitsky, Y., and Tam, M. K. Shadow Douglas–Rachford Splitting for Monotone Inclusions. *Applied Mathematics & Optimization*, 80(3):665–678, 2019.

Das Gupta, S., Van Parys, B. P. G., and Ryu, E. K. Branch-and-bound performance estimation programming: A unified methodology for constructing optimal optimization methods. *Mathematical Programming*, 2023.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training GANs with optimism. *International Conference on Learning Representations*, 2018.

Davis, D. and Yin, W. Convergence rate analysis of several splitting schemes. In Glowinski, R., Osher, S. J., and Yin, W. (eds.), *Splitting Methods in Communication, Imaging, Science and Engineering*, Chapter 4, pp. 115–163. Springer, 2016.

Diakonikolas, J. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. *Conference on Learning Theory*, 2020.

Dong, QL., Yuan, HB., Cho, YJ., and Rassias, T. M. Modified inertial Mann algorithm and inertial CQ-algorithm for nonexpansive mappings. *Optimization Letters*, 12(1): 87–102, 2018.

Drori, Y. and Taylor, A. B. Efficient first-order methods for convex minimization: A constructive approach. *Mathematical Programming*, 184(1–2):183–220, 2020.

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

Drori, Y. and Teboulle, M. An optimal variant of Kelley's cutting-plane method. *Mathematical Programming*, 160 (1–2):321–351, 2016.

Eckstein, J. and Bertsekas, D. P. On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

Gorbunov, E., Loizou, N., and Gidel, G. Extragradient method: $O(1/K)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. *International Conference on Artificial Intelligence and Statistics*, 2022.

Halpern, B. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

Ishikawa, S. Fixed points and iteration of a nonexpansive mapping in a Banach space. *Proceedings of the American Mathematical Society*, 59(1):65–71, 1976.

Jang, U., Gupta, S. D., and Ryu, E. K. Computer-assisted design of accelerated composite optimization methods: OptISTA. *arXiv:2305.15704*, 2023.

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

Kim, D. and Fessler, J. A. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1-2):81–107, 2016.

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

Kim, J. and Yang, I. Convergence analysis of ODE models for accelerated first-order methods via positive semidefinite kernels. *NeurIPS*, 2023a.

Kim, J. and Yang, I. Unifying Nesterov's accelerated gradient methods for convex and strongly convex objective functions. *International Conference on Machine Learning*, 2023b.

Kim, J., Ozdaglar, A. E., Park, C., and Ryu, E. K. Time-reversed dissipation induces duality between minimizing gradient norm and function value. *Neural Information Processing Systems*, 2023a.

Kim, J., Park, C., Ozdaglar, A., Diakonikolas, J., and Ryu, E. K. Mirror duality in convex optimization. *arXiv:2311.17296*, 2023b.

Kohlenbach, U. On quantitative versions of theorems due to F. E. Browder and R. Wittmann. *Advances in Mathematics*, 226(3):2764–2795, 2011.

Korpelevich, G. M. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12(4):747–756, 1976.

Krasnosel'skii, M. A. Two remarks on the method of successive approximations. *Uspekhi Matematicheskikh Nauk*, 10(1):123–127, 1955.

Krichene, W., Bayen, A., and Bartlett, P. L. Accelerated mirror descent in continuous and discrete time. *Neural Information Processing Systems*, 2015.

Lee, S. and Kim, D. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Neural Information Processing Systems*, 2021.

Leustean, L. Rates of asymptotic regularity for Halpern iterations of nonexpansive mappings. *Journal of Universal Computer Science*, 13(11):1680–1691, 2007.

Liang, J., Fadili, J., and Peyré, G. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159(1):403–434, 2016.

Lieder, F. *Projection Based Methods for Conic Linear Programming—Optimal First Order Complexities and Norm Constrained Quasi Newton Methods*. Doctoral dissertation, Heinrich-Heine-Universität Düsseldorf, 2018.

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

Lu, H. An $O(s^r)$-resolution ODE framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems. *Mathematical Programming*, 194(1):1061–1112, 2022.

Maingé, P.-E. Convergence theorems for inertial KM-type algorithms. *Journal of Computational and Applied Mathematics*, 219(1):223–236, 2008.

Mann, W. R. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3):506–510, 1953.

Matsushita, S.-Y. On the convergence rate of the Krasnosel'skiĭ–Mann iteration. *Bulletin of the Australian Mathematical Society*, 96(1):162–170, 2017.

Minty, G. J. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29(3):341–346, 1962.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *International Conference on Artificial Intelligence and Statistics*, 2020.

Nemirovski, A. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Nesterov, Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Nesterov, Y. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

Ouyang, Y. and Xu, Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

Park, C. and Ryu, E. K. Optimal first-order algorithms as a function of inequalities. *arXiv:2110.11035*, 2021.

Park, J. and Ryu, E. K. Exact optimal accelerated complexity for fixed-point iterations. *International Conference on Machine Learning*, 2022.

Park, J. and Ryu, E. K. Accelerated infeasibility detection of constrained optimization and fixed-point iterations. *International Conference on Machine Learning*, 2023.

Popov, L. D. A modification of the Arrow–Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. *Conference on Learning Theory*, 2013.

Reich, S., Thong, D. V., Cholamjiak, P., and Van Long, L. Inertial projection-type methods for solving pseudomonotone variational inequality problems in Hilbert space. *Numerical Algorithms*, 88(2):813–835, 2021.

Ryu, E. K. and Yin, W. *Large-Scale Convex Optimization via Monotone Operators*. Cambridge University Press, 2022.

Ryu, E. K., Yuan, K., and Yin, W. ODE analysis of stochastic gradient methods with optimism and anchoring for minimax problems and GANs. *arXiv:1905.10899*, 2019.

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

Sabach, S. and Shtern, S. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

Shehu, Y. Convergence rate analysis of inertial Krasnoselskii–Mann type iteration with applications. *Numerical Functional Analysis and Optimization*, 39(10):1077–1091, 2018.

Solodov, M. V. and Svaiter, B. F. A hybrid approximate extragradient – proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4):323–345, 1999.

Su, W., Boyd, S., and Candès, E. J. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Neural Information Processing Systems*, 2014.

Suh, J. J., Roh, G., and Ryu, E. K. Continuous-time analysis of AGM via conservation laws in dilated coordinate systems. *International Conference on Machine Learning*, 2022.

Suh, J. J., Park, J., and Ryu, E. K. Continuous-time analysis of anchor acceleration. *Neural Information Processing Systems*, 2023.

Taylor, A. and Bach, F. Stochastic first-order methods: Non-asymptotic and computer-aided analyses via potential functions. *Conference on Learning Theory*, 2019.

Taylor, A. and Drori, Y. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, 2022.

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

Tran-Dinh, Q. and Luo, Y. Halpern-type accelerated and splitting algorithms for monotone inclusions. *arXiv:2110.08150*, 2021.

Walker, H. F. and Ni, P. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

Wittmann, R. Approximation of fixed points of nonexpansive mappings. *Archiv der Mathematik*, 58(5):486–491, 1992.

Xu, H.-K. Iterative algorithms for nonlinear operators. *Journal of the London Mathematical Society*, 66(1):240–256, 2002.

Yoon, T. and Ryu, E. K. Accelerated minimax algorithms flock together. *SIAM Journal on Optimization (To appear)*.

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

Zhang, J., O'Donoghue, B., and Boyd, S. Globally convergent type-I Anderson acceleration for nonsmooth fixed-point iterations. *SIAM Journal on Optimization*, 30(4): 3170–3197, 2020.

A. Related work

Fixed-point algorithms. Iterative algorithms for solving fixed-point problems (1) have been extensively studied over a long period, and among them, Picard iteration, Krasnosel'skii–Mann (KM) Iteration and Halpern Iteration stand out as representative classes, each defined by:

$$y_{k+1} = \mathbb{T}y_k, \tag{Picard}$$
$$y_{k+1} = \lambda_{k+1}y_k + (1-\lambda_{k+1})\mathbb{T}y_k, \tag{KM}$$
$$y_{k+1} = \lambda_{k+1}y_0 + (1-\lambda_{k+1})\mathbb{T}y_k \tag{Halpern}$$

The formal study of Picard iteration dates up to Banach (1922). The class KM generalizes the works of Krasnosel'skii (1955) and Mann (1953); early works (Ishikawa, 1976; Borwein et al., 1992) focused on its asymptotic convergence $\|y_k - \mathbb{T}y_k\|^2 \to 0$, while quantitative rates of $\mathcal{O}(1/k)$ to $o(1/k)$ were respectively demonstrated by Cominetti et al. (2014); Liang et al. (2016); Bravo & Cominetti (2018) and Baillon & Bruck (1992); Davis & Yin (2016); Matsushita (2017). For the Halpern class, devised by Halpern (1967), Wittmann (1992); Xu (2002) established asymptotic convergence. Concerning quantitative convergence, (Leustean, 2007) provided the rate $\|y_k - \mathbb{T}y_k\|^2 = \mathcal{O}(1/(\log k)^2)$, which was later improved to $\mathcal{O}(1/k)$ (Kohlenbach, 2011), and to $\mathcal{O}(1/k^2)$, first by Sabach & Shtern (2017), and then by Lieder (2021) with a tighter constant, using the choice $\lambda_k = \frac{1}{k+1}$. Park & Ryu (2022) established the exact optimality of the convergence rate from Lieder (2021) (and the independently discovered equivalent result of Kim (2021)) by constructing a matching complexity lower bound. Some notable acceleration results for fixed-point problems not covered by the above list include Anderson-type acceleration (Anderson, 1965; Walker & Ni, 2011; Zhang et al., 2020) and inertial-type acceleration (Maingé, 2008; Dong et al., 2018; Shehu, 2018; Reich et al., 2021).

Minimax optimization algorithms. Smooth convex-concave minimax optimization is a classical problem, whose investigation dates up to Korpelevich (1976); Popov (1980), which respectively first developed the Extragradient (EG) and optimistic gradient (OG) whose variants have been studied extensively (Solodov & Svaiter, 1999; Nemirovski, 2004; Nesterov, 2007; Rakhlin & Sridharan, 2013; Daskalakis et al., 2018) across the optimization and machine learning communities. Recently there has been rapid development in the theory of reducing $\|\nabla \mathbf{L}(\cdot)\|^2$; while EG and OG converge at the rate of $\|\nabla \mathbf{L}(u_k, v_k)\|^2 = \mathcal{O}(1/k)$ (Gorbunov et al., 2022; Cai et al., 2022), using the anchoring mechanism (Ryu et al., 2019) motivated by the optimal Halpern iteration (Halpern, 1967; Kim, 2021; Lieder, 2021), Diakonikolas (2020) achieved the near-optimal rate $\tilde{\mathcal{O}}(1/k^2)$, which was finally accelerated to $\mathcal{O}(1/k^2)$ via Extra Anchored Gradient (EAG) algorithm (Yoon & Ryu, 2021). Lee & Kim (2021) provided a constant factor improvement over EAG while generalizing the acceleration to weakly nonconvex-nonconcave problems, and Tran-Dinh & Luo (2021); Cai & Zheng (2023) presented single-call versions of the acceleration. Boţ et al. (2023) achieved asymptotic $o(1/k^2)$ convergence based on the continuous-time perspective. The conceptual connection between smooth minimax optimization and fixed-point problems (proximal algorithms) have been formally studied in Mokhtari et al. (2020); Yoon & Ryu.

Continuous-time analyses. Taking the continuous-time limit of an iterative algorithm results in an ordinary differential equation (ODE), and they often more easily reveal the structure of convergence analysis. The ODE models for OG and EG were respectively studied by Csetnek et al. (2019); Lu (2022). The continuous-time analysis of accelerated algorithms was initiated by Su et al. (2014); Krichene et al. (2015), and the anchor acceleration ODE for fixed-point and minimax problems was first introduced by Ryu et al. (2019) and then generalized to a broader family of differential inclusion with more rigorous treatment by Suh et al. (2023). Boţ et al. (2023) studied a different family of ODE that achieves acceleration asymptotically. An ODE involving the coefficient of the form $\frac{1}{T-t}$ with fixed terminal time T was first presented in Suh et al. (2022), as continuous-time model of the OGM-G algorithm (Kim & Fessler, 2021). The formal concept of H-duality in continuous-time has been proposed by Kim et al. (2023a), while Kim & Yang (2023a;b) also presents a similar result.

Performance estimation problem (PEP) technique. At a high level, the PEP methodology (Drori & Teboulle, 2014; Taylor et al., 2017; Taylor & Bach, 2019; Ryu et al., 2020; Das Gupta et al., 2023) provides a computer-assisted framework for finding tight convergence proofs and optimizing the step-sizes of optimization of algorithms, and many efficient algorithms and novel proofs have recently been discovered with the aid of this methodology (Kim & Fessler, 2016; Lieder, 2021; Kim, 2021; Kim & Fessler, 2021; Park & Ryu, 2021; 2022; Gorbunov et al., 2022; Taylor & Drori, 2022; Jang et al., 2023; Park & Ryu, 2023).

Specifically relevant to our work are the prior, concurrent work of Kim (2021) presenting APPM and of Lieder (2021) presenting OHM. It was later shown that APPM and OHM are equivalent, and they represent one element within the family of exact optimal algorithms that we present in this work. We find it somewhat surprising that the two authors independently found the same algorithm when there is an infinitude of answers. Therefore, we personally asked Kim and Lieder to understand how their processes led to their discovery and not any other choice.

In a personal communication, Donghwan Kim explained that the search process in (Kim & Fessler, 2016; Kim, 2021), at a high level, involved the following steps:

1. Identify the set of inequalities needed to establish a convergence guarantee for a simpler "reference algorithm" (such as gradient descent or proximal point method).
2. Set the algorithm step-sizes as additional variables within the PEP formulation, and perform joint minimization of the convergence rate (worst-case complexity) with respect to the step-sizes. This may result in a nonconvex formulation, requiring heuristics such as alternating minimization and repetition of local minimization followed by perturbation (to escape from local optimum).
3. Iterate Step 2 while restricting the set of inequalities available within the proof (that PEP formulation uses) based on the pattern detected from Step 1. This often facilitates convergence to the global minimum and encourages numerically cleaner solutions that are conducive to analysis by humans. In the search of APPM (Kim, 2021), for instance, only the inequalities $\left\langle \tilde{\mathbf{A}}x_{i+1} - \tilde{\mathbf{A}}x_i, x_{i+1} - x_i \right\rangle \geq 0$ for $i = 1, \dots, N-1$ and $\left\langle \tilde{\mathbf{A}}x_N, x_N - y_\star \right\rangle \geq 0$ were used.

While the strategy of using a constrained set of inequalities is effective for identifying *an* optimal algorithm, it would prevent the discovery of other optimal algorithms that require distinct proof structures.

In another personal communication, Felix Lieder explained that the search process of OHM in Lieder (2021) involved the optimization of step-sizes as variables within the PEP formulation (as in the Step 2 in the case of APPM), but unlike (Kim, 2021), the entire set of available inequalities was used (without restriction). Local nonlinear programming solvers were employed to tackle this nonconvex optimization, and within multiple trials with random initialization, the step-sizes often converged near the step-size values for OHM (as detailed in (Lieder, 2018)). In this case, despite the numerical indication that optimal step-sizes were not unique, the authors' main goal was to identify the simplest, hence most easily interpretable solution, which turned out to be OHM in the end.

In our work, we take a mixture of deductive and numerical approaches. Our initial motivation to investigate the H-dual of OHM (namely Dual-OHM) came from the convex H-duality theory (Kim et al., 2023a), and PEP was mainly utilized for quick search of proofs (i.e. the correct way of combining the inequalities) and for numerical verification of the convergence guarantees.

Non-uniqueness of exact optimal algorithms. In the setup of minimizing a non-smooth convex function f whose subgradient magnitude is bounded by M (so the objective function is M-Lipschitz continuous), there are at least 4 different algorithms achieving the exact optimal complexity. Suppose that $\|x_0 - x_\star\| \leq R$, where $x_\star$ is a minimizer of f. The subgradient method of (Nesterov, 2004), given by $x_{k+1} = x_k - \frac{R}{\sqrt{N+1}} \frac{g_k}{\|g_k\|}$ where $g_k \in \partial f(x_k)$, exhibits the rate $f(x_N) - f_\star \leq \frac{MR}{\sqrt{N+1}}$ where $f_\star = f(x_\star)$. The exact same rate is achieved by the algorithm of (Drori & Teboulle, 2016) (which is a variant of the cutting-plane method), a fixed step-size algorithm of (Drori & Taylor, 2020) (without gradient normalization), and its line-search variant. As there is a matching lower bound $f(x_N) - f_\star \geq \frac{MR}{\sqrt{N+1}}$ (Drori & Teboulle, 2016), all these methods are exactly optimal in terms of the worst-case complexity.

In the fixed-point setup (or the equivalent monotone inclusion setup), a Halpern-type algorithm by Suh et al. (2023) that uses adaptive interpolation (anchoring) coefficients also achieves the exact optimal complexity (same rate as OHM). Hence, in a strict sense, our work is not the first discovery of the non-uniqueness of exact optimal algorithms. However, to the best of our knowledge, our work is the first to identify distinct *fixed step-size* algorithms sharing the same exact optimal complexity for fixed-point problems. Additionally, while the adaptive algorithm of (Suh et al., 2023) relies on a mechanism that is similar to OHM, our dual algorithms use an arguably different mechanism from that underlying OHM.

B. Equivalence of algorithms

In this section, we show equivalences between distinct forms of algorithms.

B.1. OHM

Form 1 (Anchoring form).

$$y_{k+1} = \frac{k+1}{k+2}\mathbb{T}y_k + \frac{1}{k+2}y_0 \tag{9}$$

Form 2 (Momentum form).

$$y_{k+1} = y_k - \frac{1}{k+2}(y_k - \mathbb{T}y_k) + \frac{k}{k+2}(\mathbb{T}y_k - \mathbb{T}y_{k-1}) \tag{10}$$

where $\mathbb{T}y_{-1} = x_0$.

Form 3 (H-matrix form).

$$y_{k+1} = y_k - \sum_{j=0}^{k} h_{k+1,j+1}(y_j - \mathbb{T}y_j)$$

where

$$h_{k,j} = \begin{cases} -\frac{j}{k(k+1)} & \text{if } j < k, \\ \frac{k}{k+1} & \text{if } j = k. \end{cases} \tag{11}$$

Form 4 (APPM).

$$\begin{aligned} x_{k+1} &= \mathbb{J}_{\mathbb{A}}(y_k) \\ y_{k+1} &= x_{k+1} + \frac{k}{k+2}(x_{k+1} - x_k) - \frac{k}{k+2}(x_k - y_{k-1}) \end{aligned} \tag{12}$$

where $y_{-1} = x_0 = y_0$ and $\mathbb{A} = 2(\mathbb{T}+\mathbb{I})^{-1} - \mathbb{I} \iff \mathbb{T} = 2\mathbb{J}_{\mathbb{A}} - \mathbb{I}$.

Form 1 $\implies$ Form 2. Multiplying $k+2$ to (9) gives

$$(k+2)y_{k+1} = (k+1)\mathbb{T}y_k + y_0.$$

Thus $y_0 = (k+2)y_{k+1} - (k+1)\mathbb{T}y_k$. Substituting this into $y_0 = (k+1)y_k - k\mathbb{T}y_{k-1}$ and rearranging:

$$0 = (k+2)y_{k+1} - (k+1)\mathbb{T}y_k - (k+1)y_k + k\mathbb{T}y_{k-1} \iff \text{[Form 2]}.$$

Form 2 $\iff$ Form 4. Direct substitution $x_{k+1} = \frac{1}{2}(y_k + \mathbb{T}y_k)$ (which follows from $\mathbb{J}_{\mathbb{A}} = \frac{1}{2}(\mathbb{I}+\mathbb{T})$) shows that (10) and (12) are the identical update rules.

Form 4 $\implies$ Form 1. Multiplying $k+2$ throughout the second line of (12) gives

$$\begin{aligned} &(k+2)y_{k+1} = (2k+2)x_{k+1} - 2kx_k + ky_{k-1} \\ \iff &(k+2)y_{k+1} - ky_{k-1} = (2k+2)x_{k+1} - 2kx_k. \end{aligned}$$

Summing up the last line, we see that the terms telescope:

$$\begin{aligned} &\sum_{j=0}^{k}[(j+2)y_{j+1} - jy_{j-1}] = \sum_{j=0}^{k}[(2j+2)x_{j+1} - 2jx_j] \\ \iff &(k+2)y_{k+1} + (k+1)y_k - y_0 = (2k+2)x_{k+1} \\ \iff &(k+2)y_{k+1} = y_0 + (k+1)(2x_{k+1} - y_k) = y_0 + (k+1)(2\mathbb{J}_{\mathbb{A}}y_k - y_k) \\ \iff &y_{k+1} = \frac{1}{k+2}y_0 + \frac{k+1}{k+2}(2\mathbb{J}_{\mathbb{A}} - \mathbb{I})(y_k) = \frac{k+1}{k+2}\mathbb{T}y_k + \frac{1}{k+2}y_0. \end{aligned}$$

The last line is precisely (9).

Form 1 $\implies$ Form 3. Observe that $y_1 = \frac{1}{2}\mathbb{T}y_0 + \frac{1}{2}y_0 = y_0 - \frac{1}{2}(y_0 - \mathbb{T}y_0)$, so we have $h_{1,1} = \frac{1}{2}$, which is agrees with (11). We use induction on k: Let $k \geq 1$ and suppose that (9) satisfies the representation (11) up to $y_0, \ldots, y_k$. Then

$$\begin{aligned} y_{k+1} &= \frac{k+1}{k+2}\mathbb{T}y_k + \frac{1}{k+2}y_0 \\ &= y_k - \frac{k+1}{k+2}(y_k - \mathbb{T}y_k) + \frac{1}{k+2}(y_0 - y_k), \end{aligned} \tag{13}$$

and

$$\begin{aligned} y_0 - y_k &= \sum_{i=1}^{k}(y_{i-1} - y_i) \\ &= \sum_{i=1}^{k}\sum_{j=0}^{i-1} h_{i,j+1}(y_j - \mathbb{T}y_j) \\ &= \sum_{i=1}^{k}\left(\sum_{j=0}^{i-2} -\frac{j+1}{i(i+1)}(y_j - \mathbb{T}y_j) + \frac{i}{i+1}(y_{i-1} - \mathbb{T}y_{i-1})\right) \\ &= \sum_{i=0}^{k-1}\left(\sum_{j=0}^{i-1} -\frac{j+1}{(i+1)(i+2)}(y_j - \mathbb{T}y_j)\right) + \sum_{j=0}^{k-1}\frac{j+1}{j+2}(y_j - \mathbb{T}y_j) \\ &= \frac{k}{k+1}(y_{k-1} - \mathbb{T}y_{k-1}) + \sum_{j=0}^{k-2}\left(\frac{j+1}{j+2} - \sum_{i=j+1}^{k-1}\frac{j+1}{(i+1)(i+2)}\right)(y_j - \mathbb{T}y_j) \end{aligned} \tag{14}$$

where we change the order of double summation to obtain the last identity. Applying the following formula

$$\sum_{i=j+1}^{k-1}\frac{j+1}{(i+1)(i+2)} = (j+1)\sum_{i=j+1}^{k-1}\left(\frac{1}{i+1} - \frac{1}{i+2}\right) = (j+1)\left(\frac{1}{j+2} - \frac{1}{k+1}\right)$$

to (14) gives

$$\begin{aligned} y_0 - y_k &= \frac{k}{k+1}(y_{k-1} - \mathbb{T}y_{k-1}) + \sum_{j=0}^{k-2}\frac{j+1}{k+1}(y_j - \mathbb{T}y_j) \\ &= \sum_{j=0}^{k-1}\frac{j+1}{k+1}(y_j - \mathbb{T}y_j). \end{aligned}$$

Finally, plugging the last expression into (13), we obtain

$$y_{k+1} = y_k - \frac{k+1}{k+2}(y_k - \mathbb{T}y_k) + \sum_{j=0}^{k-1}\frac{j+1}{(k+1)(k+2)}(y_j - \mathbb{T}y_j),$$

so $h_{k+1,j+1} = -\frac{j+1}{(k+1)(k+2)}$ for $j = 0, \ldots, k-1$ and $h_{k+1,k+1} = \frac{k+1}{k+2}$, completing the induction.

Form 3 $\implies$ Form 4. We have

$$y_1 = \frac{1}{2}y_0 + \frac{1}{2}\mathbb{T}y_0 = \frac{1}{2}y_0 + \frac{1}{2}(2x_1 - y_0) = x_1,$$

which agrees with (12) with $k = 0$.

Now let $k \geq 1$. Let $\tilde{\mathbf{A}}x_{j+1} = y_j - x_{j+1}$ for $j = 0, 1, \ldots$. Then $y_j - \mathbb{T}y_j = y_j - (2x_{j+1} - y_j) = 2\tilde{\mathbf{A}}x_{j+1}$, and

$$\begin{aligned} y_{k+1} &= y_k - \sum_{j=0}^{k} 2h_{k+1,j+1}\tilde{\mathbf{A}}x_{j+1} \\ &= y_k + \sum_{j=0}^{k-1} \frac{2(j+1)}{(k+1)(k+2)}\tilde{\mathbf{A}}x_{j+1} - \frac{2(k+1)}{k+2}\tilde{\mathbf{A}}x_{k+1}, \end{aligned} \tag{15}$$

and

$$\begin{aligned} x_{k+1} - x_k &= (y_k - \tilde{\mathbf{A}}x_{k+1}) - (y_{k-1} - \tilde{\mathbf{A}}x_k) \\ &= y_k - y_{k-1} - \tilde{\mathbf{A}}x_{k+1} + \tilde{\mathbf{A}}x_k \\ &= -\sum_{j=0}^{k-1} 2h_{k,j+1}\tilde{\mathbf{A}}x_{j+1} - \tilde{\mathbf{A}}x_{k+1} + \tilde{\mathbf{A}}x_k \\ &= \sum_{j=0}^{k-2} \frac{2(j+1)}{k(k+1)}\tilde{\mathbf{A}}x_{j+1} + \left(1 - \frac{2k}{k+1}\right)\tilde{\mathbf{A}}x_k - \tilde{\mathbf{A}}x_{k+1} \\ &= \sum_{j=0}^{k-2} \frac{2(j+1)}{k(k+1)}\tilde{\mathbf{A}}x_{j+1} - \frac{k-1}{k+1}\tilde{\mathbf{A}}x_k - \tilde{\mathbf{A}}x_{k+1}. \end{aligned}$$

Thus,

$$\begin{aligned} &x_{k+1} + \frac{k}{k+2}(x_{k+1} - x_k) - \frac{k}{k+2}(x_k - y_{k-1}) \\ &= y_k - \tilde{\mathbf{A}}x_{k+1} + \frac{k}{k+2}\left(\sum_{j=0}^{k-2} \frac{2(j+1)}{k(k+1)}\tilde{\mathbf{A}}x_{j+1} - \frac{k-1}{k+1}\tilde{\mathbf{A}}x_k - \tilde{\mathbf{A}}x_{k+1}\right) + \frac{k}{k+2}\tilde{\mathbf{A}}x_k \\ &= y_k + \sum_{j=0}^{k-2} \frac{2(j+1)}{(k+1)(k+2)}\tilde{\mathbf{A}}x_{j+1} + \left(\frac{k}{k+2} - \frac{k(k-1)}{(k+1)(k+2)}\right)\tilde{\mathbf{A}}x_k - \frac{2k+2}{k+2}\tilde{\mathbf{A}}x_{k+1} \\ &= y_k + \sum_{j=0}^{k-1} \frac{2(j+1)}{(k+1)(k+2)}\tilde{\mathbf{A}}x_{j+1} - \frac{2(k+1)}{k+2}\tilde{\mathbf{A}}x_{k+1} \\ &= y_{k+1} \end{aligned}$$

and the proof is complete (the last line follows from (15)).

B.2. Dual-OHM

Form 1 (Momentum form).

$$y_{k+1} = y_k + \frac{N-k-1}{N-k}\left(\mathbb{T}y_k - \mathbb{T}y_{k-1}\right)$$

for $k = 0, 1, \ldots, N-2$, where $\mathbb{T}y_{-1} = y_0$.

Form 2 (z-form).

$$\begin{aligned} z_{k+1} &= \frac{N-k-1}{N-k}z_k - \frac{1}{N-k}\left(y_k - \mathbb{T}y_k\right) \\ y_{k+1} &= \mathbb{T}y_k - z_{k+1} \end{aligned} \tag{16}$$

for $k = 0, 1, \ldots, N-2$, where $z_0 = 0$.

Form 3 (H-matrix form).

$$\begin{aligned} y_{k+1} &= y_k - \sum_{j=0}^{k} h_{k+1,j+1}(y_j - \mathbb{T}y_j) \\ &= y_k - \sum_{j=0}^{k} 2h_{k+1,j+1}\tilde{\mathbb{A}}x_{j+1} \end{aligned}$$

for $k = 0, 1, \ldots, N-2$, where

$$h_{k,j} = \begin{cases} -\frac{N-k}{(N-j)(N-j+1)} & \text{if } j < k, \\ \frac{N-k}{N-k+1} & \text{if } j = k. \end{cases} \tag{17}$$

Form 4 (Proximal form of Section 3.2).

$$\begin{aligned} x_{k+1} &= \mathbb{J}_{\mathbb{A}}(y_k) \\ y_{k+1} &= x_{k+1} + \frac{N-k-1}{N-k}(x_{k+1} - x_k) - \frac{N-k-1}{N-k}(x_k - y_{k-1}) - \frac{1}{N-k}(x_{k+1} - y_k) \end{aligned} \tag{18}$$

for $k = 0, 1, \ldots, N-2$, where $x_0 = y_0$ and $\mathbb{A} = 2(\mathbb{T} + \mathbb{I})^{-1} - \mathbb{I} \iff \mathbb{T} = 2\mathbb{J}_{\mathbb{A}} - \mathbb{I}$.

Form 1 $\implies$ Form 2. Let $z_0 = 0$ and $z_k = \mathbb{T}y_{k-1} - y_k$ for $k = 1, \ldots, N-1$, so that the second line of (16) holds by definition. Following establishes the recursion for z_k, which is the first line of (16):

$$\begin{aligned} z_{k+1} &= \mathbb{T}y_k - y_{k+1} \\ &= \mathbb{T}y_k - \left(y_k + \frac{N-k-1}{N-k}(\mathbb{T}y_k - \mathbb{T}y_{k-1}) \right) \\ &= \mathbb{T}y_k - \left(y_k + \frac{N-k-1}{N-k}(\mathbb{T}y_k - z_k - y_k) \right) \\ &= \frac{N-k-1}{N-k} z_k - \frac{1}{N-k}\left(y_k - \mathbb{T}y_k\right). \end{aligned}$$

Form 2 $\implies$ Form 3. The first line of (16) is equivalent to

$$z_{j+1} = \frac{N-j-1}{N-j} z_j - \frac{2}{N-j}\tilde{\mathbb{A}}x_{j+1}.$$

Dividing both sides by $N-j-1$ and rearranging, we obtain

$$\frac{1}{N-j-1} z_{j+1} - \frac{1}{N-j} z_j = -\frac{2}{(N-j)(N-j-1)}\tilde{\mathbb{A}}x_{j+1}.$$

Summing this up from $j = 0$ to k and multiplying $N-k-1$ to the both sides we have

$$z_{k+1} = -(N-k-1)\sum_{j=0}^{k} \frac{2}{(N-j)(N-j-1)}\tilde{\mathbb{A}}x_{j+1}$$

(note that $z_0 = 0$). Next, we substitute the last expression into the second line of (16):

$$\begin{aligned} y_{k+1} &= y_k + (\mathbb{T}y_k - y_k) - z_{k+1} \\ &= y_k - 2\tilde{\mathbb{A}}x_{k+1} + (N-k-1)\sum_{j=0}^{k} \frac{2}{(N-j)(N-j-1)}\tilde{\mathbb{A}}x_{j+1} \\ &= y_k - \frac{2(N-k-1)}{N-k}\tilde{\mathbb{A}}x_{k+1} + \sum_{j=1}^{k} \frac{2(N-k-1)}{(N-j+1)(N-j)}\tilde{\mathbb{A}}x_j. \end{aligned}$$

This shows that $h_{k+1,j} = -\frac{(N-k-1)}{(N-j)(N-j+1)}$ for $j = 1, \ldots, k$ and $h_{k+1,k+1} = \frac{N-k-1}{N-k}$, which agrees with (17).

Form 3 $\implies$ Form 4. From the definition of Form 3,

$$y_{k+1} = y_k - \frac{2(N-k-1)}{N-k}\tilde{\mathbb{A}}x_{k+1} + \sum_{j=0}^{k-1}\frac{2(N-k-1)}{(N-j-1)(N-j)}\tilde{\mathbb{A}}x_{j+1}. \tag{19}$$

Putting $k-1$ in place of k, we have

$$\begin{aligned} y_k &= y_{k-1} - \frac{2(N-k)}{N-k+1}\tilde{\mathbb{A}}x_k + \sum_{j=0}^{k-2}\frac{2(N-k)}{(N-j-1)(N-j)}\tilde{\mathbb{A}}x_{j+1} \\ &= y_{k-1} - 2\tilde{\mathbb{A}}x_k + \sum_{j=0}^{k-1}\frac{2(N-k)}{(N-j-1)(N-j)}\tilde{\mathbb{A}}x_{j+1}. \end{aligned}$$

Using the last equation, we replace the summation within (19):

$$y_{k+1} = y_k - \frac{2(N-k-1)}{N-k}\tilde{\mathbb{A}}x_{k+1} + \frac{N-k-1}{N-k}\left(y_k - y_{k-1} + 2\tilde{\mathbb{A}}x_k\right).$$

Finally, substitute $\tilde{\mathbb{A}}x_{k+1} = y_k - x_{k+1}$, $\tilde{\mathbb{A}}x_k = y_{k-1} - x_k$ and rearrange to obtain Form 4:

$$\begin{aligned} y_{k+1} &= y_k - \frac{2(N-k-1)}{N-k}\left(y_k - x_{k+1}\right) + \frac{N-k-1}{N-k}\left(y_k - y_{k-1} + 2(y_{k-1} - x_k)\right) \\ &= x_{k+1} + \frac{N-k-1}{N-k}(x_{k+1} - x_k) - \frac{N-k-1}{N-k}(x_k - y_{k-1}) - \frac{1}{N-k}(x_{k+1} - y_k). \end{aligned}$$

Form 4 $\implies$ Form 1. Simply substitute $x_{k+1} = \frac{1}{2}(y_k + \mathbb{T}y_k)$ into the second line of (18) and rearrange. In detail:

$$\begin{aligned} y_{k+1} &= x_{k+1} + \frac{N-k-1}{N-k}(x_{k+1} - x_k) - \frac{N-k-1}{N-k}(x_k - y_{k-1}) - \frac{1}{N-k}(x_{k+1} - y_k) \\ &= \frac{2(N-k-1)}{N-k}x_{k+1} - \frac{2(N-k-1)}{N-k}x_k + \frac{N-k-1}{N-k}y_{k-1} + \frac{1}{N-k}y_k \\ &= \frac{N-k-1}{N-k}\left(y_k + \mathbb{T}y_k\right) - \frac{N-k-1}{N-k}\left(y_{k-1} + \mathbb{T}y_{k-1}\right) + \frac{N-k-1}{N-k}y_{k-1} + \frac{1}{N-k}y_k \\ &= y_k + \frac{N-k-1}{N-k}\left(\mathbb{T}y_k - \mathbb{T}y_{k-1}\right). \end{aligned}$$

B.3. FEG

Form 1 (Original form).

$$\begin{aligned} x_{k+1/2} &= x_k + \frac{1}{k+1}(x_0 - x_k) - \frac{k}{k+1}\alpha\nabla_{\pm}\mathbf{L}(x_k) \\ x_{k+1} &= x_k + \frac{1}{k+1}(x_0 - x_k) - \alpha\nabla_{\pm}\mathbf{L}(x_{k+1/2}) \end{aligned} \tag{20}$$

Form 2 (H-matrix form).

$$x_{(\ell+1)/2} = x_{\ell/2} - \frac{1}{L}\sum_{i=0}^{\ell} h_{(\ell+1)/2,i/2}\nabla_{\pm}\mathbf{L}(x_{i/2})$$

where for $k = 0, 1, \ldots, N-1$,

$$\frac{1}{\alpha L}h_{(\ell+1)/2,i/2} = \begin{cases} \frac{k}{k+1} & \text{if } \ell = 2k,\ i = 2k \\ -\frac{j+1}{k(k+1)} & \text{if } \ell = 2k,\ i = 2j+1,\ j = 0, \ldots, k-1 \\ 0 & \text{if } \ell = 2k,\ i = 2j,\ j = 0, \ldots, k-1 \\ 1 & \text{if } \ell = 2k+1,\ i = 2k+1 \\ -\frac{k}{k+1} & \text{if } \ell = 2k+1,\ i = 2k \\ 0 & \text{if } \ell = 2k+1,\ i = 0, \ldots, 2k-1. \end{cases} \tag{21}$$

It suffices to inductively check that the two forms indicate the same update rule of generating $x_{k+1/2}, x_{k+1}$ provided that they are equivalent for all $x_0, \dots, x_k$. They are clearly equivalent for $k = 0$. Assume that the equivalence holds for $j = 0, \dots, k$. Now we show that the update rule by Form 1 agrees with (21). Multiplying $k+1$ to the second line of (20) and switching the index from k to j, we have

$$(j+1)x_{j+1} = jx_j + x_0 - \alpha(j+1)\nabla_{\pm}\mathbf{L}(x_{j+1/2}).$$

Summing this up from $j = 0$ to k and dividing by $k+1$ we have

$$x_{k+1} - x_0 = -\frac{1}{k+1}\alpha\sum_{j=0}^{k}(j+1)\nabla_{\pm}\mathbf{L}(x_{i+1/2}). \tag{22}$$

Also, by subtracting the first line of (20) from the second line of (20), we obtain

$$x_{k+1} = x_{k+1/2} - \alpha\nabla_{\pm}\mathbf{L}(x_{k+1/2}) + \frac{k}{k+1}\alpha\nabla_{\pm}\mathbf{L}(x_k).$$

Then, by applying (22) with k and $k \leftarrow k-1$ we can write

$$\begin{aligned}
x_{k+1/2} - x_k &= x_{k+1} + \alpha\nabla_{\pm}\mathbf{L}(x_{k+1/2}) - \frac{k}{k+1}\alpha\nabla_{\pm}\mathbf{L}(x_k) - x_k \\
&= (x_{k+1} - x_k) + \alpha\nabla_{\pm}\mathbf{L}(x_{k+1/2}) - \frac{k}{k+1}\alpha\nabla_{\pm}\mathbf{L}(x_k) \\
&= (x_{k+1} - x_0) - (x_k - x_0) + \alpha\nabla_{\pm}\mathbf{L}(x_{k+1/2}) - \frac{k}{k+1}\alpha\nabla_{\pm}\mathbf{L}(x_k) \\
&= \left(-\frac{1}{k+1}\alpha\sum_{j=0}^{k}(j+1)\nabla_{\pm}\mathbf{L}(x_{i+1/2})\right) - \left(-\frac{1}{k}\alpha\sum_{j=0}^{k-1}(j+1)\nabla_{\pm}\mathbf{L}(x_{i+1/2})\right) \\
&\quad + \alpha\nabla_{\pm}\mathbf{L}(x_{k+1/2}) - \frac{k}{k+1}\alpha\nabla_{\pm}\mathbf{L}(x_k) \\
&= -\frac{k}{k+1}\alpha\nabla_{\pm}\mathbf{L}(x_k) + \sum_{j=0}^{k-1}\left(\frac{j+1}{k} - \frac{j+1}{k+1}\right)\alpha\nabla_{\pm}\mathbf{L}(x_{j+1/2}) \\
&= -\frac{k}{k+1}\alpha\nabla_{\pm}\mathbf{L}(x_k) + \sum_{j=0}^{k-1}\frac{j+1}{k(k+1)}\alpha\nabla_{\pm}\mathbf{L}(x_{j+1/2}).
\end{aligned}$$

This shows that

$$h_{k+1/2,k} = \frac{k}{k+1}\alpha L, \quad h_{k+1/2,j+1/2} = -\frac{j+1}{k(k+1)}\alpha L, \quad h_{k+1/2,j} = 0 \quad (j = 0, \dots, k-1).$$

Next, because

$$x_{k+1} = x_{k+1/2} - \alpha\nabla_{\pm}\mathbf{L}(x_{k+1/2}) + \frac{k}{k+1}\alpha\nabla_{\pm}\mathbf{L}(x_k)$$

we obtain

$$h_{k+1,k+1/2} = \alpha L, \quad h_{k+1,k} = -\frac{k}{k+1}\alpha L, \quad h_{k+1,i/2} = 0 \quad (i = 0, \dots, 2k-1)$$

which agrees with (21).

B.4. Dual-FEG

Form 1 (Original form).

$$\begin{aligned}
x_{k+1/2} &= x_k - \alpha z_k - \alpha\nabla_{\pm}\mathbf{L}(x_k) \\
x_{k+1} &= x_{k+1/2} - \frac{N-k-1}{N-k}\alpha\left(\nabla_{\pm}\mathbf{L}(x_{k+1/2}) - \nabla_{\pm}\mathbf{L}(x_k)\right) \\
z_{k+1} &= \frac{N-k-1}{N-k}z_k - \frac{1}{N-k}\nabla_{\pm}\mathbf{L}(x_{k+1/2})
\end{aligned} \tag{23}$$

for $k = 0, 1, \ldots, N-1$, where $z_0 = 0$.

Form 2 (H-matrix form).

$$x_{(\ell+1)/2} = x_{\ell/2} - \frac{1}{L}\sum_{i=0}^{\ell} h_{(\ell+1)/2,i/2}\nabla_{\pm}\mathbf{L}(x_{i/2})$$

where

$$\frac{1}{\alpha L}h_{(\ell+1)/2,i/2} = \begin{cases} 1 & \text{if } \ell = 2k,\ i = 2k \\ -\frac{N-k}{(N-j-1)(N-j)} & \text{if } \ell = 2k,\ i = 2j+1,\ j = 0,\ldots,k-1 \\ 0 & \text{if } \ell = 2k,\ i = 2j,\ j = 0,\ldots,k-1 \\ \frac{N-k-1}{N-k} & \text{if } \ell = 2k+1,\ i = 2k+1 \\ -\frac{N-k-1}{N-k} & \text{if } \ell = 2k+1,\ i = 2k \\ 0 & \text{if } \ell = 2k+1,\ i = 0,\ldots,2k-1. \end{cases} \tag{24}$$

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

As in the case of FEG, we check that update rule (23) of Form 1 defines the identical update rule for $x_{k+1/2}, x_{k+1}$ provided that they are equivalent for all $x_0, \ldots, x_k$. Dividing the the third line of (23) by $N-k$ and switching the index from k to j we obtain

$$\frac{1}{N-j}z_j = \frac{1}{N-j+1}z_{j-1} - \frac{1}{(N-j)(N-j+1)}\nabla_{\pm}\mathbf{L}(x_{j-1/2}).$$

Summing this up from $j = 0$ to $k-1$ and multiplying $N-k$ to both sides gives

$$z_k = -\sum_{j=0}^{k-1}\frac{N-k}{(N-j-1)(N-j)}\nabla_{\pm}\mathbf{L}(x_{j+1/2})$$

(note that $z_0 = 0$). Now substituting the above expression for z_k into the first line of (23) gives

$$\begin{aligned} x_{k+1/2} &= x_k - \alpha z_k - \alpha\nabla_{\pm}\mathbf{L}(x_k) \\ &= x_k - \alpha\nabla_{\pm}\mathbf{L}(x_k) + \sum_{j=0}^{k-1}\frac{N-k}{(N-j-1)(N-j)}\alpha\nabla_{\pm}\mathbf{L}(x_{j+1/2}) \end{aligned}$$

and thus,

$$h_{k+1/2,k} = \alpha L, \quad h_{k+1/2,j+1/2} = -\frac{N-k}{(N-j-1)(N-j)}\alpha L, \quad h_{k+1/2,j} = 0 \quad (j = 0,\ldots,k-1).$$

Finally, the second line of (23) is

$$x_{k+1} = x_{k+1/2} - \frac{N-k-1}{N-k}\alpha\nabla_{\pm}\mathbf{L}(x_{k+1/2}) + \frac{N-k-1}{N-k}\alpha\nabla_{\pm}\mathbf{L}(x_k)$$

which gives

$$h_{k+1,k+1/2} = \frac{N-k-1}{N-k}\alpha L, \quad h_{k+1,k} = -\frac{N-k-1}{N-k}\alpha L, \quad h_{k+1,i/2} = 0 \quad (i = 0,\ldots,2k-1).$$

This is precisely (24).

B.5. Anchor ODE (4)

Form 1 (Original form).

$$\dot{X}(t) = -\mathbb{A}(X(t)) - \frac{1}{t}(X(t) - X_0) \tag{25}$$

where $X(0) = X_0$.

Form 2 (Second-order form).

$$\ddot{X}(t) + \frac{2}{t}\dot{X}(t) + \frac{1}{t}\mathbb{A}(X(t)) + \frac{d}{dt}\mathbb{A}(X(t)) = 0 \tag{26}$$

where $X(0) = X_0$ and $\dot{X}(0) = -A(X_0)$.

Form 1 $\implies$ Form 2. Let $X\colon [0,\infty) \to \mathbb{R}^d$ be the solution to (25) with initial condition $X(0) = X_0$. First observe that taking the limit $t \to 0^+$ in (25) gives $\dot{X}(0) = -\mathbb{A}(X_0) - \dot{X}(0)$, which is equivalent to the initial velocity condition $\dot{X}(0) = -\frac{1}{2}\mathbb{A}(X_0)$ for (26). Differentiating both sides of (25), we have

$$\ddot{X}(t) = -\frac{d}{dt}\mathbb{A}(X(t)) - \frac{1}{t}\dot{X}(t) + \frac{1}{t^2}(X(t) - X_0).$$

Rearranging the defining equation (25) gives $\frac{1}{t^2}(X(t) - X_0) = -\frac{1}{t}\dot{X}(t) - \frac{1}{t}\mathbb{A}(X(t))$. Substituting this into the last equation and reorganizing, we obtain (26):

$$\begin{aligned}\ddot{X}(t) &= -\frac{d}{dt}\mathbb{A}(X(t)) - \frac{1}{t}\dot{X}(t) + \left(-\frac{1}{t}\dot{X}(t) - \frac{1}{t}\mathbb{A}(X(t))\right)\\ &\iff \ddot{X}(t) + \frac{2}{t}\dot{X}(t) + \frac{1}{t}\mathbb{A}(X(t)) + \frac{d}{dt}\mathbb{A}(X(t)) = 0.\end{aligned}$$

Form 2 $\implies$ Form 1. Suppose $X\colon [0,\infty) \to \mathbb{R}^d$ is the solution to (26) with initial conditions $X(0) = X_0$ and $\dot{X}(0) = -\frac{1}{2}\mathbb{A}(X_0)$. Multiplying t throughout (26), we have

$$0 = t\ddot{X}(t) + 2\dot{X}(t) + \mathbb{A}(X(t)) + t\frac{d}{dt}\mathbb{A}(X(t)) = \frac{d}{dt}\left(t\dot{X}(t)\right) + \dot{X}(t) + \frac{d}{dt}\left(t\mathbb{A}(X(t))\right).$$

Integrating both sides from 0 to t gives

$$0 = t\dot{X}(t) + X(t) - X_0 + t\mathbb{A}(X(t)).$$

Dividing both sides by t and reorganizing, we get (25).

The above result holds given a minimal assumption that $\mathbb{A}$ is Lipschitz continuous (with the equalities holding for almost every t if differentiability is not assumed). For the rigorous discussion on this point, we refer the readers to (Suh et al., 2023, Appendix B).

B.6. Dual-Anchor ODE (5)

Form 1 (Original form).

$$\begin{aligned}\dot{X}(t) &= -Z(t) - \mathbb{A}(X(t))\\ \dot{Z}(t) &= -\frac{1}{T-t}Z(t) - \frac{1}{T-t}\mathbb{A}(X(t))\end{aligned} \tag{27}$$

where $X(0) = X_0$ and $Z(0) = 0$.

Form 2 (Second-order form).

$$\ddot{X}(t) + \frac{1}{T-t}\dot{X}(t) - \frac{d}{dt}\mathbb{A}(X(t)) = 0 \tag{28}$$

where $X(0) = X_0$ and $\dot{X}(0) = -\mathbb{A}(X_0)$.

Form 1 $\implies$ Form 2. Let $\begin{pmatrix} X \\ Z \end{pmatrix} : [0,T) \to \mathbb{R}^d \times \mathbb{R}^d$ be the solution to (27) with initial conditions $X(0) = X_0$ and $Z(0) = 0$. Plugging $t = 0$ into the first line of (5) gives $\dot{X}(0) = 0 - \mathbb{A}(X_0)$, which is the initial velocity condition for (28). Now observe that

$$\dot{Z}(t) = \frac{1}{T-t}\left(-Z(t) - \mathbb{A}(X(t))\right) = \frac{1}{T-t}\dot{X}(t)$$

where the last equality comes from the first line of (27). Differentiating the first line of (27) and plugging in the above identity we obtain Form 2:

$$0 = \ddot{X}(t) + \dot{Z}(t) + \frac{d}{dt}\mathbb{A}(X(t)) = \ddot{X}(t) + \frac{1}{T-t}\dot{X}(t) - \frac{d}{dt}\mathbb{A}(X(t)).$$

Form 2 $\implies$ Form 1. Suppose $X\colon [0,T) \to \mathbb{R}^d$ is the solution to (28) with initial conditions $X(0) = X_0$ and $\dot{X}(0) = -\mathbb{A}(X_0)$. Define $Z\colon [0,T) \to \mathbb{R}^d$ by $Z(t) = -\dot{X}(t) - \mathbb{A}(X(t))$. Then $\dot{X}(t) = -Z(t) - \mathbb{A}(X(t))$ by definition (this is the first line of (27)). Also note that $Z(0) = -\dot{X}(0) - \mathbb{A}(X_0) = 0$, which is the Z-initial condition for (27). Now differentiating Z we have

$$\dot{Z}(t) = -\ddot{X}(t) - \frac{d}{dt}\mathbb{A}(X(t)) = \frac{1}{T-t}\dot{X}(t) = -\frac{1}{T-t}\left(Z(t) + \mathbb{A}(X(t))\right).$$

The second equality directly follows from the defining equation (28). This shows that $\begin{pmatrix} X \\ Z \end{pmatrix}$ is the solution of (27).

The above result holds given a minimal assumption that $\mathbb{A}$ is Lipschitz continuous (with the equalities holding for almost every t if differentiability is not assumed). For the rigorous discussion on existence, uniqueness of the solutions and almost everywhere differntiability of the involved quantities, we refer the readers to Appendix H.

C. Lyapunov analysis of Dual-OHM

Recall the following form of Dual-OHM:

$$z_{k+1} = \frac{N-k-1}{N-k} z_k - \frac{1}{N-k}\left(y_k - \mathbb{T}y_k\right) \tag{29}$$

$$y_{k+1} = \mathbb{T}y_k - z_{k+1}. \tag{30}$$

Substituting (29) into (30) we get

$$\begin{aligned} y_{k+1} &= \mathbb{T}y_k - \frac{N-k-1}{N-k} z_k + \frac{1}{N-k}\left(y_k - \mathbb{T}y_k\right) \\ &= \frac{1}{N-k} y_k + \frac{N-k-1}{N-k}\mathbb{T}y_k - \frac{N-k-1}{N-k} z_k \end{aligned} \tag{31}$$

With the substitution $\mathbb{T} = 2\mathbb{J}_{\mathbb{A}} - \mathbb{I}$ and $x_{k+1} = \mathbb{J}_{\mathbb{A}} y_k$ we can write (29) and (31) as

$$z_{k+1} = \frac{N-k-1}{N-k} z_k - \frac{2}{N-k}\tilde{\mathbb{A}} x_{k+1} \tag{32}$$

$$\begin{aligned} y_{k+1} &= \frac{1}{N-k} y_k + \frac{N-k-1}{N-k}\left(y_k - 2\tilde{\mathbb{A}}x_{k+1}\right) - \frac{N-k-1}{N-k} z_k \\ &= y_k - \frac{2(N-k-1)}{N-k}\tilde{\mathbb{A}}x_{k+1} - \frac{N-k-1}{N-k} z_k \end{aligned} \tag{33}$$

where we have used $\mathbb{T}y_k = 2\mathbb{J}_{\mathbb{A}} y_k - y_k = 2x_{k+1} - y_k = y_k - 2(y_k - x_{k+1}) = y_k - 2\tilde{\mathbb{A}}x_{k+1}$. For simplicity, write $g_j = \tilde{\mathbb{A}}x_j$ for $j = 1, \ldots, N$. To complete the proof, it remains to show that for

$$V_k = \underbrace{-\frac{N-k-1}{N-k}\left\|z_k + 2g_N\right\|^2}_{:=V_k^{(1)}} + \underbrace{\frac{2}{N-k}\left\langle z_k + 2g_N, y_k - y_{N-1}\right\rangle}_{:=V_k^{(2)}}$$

the following holds:

$$V_k - V_{k+1} = \frac{4}{(N-k)(N-k-1)}\left\langle x_{k+1} - x_N, g_{k+1} - g_N\right\rangle.$$

Rewriting the right hand side, we have

$$\begin{aligned} &\frac{4}{(N-k)(N-k-1)}\left\langle x_{k+1} - x_N, g_{k+1} - g_N\right\rangle \\ &= \frac{4}{(N-k)(N-k-1)}\left\langle y_k - y_{N-1} - (g_{k+1} - g_N), g_{k+1} - g_N\right\rangle \\ &= \frac{2}{N-k-1}\left\langle y_k - y_{N-1}, \frac{2}{N-k} g_{k+1} - \frac{2}{N-k} g_N\right\rangle - \frac{4}{(N-k)(N-k-1)}\left\|g_{k+1} - g_N\right\|^2 \\ &\overset{(32)}{=} \frac{2}{N-k-1}\left\langle y_k - y_{N-1}, \frac{N-k-1}{N-k}(z_k + 2g_N) - (z_{k+1} + 2g_N)\right\rangle - \frac{4}{(N-k)(N-k-1)}\left\|g_{k+1} - g_N\right\|^2 \\ &= \frac{2}{N-k}\left\langle y_k - y_{N-1}, z_k + 2g_N\right\rangle - \frac{2}{N-k-1}\left\langle y_k - y_{N-1}, z_{k+1} + 2g_N\right\rangle - \frac{4}{(N-k)(N-k-1)}\left\|g_{k+1} - g_N\right\|^2 \\ &\overset{(33)}{=} V_k^{(2)} - \frac{2}{N-k-1}\left\langle y_{k+1} - y_{N-1} + \frac{2(N-k-1)}{N-k} g_{k+1} + \frac{N-k-1}{N-k} z_k, z_{k+1} + 2g_N\right\rangle \\ &\quad - \frac{4}{(N-k)(N-k-1)}\left\|g_{k+1} - g_N\right\|^2 \\ &= V_k^{(2)} - V_{k+1}^{(2)} \underbrace{- \frac{2}{N-k}\left\langle 2g_{k+1} + z_k, z_{k+1} + 2g_N\right\rangle - \frac{1}{(N-k)(N-k-1)}\left\|2g_{k+1} - 2g_N\right\|^2}_{:=R_k} \end{aligned}$$

and the proof is done once we show $R_k = V_k^{(1)} - V_{k+1}^{(1)}$. From (32) we have $2g_{k+1} = (N-k-1)z_k - (N-k)z_{k+1}$, and plugging this into R_k we obtain

$$\begin{aligned}
R_k &= -2\left\langle z_{k+1} - z_k, z_{k+1} + 2g_N \right\rangle - \frac{1}{(N-k)(N-k-1)} \left\| (N-k-1)z_k - (N-k)z_{k+1} - 2g_N \right\|^2 \\
&= 2\left\langle (z_{k+1} + 2g_N) - (z_k + 2g_N), z_{k+1} + 2g_N \right\rangle \\
&\quad - \frac{1}{(N-k)(N-k-1)} \left\| (N-k-1)(z_k + 2g_N) - (N-k)(z_{k+1} + 2g_N) \right\|^2 \\
&= 2\left\| z_{k+1} + 2g_N \right\|^2 - \frac{N-k-1}{N-k} \left\| z_k + 2g_N \right\|^2 - \frac{N-k}{N-k-1} \left\| z_{k+1} + 2g_N \right\|^2 \\
&= -\frac{N-k-1}{N-k} \left\| z_k + 2g_N \right\|^2 + \frac{N-k-2}{N-k-1} \left\| z_{k+1} + 2g_N \right\|^2 \\
&= V_k^{(1)} - V_{k+1}^{(1)},
\end{aligned}$$

which proves that indeed, $R_k = V_k^{(1)} - V_{k+1}^{(1)}$.

D. Fixed-point H-duality theory

D.1. The precise statement of H-duality theorem

To state the theorem, we first need to set up some notations and concepts.

Primal Lyapunov structure. Consider the H-matrix representation of an algorithm

$$y_{k+1} = y_k - \sum_{j=0}^{k} h_{k+1,j+1}(y_j - \mathbb{T}y_j) = y_k - \sum_{j=0}^{k} 2h_{k+1,j+1}\tilde{\mathbb{A}}x_{j+1} \tag{34}$$

for $k = 0, 1, \ldots, N-2$, where $x_{k+1} = \mathbb{J}_{\mathbb{A}}(y_k)$ for $k = 0, 1, \ldots, N-1$. Consider a convergence proof for (34) with respect to the performance measure $\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 = 4\left\|\tilde{\mathbb{A}}x_N\right\|^2$, structured in the following way: Take a sequence $\{u_j\}_{j=1}^{N-1}$ of positive numbers, and define the sequence $\{U_j\}_{j=1}^{N}$ by $U_1 = 0$ and

$$U_{j+1} = U_j - u_j \left\langle x_{j+1} - x_j, \tilde{\mathbb{A}}x_{j+1} - \tilde{\mathbb{A}}x_j \right\rangle$$

for $j = 1, \ldots, N-1$. As $\mathbb{A}$ is monotone, $\{U_j\}_{j=1}^{N}$ is a nonincreasing sequence. To clarify, while we restrict the proof structure to use a specific combination of monotonicity inequalities, we let $\{u_j\}_{j=1}^{N-1}$ as free variables which can be appropriately chosen to make the convergence analysis work.

Assume that we can show that for some $\tau_U > 0$,

$$\tau_U \left\|\tilde{\mathbb{A}}x_N\right\|^2 + \left\langle \tilde{\mathbb{A}}x_N, x_N - y_0 \right\rangle \le U_N \tag{C1}$$

holds for arbitrary $\tilde{\mathbb{A}}x_1, \ldots, \tilde{\mathbb{A}}x_N$. To put this precisely, (RHS) − (LHS) in (C1) is a "vector quadratic form" of $\{\tilde{\mathbb{A}}x_j\}_{j=1}^{N}$, i.e., a function $\mathcal{Q}\colon \prod_{j=1}^{N}\mathbb{R}^d \to \mathbb{R}$ of the form

$$\mathcal{Q}(g_1, \ldots, g_N) = \sum_{i=1}^{N}\sum_{j=1}^{N} s_{i,j}\langle g_i, g_j\rangle = \mathrm{Trace}\left(\mathbf{G}^{\intercal}\mathbf{S}_N\mathbf{G}\right)$$

where $\mathbf{G} = \begin{bmatrix} g_1 & \cdots & g_N \end{bmatrix}, \mathbf{S}_N = (s_{i,j})_{1\le i,j\le N} \in \mathbb{R}^{N\times N}$. Here $s_{i,j} = s_{i,j}\left(H, \{u_j\}_{j=1}^{N-1}, \tau_U\right)$ has a hidden dependency on the entries of H-matrix, u_j's and τ_U. If $\mathcal{Q}(g_1, \ldots, g_N) \ge 0$ for any $g_1, \ldots, g_N \in \mathbb{R}^d$ (which is equivalent to $\mathbf{S}_N \succeq 0$) then we informally say $\mathcal{Q} \ge 0 \iff$ (C1). If that is the case, we can establish

$$\tau_U \left\|\mathbb{A}x_N\right\|^2 + \left\langle \tilde{\mathbb{A}}x_N, x_N - y_0 \right\rangle \le U_N \le \cdots \le U_1 = 0.$$

By Lemma 3.2, this implies $\left\|\tilde{\mathbb{A}}x_N\right\|^2 \le \frac{\|y_0 - x_\star\|^2}{\tau_U^2}$. OHM is an example where this holds; selecting $\tau_U = N$ and $u_j = \frac{j(j+1)}{N}$ for $j = 1, \ldots, N-1$ ensures (C1), leading to the final convergence rate $\left\|\tilde{\mathbb{A}}x_N\right\|^2 \le \frac{\|y_0 - x_\star\|^2}{N^2}$. We refer to this proof strategy as *primal Lyapunov proof*.

Dual Lyapunov structure. Consider a convergence proof for (34) with respect to the same performance measure $\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 = 4\left\|\tilde{\mathbb{A}}x_N\right\|^2$, but now using a positive sequence $\{v_j\}_{j=1}^{N-1}$ and $\{V_j\}_{j=0}^{N-1}$, defined by $V_{N-1} = 0$ and the backward recursion

$$V_j = V_{j+1} + v_{j+1}\left\langle x_N - x_{j+1}, \tilde{\mathbb{A}}x_N - \tilde{\mathbb{A}}x_{j+1}\right\rangle$$

for $j = 0, \ldots, N-2$ (note that we are using a different set of inequalities than U_j). Then $\{V_j\}_{j=0}^{N-1}$ is nonincreasing because $\mathbb{A}$ is monotone. As before, $\{v_j\}_{j=1}^{N-1}$ can be seen as free variables. Assume that we can show that for some $\tau_V > 0$,

$$V_0 \le -\tau_V \left\|\tilde{\mathbb{A}}x_N\right\|^2 - \left\langle \tilde{\mathbb{A}}x_N, x_N - y_0 \right\rangle \tag{C2}$$

holds (in the sense of vector quadratic form with coefficients on H, v_j's and τ_V). Then

$$0 = V_{N-1} \le \cdots \le V_0 \le -\tau_V \left\|\tilde{\mathbb{A}}x_N\right\|^2 - \left\langle \tilde{\mathbb{A}}x_N, x_N - y_0 \right\rangle.$$

By Lemma 3.2, this implies $\left\|\tilde{\mathbf{A}}x_N\right\|^2 \leq \frac{\|y_0-x_\star\|^2}{\tau_V^2}$. Dual-OHM is an example satisfying this; selecting $\tau_V = N$ and $v_j = \frac{N}{(N-j)(N-j+1)}$ for $j = 1, \dots, N-1$ ensures (C2), which implies $\left\|\tilde{\mathbf{A}}x_N\right\|^2 \leq \frac{\|y_0-x_\star\|^2}{N^2}$. We refer to this proof strategy as *dual Lyapunov proof*.

Finally, we are ready to state the H-duality theorem, which shows that the primal Lyapunov proof for a "primal" algorithm can be transformed into a dual Lyapunov proof establishing the same rate for its H-dual algorithm.

Theorem D.1 (H-duality)**.** *Consider sequences of positive real numbers $\{u_j\}_{j=1}^{N-1}$ and $\{v_j\}_{j=1}^{N-1}$ related through $v_j = \frac{1}{u_{N-j}}$ for $j = 1, \dots, N-1$. Let $H \in \mathbb{R}^{(N-1)\times(N-1)}$ be a lower triangular matrix and $\tau > 0$. Then*

$$\left[\text{(C1) is satisfied with } H, \{u_j\}_{j=1}^{N-1} \text{ and } \tau_U = \tau\right] \quad \Leftrightarrow \quad \left[\text{(C2) is satisfied with } H^A, \{v_j\}_{j=1}^{N-1} \text{ and } \tau_V = \tau\right].$$

D.2. Proof of Theorem D.1

Define the two vector quadratic forms $\mathcal{S}$ and $\mathcal{T}$ by

$$\begin{aligned}
\mathcal{S}(\tilde{\mathbf{A}}x_1, \dots, \tilde{\mathbf{A}}x_N) &= U_N\left(H, \{u_j\}_{j=1}^{N-1}\right) - \tau\left\|\tilde{\mathbf{A}}x_N\right\|^2 - \left\langle \tilde{\mathbf{A}}x_N, x_N(H) - y_0\right\rangle \\
\mathcal{T}(\tilde{\mathbf{A}}x_1, \dots, \tilde{\mathbf{A}}x_N) &= -V_0\left(H^{\mathrm{A}}, \{v_j\}_{j=1}^{N-1}\right) - \tau\left\|\tilde{\mathbf{A}}x_N\right\|^2 - \left\langle \tilde{\mathbf{A}}x_N, x_N\left(H^{\mathrm{A}}\right) - y_0\right\rangle,
\end{aligned}$$

where we write $U_N\left(H, \{u_j\}_{j=1}^{N-1}\right)$ and $x_N(H)$ to specify that these quantities are defined through the rules of Appendix D.1 while using the H-matrix H to define $x_1, \dots, x_N$ (according to (34)) and using those $x_1, \dots, x_N$ and the specified sequence $\{u_j\}_{j=1}^{N-1}$ to define U_N. Similarly, $V_0\left(H^{\mathrm{A}}, \{v_j\}_{j=1}^{N-1}\right)$ and $x_N\left(H^{\mathrm{A}}\right)$ specifies that they are defined using H^{A} as H-matrix and using the sequence $\{v_j\}_{j=1}^{N-1}$.

Rewriting $\mathcal{S}$. We first expand $\mathcal{S}$ without explicitly expressing its dependency on H:

$$\begin{aligned}
&\mathcal{S}(\tilde{\mathbf{A}}x_1, \dots, \tilde{\mathbf{A}}x_N) \\
&= U_N - \tau\left\|\tilde{\mathbf{A}}x_N\right\|^2 - \left\langle \tilde{\mathbf{A}}x_N, x_N - y_0\right\rangle \\
&= -\sum_{k=1}^{N-1} u_k \left\langle \tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k, x_{k+1} - x_k\right\rangle - \tau\left\|\tilde{\mathbf{A}}x_N\right\|^2 - \left\langle \tilde{\mathbf{A}}x_N, x_N - y_0\right\rangle \\
&= -\tau\left\|\tilde{\mathbf{A}}x_N\right\|^2 - \left\langle \tilde{\mathbf{A}}x_N, y_{N-1} - \tilde{\mathbf{A}}x_N - y_0\right\rangle - \sum_{k=1}^{N-1} u_k \left\langle \tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k, y_k - \tilde{\mathbf{A}}x_{k+1} - y_{k-1} + \tilde{\mathbf{A}}x_k\right\rangle \\
&= -(\tau-1)\left\|\tilde{\mathbf{A}}x_N\right\|^2 + \left\langle \tilde{\mathbf{A}}x_N, y_0 - y_{N-1}\right\rangle + \sum_{k=1}^{N-1} u_k \left\|\tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k\right\|^2 - \sum_{k=1}^{N-1} u_k \left\langle \tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k, y_k - y_{k-1}\right\rangle \quad (35)
\end{aligned}$$

Now rewrite $\left\langle \tilde{\mathbf{A}}x_N, y_0 - y_{N-1}\right\rangle = -\sum_{k=1}^{N-1}\left\langle \tilde{\mathbf{A}}x_N, y_k - y_{k-1}\right\rangle$ and group this summation with the last summation within (35) to obtain

$$\begin{aligned}
&\mathcal{S}(\tilde{\mathbf{A}}x_1, \dots, \tilde{\mathbf{A}}x_N) \\
&= -(\tau-1)\left\|\tilde{\mathbf{A}}x_N\right\|^2 + \sum_{k=1}^{N-1} u_k \left\|\tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k\right\|^2 - \sum_{k=1}^{N-1}\left\langle u_k(\tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k) + \tilde{\mathbf{A}}x_N, y_k - y_{k-1}\right\rangle \\
&= -(\tau-1)\left\|\tilde{\mathbf{A}}x_N\right\|^2 + \sum_{k=1}^{N-1} u_k \left\|\tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k\right\|^2 + \sum_{k=1}^{N-1}\left\langle u_k(\tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k) + \tilde{\mathbf{A}}x_N, \sum_{j=0}^{k-1} 2h_{k,j+1}\tilde{\mathbf{A}}x_{j+1}\right\rangle \quad (36)
\end{aligned}$$

where in the last line we substitute $y_k - y_{k-1}$ using the update rule (34) (which finally reveals the H-dependency of $\mathcal{S}$).

Rewriting $\mathcal{T}$. We similarly expand $\mathcal{T}$:

$$\begin{aligned}
&\mathcal{T}(\tilde{\mathbf{A}}x_1,\dots,\tilde{\mathbf{A}}x_N)\\
&= -V_0 - \tau\left\|\tilde{\mathbf{A}}x_N\right\|^2 - \left\langle \tilde{\mathbf{A}}x_N, x_N - y_0\right\rangle\\
&= -\sum_{k=1}^{N-1} v_k\left\langle \tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k, x_N - x_k\right\rangle - \tau\left\|\tilde{\mathbf{A}}x_N\right\|^2 - \left\langle \tilde{\mathbf{A}}x_N, x_N - y_0\right\rangle\\
&= -\tau\left\|\tilde{\mathbf{A}}x_N\right\|^2 - \left\langle \tilde{\mathbf{A}}x_N, y_{N-1} - \tilde{\mathbf{A}}x_N - y_0\right\rangle - \sum_{k=1}^{N-1} v_k\left\langle \tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k, y_{N-1} - \tilde{\mathbf{A}}x_N - y_{k-1} + \tilde{\mathbf{A}}x_k\right\rangle\\
&= -(\tau-1)\left\|\tilde{\mathbf{A}}x_N\right\|^2 + \left\langle \tilde{\mathbf{A}}x_N, y_0 - y_{N-1}\right\rangle + \sum_{k=1}^{N-1} v_k\left\|\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k\right\|^2 - \sum_{k=1}^{N-1} v_k\left\langle \tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k, y_{N-1} - y_{k-1}\right\rangle \quad (37)
\end{aligned}$$

We can rewrite the last summation in (37) as

$$\begin{aligned}
-\sum_{k=1}^{N-1} v_k\left\langle \tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k, y_{N-1} - y_{k-1}\right\rangle &= -\sum_{k=1}^{N-1}\sum_{j=k}^{N-1}\left\langle v_k(\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k), y_j - y_{j-1}\right\rangle\\
&= -\sum_{j=1}^{N-1}\sum_{k=1}^{j}\left\langle v_k(\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k), y_j - y_{j-1}\right\rangle\\
&= -\sum_{k=1}^{N-1}\sum_{j=1}^{k}\left\langle v_j(\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k), y_k - y_{k-1}\right\rangle\\
&= -\sum_{k=1}^{N-1}\left\langle v_1(\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_1) + \cdots + v_k(\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k), y_k - y_{k-1}\right\rangle
\end{aligned}$$

where in the second-last equality, we switch the roles of indices k and j. Now plug this back into (37) together with the identity $\left\langle \tilde{\mathbf{A}}x_N, y_0 - y_{N-1}\right\rangle = -\sum_{k=1}^{N-1}\left\langle \tilde{\mathbf{A}}x_N, y_k - y_{k-1}\right\rangle$ to obtain

$$\begin{aligned}
&\mathcal{T}(\tilde{\mathbf{A}}x_1,\dots,\tilde{\mathbf{A}}x_N)\\
&= -(\tau-1)\left\|\tilde{\mathbf{A}}x_N\right\|^2 + \sum_{k=1}^{N-1} v_k\left\|\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k\right\|^2\\
&\quad - \sum_{k=1}^{N-1}\left\langle \tilde{\mathbf{A}}x_N + v_1(\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_1) + \cdots + v_k(\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k), y_k - y_{k-1}\right\rangle\\
&= -(\tau-1)\left\|\tilde{\mathbf{A}}x_N\right\|^2 + \sum_{k=1}^{N-1} v_k\left\|\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k\right\|^2\\
&\quad + \sum_{k=1}^{N-1}\left\langle \tilde{\mathbf{A}}x_N + v_1(\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_1) + \cdots + v_k(\tilde{\mathbf{A}}x_N - \tilde{\mathbf{A}}x_k), \sum_{j=0}^{k-1} 2h_{N-j-1,N-k}\tilde{\mathbf{A}}x_{j+1}\right\rangle \quad (38)
\end{aligned}$$

where in the last line we substitute $y_k - y_{k-1}$ using the update rule (34) but with entries of H^{A}, i.e., $h_{N-j-1,N-k-1}$ in place of $h_{k+1,j+1}$, which reveals the H^{A}-dependency of $\mathcal{T}$.

Equivalence of $\mathcal{S}\geq 0$ and $\mathcal{T}\geq 0$. Now we construct a one-to-one correspondence $\mathcal{F}\colon \prod_{j=1}^N \mathbb{R}^d \to \prod_{j=1}^N \mathbb{R}^d$ satisfying $\mathcal{S}(g_1,\dots,g_N) = \mathcal{T}(\mathcal{F}(g_1,\dots,g_N))$. Note that once this is established, it immediately follows that

$$\left[\mathcal{S}(g_1,\dots,g_N)\geq 0,\quad \forall\, g_1,\dots,g_N\in\mathbb{R}^d\right] \quad\Longleftrightarrow\quad \left[\mathcal{T}(g_1,\dots,g_N)\geq 0,\quad \forall\, g_1,\dots,g_N\in\mathbb{R}^d\right].$$

Specifically, define

$$\mathcal{F}\left(g_1,\dots,g_N\right) = \left(u_{N-1}(g_N - g_{N-1}) + g_N,\dots,u_1(g_2-g_1)+g_N, g_N\right) = (g_1',\dots,g_N').$$

Then $\mathcal{F}$ is a bijection because it has the explicit inverse mapping

$$\mathcal{F}^{-1}(g'_1,\dots,g'_N)=\left(g'_N+\sum_{k=1}^{N-1}\frac{1}{u_{N-k}}(g'_N-g'_k),\dots,g'_N+\frac{1}{u_{N-1}}(g'_N-g'_1),g'_N\right).$$

It remains to verify that $\mathcal{S}(g_1,\dots,g_N)=\mathcal{T}(\mathcal{F}(g_1,\dots,g_N))$ indeed holds true. We directly plug in the transformed set of vectors $g'_1,\dots,g'_N$ to the expansion (38) of $\mathcal{T}$, and substitute $v_k=\frac{1}{u_{N-k}}$ for $k=1,\dots,N$:

$$\begin{aligned}\mathcal{T}(g'_1,\dots,g'_N)&=-(\tau-1)\left\|g'_N\right\|^2+\sum_{k=1}^{N-1}v_k\left\|g'_N-g'_k\right\|^2\\&\quad+\sum_{k=1}^{N-1}\left\langle g'_N+v_1(g'_N-g'_1)+\dots+v_k(g'_N-g'_k),\sum_{j=0}^{k-1}2h_{N-j-1,N-k}g'_{j+1}\right\rangle\\&=-(\tau-1)\left\|g'_N\right\|^2+\sum_{k=1}^{N-1}\frac{1}{u_{N-k}}\left\|g'_N-g'_k\right\|^2\\&\quad+\sum_{k=1}^{N-1}\left\langle g'_N+\frac{1}{u_{N-1}}(g'_N-g'_1)+\dots+\frac{1}{u_{N-k}}(g'_N-g'_k),\sum_{j=0}^{k-1}2h_{N-j-1,N-k}g'_{j+1}\right\rangle\end{aligned}\tag{39}$$

Note that for $k=1,\dots,N-1$, it follows directly from the definition of $\mathcal{F}$ that $g'_N-g'_k=u_{N-k}(g_{N-k}-g_{N-k+1})$. Therefore, the first two terms of (39) can be rewritten as

$$\begin{aligned}-(\tau-1)\left\|g'_N\right\|^2+\sum_{k=1}^{N-1}\frac{1}{u_{N-k}}\left\|g'_N-g'_k\right\|^2&=-(\tau-1)\left\|g_N\right\|^2+\sum_{k=1}^{N-1}u_{N-k}\left\|g_{N-k}-g_{N-k+1}\right\|^2\\&=-(\tau-1)\left\|g_N\right\|^2+\sum_{k=1}^{N-1}u_k\left\|g_{k+1}-g_k\right\|^2\end{aligned}$$

and the last expression coincides with the first two terms within (36). Next, rewrite the second line of (39) as following:

$$\begin{aligned}&\sum_{k=1}^{N-1}\left\langle g'_N+\frac{1}{u_{N-1}}(g'_N-g'_1)+\dots+\frac{1}{u_{N-k}}(g'_N-g'_k),\sum_{j=0}^{k-1}2h_{N-j-1,N-k}g'_{j+1}\right\rangle\\&=\sum_{k=1}^{N-1}\left\langle g_N+(g_{N-1}-g_N)+\dots+(g_{N-k}-g_{N-k+1}),\sum_{j=0}^{k-1}2h_{N-j-1,N-k}g'_{j+1}\right\rangle\\&=\sum_{k=1}^{N-1}\left\langle g_{N-k},\sum_{j=0}^{k-1}2h_{N-j-1,N-k}g'_{j+1}\right\rangle\\&=\sum_{k=1}^{N-1}\sum_{j=0}^{k-1}2h_{N-j-1,N-k}\left\langle g_{N-k},g_N+u_{N-j-1}(g_{N-j}-g_{N-j-1})\right\rangle\\&=\sum_{k=1}^{N-1}\sum_{i=N-k}^{N-1}2h_{i,N-k}\left\langle g_{N-k},g_N+u_i(g_{i+1}-g_i)\right\rangle\end{aligned}$$

where to obtain the last equality, we make an index substitution $i=N-j-1$. From the last line, make another substitution of index $\ell=N-k$:

$$\begin{aligned}\sum_{k=1}^{N-1}\sum_{i=N-k}^{N-1}2h_{i,N-k}\left\langle g_{N-k},g_N+u_i(g_{i+1}-g_i)\right\rangle&=\sum_{\ell=1}^{N-1}\sum_{i=\ell}^{N-1}2h_{i,\ell}\left\langle g_\ell,g_N+u_i(g_{i+1}-g_i)\right\rangle\\&=\sum_{i=1}^{N-1}\sum_{\ell=1}^{i}2h_{i,\ell}\left\langle g_\ell,g_N+u_i(g_{i+1}-g_i)\right\rangle.\end{aligned}\tag{40}$$

Finally, changing the name of the indices (i, ℓ) to $(k, j+1)$ in (40) gives

$$\begin{aligned}\sum_{i=1}^{N-1}\sum_{\ell=1}^{i} 2h_{i,\ell}\left\langle g_\ell, g_N + u_i(g_{i+1}-g_i)\right\rangle &= \sum_{k=1}^{N-1}\sum_{j=0}^{k-1} 2h_{k,j+1}\left\langle g_{j+1}, g_N + u_k(g_{k+1}-g_k)\right\rangle \\ &= \sum_{k=1}^{N-1}\left\langle g_N + u_k(g_{k+1}-g_k), \sum_{j=0}^{k-1} 2h_{k,j+1}g_{j+1}\right\rangle\end{aligned}$$

where the last expression coincides with the last summation within (36). This shows that $\mathcal{T}(\mathcal{F}(g_1,\dots,g_N)) = \mathcal{T}(g'_1,\dots,g'_N) = \mathcal{S}(g_1,\dots,g_N)$, completing the proof of Theorem D.1.

E. Proof of the optimal family theorem

E.1. Overview of the proof and description of the parametrization $\Phi\colon C \to \mathbb{R}^{(N-1)\times(N-1)}$

In this section, we prove Theorem 4.1. The full proof is long and complicated, so we first outline the structure of the proof.

1. We consider the following proof strategy: If there is some index set $I \subseteq \{1,\dots,N\} \times \{1,\dots,N\}$ and nonnegative real numbers $\lambda_{i,j}$ for each $(i,j) \in I$ such that

$$0 = \left\langle \tilde{\mathbb{A}} x_N, x_N - y_0 \right\rangle + N\|\tilde{\mathbb{A}} x_N\|^2 + \sum_{(i,j)\in I} \lambda_{i,j} \left\langle \tilde{\mathbb{A}} x_i - \tilde{\mathbb{A}} x_j, x_i - x_j \right\rangle \tag{41}$$

   then the algorithm exhibits the rate

$$\|y_{N-1} - \mathbb{T} y_{N-1}\|^2 = 4 \left\|\tilde{\mathbb{A}} x_N\right\|^2 \leq \frac{4\|y_0 - y_\star\|^2}{N^2}.$$

   The final convergence rate is a direct consequence of (41), monotonicity of $\mathbb{A}$ and Lemma 3.2.

2. We choose $I = \{(k+1,k) \,|\, k = 1,\dots,N-1\} \cup \{(N,k) \,|\, k = 1,\dots,N-1\}$, and we use the H-matrix representation (34) to eliminate $x_1,\dots,x_N$ within (41). Then (41) becomes a vector quadratic form in $\tilde{\mathbb{A}} x_1,\dots,\tilde{\mathbb{A}} x_N$, i.e.,

$$\left\langle \tilde{\mathbb{A}} x_N, x_N - y_0 \right\rangle + N\|\tilde{\mathbb{A}} x_N\|^2 + \sum_{(i,j)\in I} \lambda_{i,j} \left\langle \tilde{\mathbb{A}} x_i - \tilde{\mathbb{A}} x_j, x_i - x_j \right\rangle = \sum_{k=1}^{N} \sum_{\ell=k}^{N} s_{\ell,k} \left\langle \tilde{\mathbb{A}} x_\ell, \tilde{\mathbb{A}} x_k \right\rangle$$

   where the coefficients $s_{\ell,k} = s_{\ell,k}(H,\lambda)$ are functions depending on H and $\lambda = (\lambda_{i,j})_{(i,j)\in I}$ (note that we only keep $s_{\ell,k}$ with $\ell \geq k$ to avoid redundancy). We characterize the explicit expressions for $s_{\ell,k}$ in terms of λ and entries of H, reducing the problem of establishing the identity (41) to the problem of solving the system of equations

$$s_{\ell,k}(H,\lambda) = 0 \quad (k = 1,\dots,N,\ \ell = k,\dots,N). \tag{42}$$

3. We provide explicit solutions for λ, in terms of diagonal entries $h_{k,k}$ $(k = 1,\dots,N-1)$ of the H-matrix. With these explicit values of λ, the system (42) becomes nonlinear in $h_{k,k}$ $(k = 1,\dots,N-1)$, but it is a linear system in non-diagonal entries $h_{\ell,k}$ $(\ell = k+1,\dots,N-1)$ of H.

4. (This is the most technical core step of analysis.) We show that with the expressions of λ from the previous step, under certain conditions on $h_{k,k}$ $(k = 1,\dots,N-1)$, the linear system (42) is uniquely solvable in the non-diagonal H-matrix entries $h_{\ell,k}$ $(\ell = k+1,\dots,N-1)$, and these unique solutions

$$h^*_{\ell,k} = h^*_{\ell,k}(h_{1,1}, h_{2,2}, \dots, h_{N-1,N-1}) \tag{43}$$

   can be expressed as continuous functions of the diagonal $h_{k,k}$ $(k = 1,\dots,N-1)$.

We now describe how $\Phi\colon C \to \mathbb{R}^{(N-1)\times(N-1)}$ is constructed. Define

$$p_k = \prod_{\ell=k}^{N-1} h_{\ell,\ell}, \quad k = 1,\dots,N-1 \tag{44}$$

(these quantities play significant role in all analyses of this section). It turns out that the affine constraints

$$p_1 = \frac{1}{N} \tag{45}$$

$$p_k \geq \frac{1}{N-k+1} \quad (k = 2,\dots,N-1) \tag{46}$$

$$p_k \geq \frac{N-k}{N-k-1} p_{k+1} - \frac{1}{N-k-1} \quad (k = 1,\dots,N-2) \tag{47}$$

are the key conditions making the construction of Steps 1 through 4 work. Specifically, we have $\lambda_{i,j} > 0$ for explicit expressions of λ from Step 3, if the constraints (45) holds, and (46) and (47) holds with strict inequality. The set C of all $(p_2, \dots, p_{N-1})$ satisfying the inequality constraints (46) and (47) strictly is an open convex subset of $\mathbb{R}^{N-2}$. To check that $C \neq \emptyset$, consider the H-matrices of OHM and Dual-OHM. From OHM, we have

$$p_k^{(0)} = \prod_{\ell=k}^{N-1} h_{\ell,\ell} = \prod_{\ell=k}^{N-1} \frac{\ell}{\ell+1} = \frac{k}{N}$$

and from Dual-OHM,

$$p_k^{(1)} = \prod_{\ell=k}^{N-1} h_{\ell,\ell} = \prod_{\ell=k}^{N-1} \frac{N-\ell}{N-\ell+1} = \frac{1}{N-k+1}.$$

It can be checked by direct calculations that $p_1^{(0)} = \frac{1}{N} = p_1^{(1)}$ and that $p_k^{(0)}$'s satisfy (46) strictly and (47) with equality, while $p_k^{(1)}$'s satisfy (46) with equality and (47) strictly. Therefore, $\left(p_2^{(0)}, \dots, p_{N-1}^{(0)}\right), \left(p_2^{(0)}, \dots, p_{N-1}^{(0)}\right) \in \partial C$. Additionally, if we define $p_k^{(\gamma)} := \gamma p_k^{(0)} + (1-\gamma) p_k^{(1)}$ for $\gamma \in (0,1)$ and $k = 1, \dots, N-1$, then $p_1^{(\gamma)} = \frac{1}{N}$, and $\left(p_2^{(\gamma)}, \dots, p_{N-1}^{(\gamma)}\right)$ satisfies both (46), (47) strictly, i.e.,

$$\left(p_2^{(\gamma)}, \dots, p_{N-1}^{(\gamma)}\right) \in C.$$

This shows that C is nonempty.

Note that provided that $p_1 = \frac{1}{N}$, we can recover the diagonal $h_{k,k}$'s from $(p_2, \dots, p_{N-1}) \in C$ as $h_{1,1} = \frac{p_1}{p_2} = \frac{1}{Np_2}$, $h_{k,k} = \frac{p_k}{p_{k+1}}$ for $k = 2, \dots, N-2$ and $h_{N-1,N-1} = p_{N-1}$. Now assuming that Step 4 is done so that we have $h^*_{\ell,k}$ determined as functions of $h_{1,1}, \dots, h_{N-1,N-1}$ as in (43), we define $\Phi : C \to \mathbb{R}^{(N-1)\times(N-1)}$ by

$$\Phi(p_2, \dots, p_{N-1}) = \begin{bmatrix} \frac{1}{Np_2} & 0 & \cdots & 0 \\ h^*_{2,1}\left(\frac{1}{Np_2}, \frac{p_2}{p_3}, \dots, p_{N-1}\right) & \frac{p_2}{p_3} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h^*_{N-1,1}\left(\frac{1}{Np_2}, \frac{p_2}{p_3}, \dots, p_{N-1}\right) & h^*_{N-1,2}\left(\frac{1}{Np_2}, \frac{p_2}{p_3}, \dots, p_{N-1}\right) & \cdots & p_{N-1} \end{bmatrix}.$$

This map is injective because one can recover the values of $p_2, \dots, p_{N-1}$ from the diagonal entries of $\Phi(p_2, \dots, p_{N-1})$. Additionally, it is continuous provided that $h^*_{\ell,k}$ are continuous. Because $\Phi(w)$ is designed to satisfy (41) (when it is used as H-matrix) for any $w = (p_2, \dots, p_{N-1}) \in C$, by monotonicity of $\mathbb{A}$, the algorithm with $\Phi(w)$ as H-matrix exhibits the exact optimal rate of $\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 = 4\left\|\tilde{\mathbb{A}}x_N\right\|^2 \leq \frac{4\|y_0 - y_\star\|^2}{N^2}$ by Step 1. In the subsequent sections, we go through each steps outlined above, and explain how the quantities p_k defined in (44) and the constraints (45), (46) and (47) become relevant to analysis.

E.2. Step 2: Computation of coefficient functions $s_{\ell,k}$ in vector quadratic form

We rewrite the terms $x_N - y_0$ and $x_i - x_j$ in the right hand side of (41) as linear combinations of $\tilde{\mathbb{A}}x_k$ $(k = 1, \dots, N)$ according to the definition (6) and $x_k = \mathbb{J}_{\mathbb{A}}(y_{k-1}) = y_{k-1} - \tilde{\mathbb{A}}x_k$, and then expand everything. Denoting $g_k = \tilde{\mathbb{A}}x_k$ for

simplicity, the expansion goes:

$$
\begin{aligned}
&\left\langle \tilde{\mathbb{A}}x_N, x_N - y_0 \right\rangle + N\|\tilde{\mathbb{A}}x_N\|^2 + \sum_{(i,j)\in I} \lambda_{i,j} \left\langle \tilde{\mathbb{A}}x_i - \tilde{\mathbb{A}}x_j, x_i - x_j \right\rangle \\
&= \left\langle \tilde{\mathbb{A}}x_N, x_N - y_0 \right\rangle + N\|\tilde{\mathbb{A}}x_N\|^2 + \sum_{k=1}^{N-2} \lambda_{k+1,k} \left\langle \tilde{\mathbb{A}}x_{k+1} - \tilde{\mathbb{A}}x_k, x_{k+1} - x_k \right\rangle + \sum_{k=1}^{N-2} \lambda_{N,k} \left\langle \tilde{\mathbb{A}}x_N - \tilde{\mathbb{A}}x_k, x_N - x_k \right\rangle \\
&\quad + \lambda_{N,N-1} \left\langle \tilde{\mathbb{A}}x_N - \tilde{\mathbb{A}}x_{N-1}, x_N - x_{N-1} \right\rangle \\
&= \left\langle g_N, -\sum_{k=1}^{N-1}\sum_{j=1}^{k} 2h_{k,j}g_j - g_N \right\rangle + N\|g_N\|^2 + \sum_{k=1}^{N-2} \lambda_{k+1,k} \left\langle g_{k+1} - g_k, -\sum_{j=1}^{k} 2h_{k,j}g_j - (g_{k+1} - g_k) \right\rangle \\
&+ \sum_{k=1}^{N-2} \lambda_{N,k} \left\langle g_N - g_k, -\sum_{i=k}^{N-1}\sum_{j=1}^{i} 2h_{i,j}g_j - (g_N - g_k) \right\rangle + \lambda_{N,N-1} \left\langle g_N - g_{N-1}, -\sum_{j=1}^{N-1} 2h_{N-1,j}g_j - (g_N - g_{N-1}) \right\rangle \\
&= \left\langle g_N, -\sum_{j=1}^{N-1}\sum_{k=j}^{N-1} 2h_{k,j}g_j - g_N \right\rangle + N\|g_N\|^2 - \sum_{k=1}^{N-2} \lambda_{k+1,k}\|g_{k+1} - g_k\|^2 - \sum_{k=1}^{N-2} \lambda_{N,k}\|g_N - g_k\|^2 \\
&\quad - \lambda_{N,N-1}\|g_N - g_{N-1}\|^2 - \sum_{k=1}^{N-2} \lambda_{N,k} \left\langle g_N - g_k, \sum_{i=k}^{N-1}\sum_{j=1}^{i} 2h_{i,j}g_j \right\rangle - \sum_{j=1}^{N-2} \left\langle \sum_{k=j}^{N-2} 2h_{k,j}\lambda_{k+1,k}(g_{k+1} - g_k), g_j \right\rangle \\
&\quad - \lambda_{N,N-1} \left\langle g_N - g_{N-1}, \sum_{j=1}^{N-1} 2h_{N-1,j}g_j \right\rangle.
\end{aligned}
$$

We carefully gather the terms within the above expansion and group the coefficients attached to the same inner product terms, into the form

$$
\sum_{k=1}^{N}\sum_{\ell=k}^{N} s_{\ell,k} \left\langle g_\ell, g_k \right\rangle. \tag{48}
$$

The result of the computation is as following:

$$
\begin{aligned}
s_{N,N} &= N - 1 - \sum_{k=1}^{N-1} \lambda_{N,k} \\
s_{N-1,N-1} &= -\lambda_{N-1,N-2} + \lambda_{N,N-1}(2h_{N-1,N-1} - 1) \\
s_{N,N-1} &= -2h_{N-1,N-1} - \sum_{k=1}^{N-2} 2\lambda_{N,k}h_{N-1,N-1} - 2\lambda_{N,N-1}(h_{N-1,N-1} - 1)
\end{aligned} \tag{49}
$$

and for $k = 1, \ldots, N-2$,

$$
\begin{aligned}
s_{k,k} &= \begin{cases} \lambda_{k+1,k}(2h_{k,k} - 1) - \lambda_{k,k-1} + \lambda_{N,k}\left(\sum_{i=k}^{N-1} 2h_{i,k} - 1\right) & \text{if } k > 1 \\ \lambda_{k+1,k}(2h_{k,k} - 1) + \lambda_{N,k}\left(\sum_{i=k}^{N-1} 2h_{i,k} - 1\right) & \text{if } k = 1 \end{cases} \\
s_{k+1,k} &= 2\lambda_{k+1,k}(1 - h_{k,k}) + 2\lambda_{k+2,k+1}h_{k+1,k} + 2\lambda_{N,k}\sum_{i=k+1}^{N-1} h_{i,k+1} + 2\lambda_{N,k+1}\sum_{i=k+1}^{N-1} h_{i,k} \quad (k < N-2) \\
s_{\ell,k} &= -2\lambda_{\ell,\ell-1}h_{\ell-1,k} + 2\lambda_{\ell+1,\ell}h_{\ell,k} + 2\lambda_{N,k}\sum_{i=\ell}^{N-1} h_{i,\ell} + 2\lambda_{N,\ell}\sum_{i=\ell}^{N-1} h_{i,k} \quad (\ell = k+2, \ldots, N-2) \\
s_{N-1,k} &= 2\lambda_{N,k}h_{N-1,N-1} - 2\lambda_{N-1,N-2}h_{N-2,k} + 2\lambda_{N,N-1}h_{N-1,k} \quad (k < N-2) \\
s_{N,k} &= 2\lambda_{N,k} - \sum_{\ell=k}^{N-1} \left(1 + \sum_{j=1}^{\ell} \lambda_{N,j}\right) 2h_{\ell,k}
\end{aligned} \tag{50}
$$

and finally,

$$s_{N-1,N-2} = 2\lambda_{N-1,N-2}(1 - h_{N-2,N-2}) + 2\lambda_{N,N-1}h_{N-1,N-2} + 2\lambda_{N,N-2}h_{N-1,N-1}. \tag{51}$$

E.3. Step 3: Explicit characterization of λ

Recall the definition (44): $p_k = \prod_{\ell=k}^{N-1} h_{\ell,\ell}$ for $k = 1, \dots, N-1$. Assume $p_1 = \frac{1}{N}$. We define

$$\begin{aligned}
\lambda_{N,N-1} &= Nh_{N-1,N-1} \\
\lambda_{k+1,k} &= \frac{N}{N-k-1}p_{k+1}\left((N-k)p_{k+1} - 1\right) \quad (k = 1, \dots, N-2) \\
\lambda_{N,k} &= \frac{N}{(N-k)(N-k-1)} - \frac{N}{N-k-1}p_{k+1} + \frac{N}{N-k}p_k \quad (k = 1, \dots, N-2)
\end{aligned} \tag{52}$$

We prove the following handy identities for λ.

Proposition E.1. *For λ defined as in (52), the following holds, provided that $p_1 = \frac{1}{N}$.*

$$\begin{aligned}
\sum_{i=k}^{N-1} \lambda_{N,i} &= \frac{N(N-k-1)}{N-k} + \frac{N}{N-k}p_k \quad (k = 2, \dots, N-1) \\
\sum_{i=1}^{N-1} \lambda_{N,i} &= N - 1.
\end{aligned}$$

Proof. Observe that the first identity, when $k = N-1$, is $\lambda_{N,N-1} = Np_{N-1} = Nh_{N-1,N-1}$, which holds true by definition. Next, for $k = 1, \dots, N-1$, we can rewrite

$$\lambda_{N,k} = \frac{N}{N-k-1} - \frac{N}{N-k} - \frac{N}{N-k-1}p_{k+1} + \frac{N}{N-k}p_k$$

so they telescope:

$$\begin{aligned}
\sum_{i=k}^{N-1} \lambda_{N,i} &= \lambda_{N,N-1} + \sum_{i=k}^{N-2} \lambda_{N,i} \\
&= Np_{N-1} + \sum_{i=k}^{N-2} \left(\frac{N}{N-i-1} - \frac{N}{N-i} - \frac{N}{N-i-1}p_{i+1} + \frac{N}{N-i}p_i \right) \\
&= Np_{N-1} + \left(N - \frac{N}{N-k} - Np_{N-1} + \frac{N}{N-k}p_k \right) \\
&= \frac{N(N-k-1)}{N-k} + \frac{N}{N-k}p_k.
\end{aligned}$$

In the case $k = 1$, we have

$$\sum_{i=1}^{N-1} \lambda_{N,i} = \frac{N(N-2)}{N-1} + \frac{N}{N-1}p_1 = \frac{N^2 - 2N + 1}{N-1} = N - 1.$$

□

Observe that $\lambda_{k+1,k} \geq 0$ if $p_{k+1} \geq \frac{1}{N-k}$ for $k = 1, \dots, N-2$, which is equivalent to (46). Moreover,

$$\begin{aligned}
&\lambda_{N,k} = \frac{N}{(N-k)(N-k-1)} - \frac{N}{N-k-1}p_{k+1} + \frac{N}{N-k}p_k \geq 0 \\
&\iff p_k \geq \frac{N-k}{N}\left(\frac{N}{N-k-1}p_{k+1} - \frac{N}{(N-k)(N-k-1)} \right) = \frac{N-k}{N-k-1}p_{k+1} - \frac{1}{N-k-1}
\end{aligned}$$

which is (47). Clearly, if these inequalities are satisfied strictly, then $\lambda_{k+1,k}, \lambda_{N,k} > 0$.

E.4. Step 4: Solving the linear system $s_{\ell,k} = 0$

We first observe that by definition of λ from the previous section and Proposition E.1,

$$\begin{aligned} s_{N,N} &= N - 1 - \sum_{k=1}^{N-1} \lambda_{N,k} = 0 \\ s_{N-1,N-1} &= -\lambda_{N-1,N-2} + \lambda_{N,N-1}(2h_{N-1,N-1} - 1) \\ &= -Np_{N-1}(2p_{N-1} - 1) + Nh_{N-1,N-1}(2h_{N-1,N-1} - 1) = 0 \end{aligned}$$

and

$$\begin{aligned} s_{N,N-1} &= -2h_{N-1,N-1} - \sum_{k=1}^{N-2} 2\lambda_{N,k}h_{N-1,N-1} - 2\lambda_{N,N-1}(h_{N-1,N-1} - 1) \\ &= -2h_{N-1,N-1}\left(1 + \sum_{k=1}^{N-1} \lambda_{N,k}\right) + 2\lambda_{N,N-1} \\ &= -2Nh_{N-1,N-1} + 2\lambda_{N,N-1} \\ &= 0 \end{aligned}$$

directly follows. However, for $s_{\ell,k}$ with $k \leq N-2$ and $\ell = k, \ldots, N$, we need to determine $h_{k+1,k}, \ldots, h_{N-1,k}$ that achieves $s_{\ell,k} = 0$. We first characterize the condition for $s_{k,k} = 0$.

Proposition E.2. *For $k = 1, \ldots, N-2$, provided that $p_1 = \frac{1}{N}$,*

$$s_{k,k} = 0 \iff \sum_{i=k}^{N-1} 2h_{i,k} = 1 - (N-k)p_{k+1} + (N-k+1)p_k. \tag{53}$$

Proof. Recall from (50) that

$$s_{k,k} = \begin{cases} \lambda_{k+1,k}(2h_{k,k} - 1) - \lambda_{k,k-1} + \lambda_{N,k}\left(\sum_{i=k}^{N-1} 2h_{i,k} - 1\right) & \text{if } k > 1 \\ \lambda_{k+1,k}(2h_{k,k} - 1) + \lambda_{N,k}\left(\sum_{i=k}^{N-1} 2h_{i,k} - 1\right) & \text{if } k = 1 \end{cases}$$

To eliminate the need to deal with the case $k = 1$ separately, define the dummy variable $\lambda_{1,0} = \frac{N}{N-1}p_1(Np_1 - 1)$, which is consistent with the rule (52) for defining $\lambda_{k+1,k}$ for $k = 1, \ldots, N-2$. Given that $p_1 = \frac{1}{N}$ we have $\lambda_{1,0} = 0$, so we can write

$$s_{k,k} = 0 \iff \sum_{i=k}^{N-1} 2h_{i,k} - 1 = \frac{\lambda_{k,k-1} - \lambda_{k+1,k}(2h_{k,k} - 1)}{\lambda_{N,k}} \tag{54}$$

where

$$\begin{aligned} \lambda_{k+1,k} &= \frac{N}{N-k-1}p_{k+1}\left((N-k)p_{k+1} - 1\right) \\ \lambda_{k,k-1} &= \frac{N}{N-k}p_k\left((N-k+1)p_k - 1\right) \\ \lambda_{N,k} &= \frac{N}{(N-k)(N-k-1)} - \frac{N}{N-k-1}p_{k+1} + \frac{N}{N-k}p_k \\ h_{k,k} &= \frac{\prod_{\ell=k}^{N-1} h_{\ell,\ell}}{\prod_{\ell=k+1}^{N-1} h_{\ell,\ell}} = \frac{p_k}{p_{k+1}} \end{aligned}$$

Plugging the above identities into the right hand side of (54) gives

$$\begin{aligned}\sum_{i=k}^{N-1} 2h_{i,k} - 1 &= \frac{\lambda_{k,k-1} - \lambda_{k+1,k}(2h_{k,k} - 1)}{\lambda_{N,k}} \\ &= \frac{\frac{N}{N-k}p_k\left((N-k+1)p_k - 1\right) - \frac{N}{N-k-1}p_{k+1}\left((N-k)p_{k+1} - 1\right)\left(\frac{2p_k}{p_{k+1}} - 1\right)}{\frac{N}{(N-k)(N-k-1)} - \frac{N}{N-k-1}p_{k+1} + \frac{N}{N-k}p_k}.\end{aligned}$$

Multiply $\frac{(N-k)(N-k-1)}{N}$ throughout both numerator and denominator, and rearrange the expressions using the substitution $d_k = 1 - (N-k)p_{k+1} + (N-k-1)p_k$:

$$\begin{aligned}\sum_{i=k}^{N-1} 2h_{i,k} - 1 &= \frac{(N-k-1)p_k\left((N-k+1)p_k - 1\right) - (N-k)\left((N-k)p_{k+1} - 1\right)(2p_k - p_{k+1})}{1 - (N-k)p_{k+1} + (N-k-1)p_k} \\ &= \frac{(N-k-1)p_k\left((N-k+1)p_k - 1\right) - \left((N-k)p_{k+1} - 1\right)(2(N-k)p_k - (N-k)p_{k+1})}{1 - (N-k)p_{k+1} + (N-k-1)p_k} \\ &= \frac{(N-k-1)p_k\left((N-k+1)p_k - 1\right) - \left((N-k-1)p_k - d_k\right)(2(N-k)p_k + d_k - 1 - (N-k-1)p_k)}{d_k} \\ &= \frac{(N-k-1)p_k\left((N-k+1)p_k - 1\right) - \left((N-k-1)p_k - d_k\right)\left((N-k+1)p_k + d_k - 1\right)}{d_k} \\ &= \frac{d_k^2 + (2p_k - 1)d_k}{d_k} \\ &= d_k + 2p_k - 1.\end{aligned}$$

This is again equivalent to

$$\sum_{i=k}^{N-1} 2h_{i,k} = d_k + 2p_k = 1 - (N-k)p_{k+1} + (N-k+1)p_k.$$

□

We denote the rightmost quantity from (53) by

$$q_k = 1 - (N-k)p_{k+1} + (N-k+1)p_k, \tag{55}$$

as it will be frequently used in the subsequent analysis.

In (50), we see that for each fixed $k = 1, \ldots, N-2$, the equations $s_{\ell,k} = 0$ ($\ell \geq k$) involve only $h_{i,j}$'s with $i \geq j \geq k$. Therefore, rather than trying to solve the whole system $s_{\ell,j} = 0$ ($j = 1, \ldots, N-2$, $\ell = k, \ldots, N$) at once, we iteratively solve the partial systems $s_{\ell,k} = 0$ ($\ell = k, \ldots, N$) one-by-one with fixed value of k, starting from $k = N-2$ and progressively lowering it to $k = 1$.

Proposition E.3. *For each $k = 1, \ldots, N-2$ fixed, the system of equations $s_{\ell,k} = 0$ ($\ell = k, \ldots, N$), viewing only $h_{k+1,k}, \ldots, h_{N-1,k}$ as variables, is a uniquely solvable linear system, if: all $\lambda_{i,j}$'s defined in (52) are positive and*

$$\sum_{i=j}^{N-1} 2h_{i,j} = q_j \tag{56}$$

holds for $j = k+1, \ldots, N-2$. In this case, the solutions $h^*_{k+1,k}, \ldots, h^*_{N-1,k}$ satisfying $s_{\ell,k} = 0$ ($\ell = k, \ldots, N$) can be written as continuous functions of $h_{1,1}, \ldots, h_{N-1,N-1}$, i.e.,

$$h^*_{i,k} = h^*_{i,k}\left(h_{1,1}, \ldots, h_{N-1,N-1}\right).$$

Proof. First consider the case $k = N-2$, which requires a separate treatment because $s_{N-1,N-2}$ is defined separately as (51). In this case, we have 3 equations to solve, namely

$$s_{N-2,N-2} = 0, \quad s_{N-1,N-2} = 0, \quad s_{N,N-2} = 0. \tag{57}$$

By Proposition E.2, we have

$$\begin{aligned} s_{N-2,N-2} = 0 &\iff 2(h_{N-2,N-2} + h_{N-1,N-2}) = q_{N-1} = 1 - 2p_{N-1} + 3p_{N-2} \\ &\iff h_{N-1,N-2} = \frac{1 - 2h_{N-1} + 3h_{N-2,N-2}h_{N-1,N-1}}{2} - h_{N-2,N-2}. \end{aligned} \tag{58}$$

Plugging the last expression for $h_{N-1,N-2}$ into

$$\begin{aligned} s_{N-1,N-2} &= 2\lambda_{N-1,N-2}(1 - h_{N-2,N-2}) + 2\lambda_{N,N-1}h_{N-1,N-2} + 2\lambda_{N,N-2}h_{N-1,N-1} \\ &= 2Nh_{N-1,N-1}(2h_{N-1,N-1} - 1)(1 - h_{N-2,N-2}) + 2Nh_{N-1,N-1}h_{N-1,N-2} \\ &\quad + 2\left(\frac{N}{2} - Nh_{N-1,N-1} + \frac{N}{2}h_{N-2,N-2}h_9N-1,N-1\right)h_{N-1,N-1} \end{aligned}$$

gives $s_{N-1,N-2} = 0$. Next, provided that (58) holds true, we have

$$\begin{aligned} s_{N,N-2} &= 2\lambda_{N,N-2} - \left(1 + \sum_{i=1}^{N-2}\lambda_{N,i}\right)2h_{N-2,k} - \left(1 + \sum_{i=1}^{N-1}\lambda_{N,i}\right)2h_{N-1,k} \\ &= 2\lambda_{N,N-2} - (N - \lambda_{N,N-1})2h_{N-2,N-2} - 2Nh_{N-1,N-2} \\ &= 2\left(\frac{N}{2} - Nh_{N-1,N-1} + \frac{N}{2}h_{N-2,N-2}h_{N-1,N-1}\right) + 2\lambda_{N,N-1}h_{N-2,N-2} - N(2h_{N-2,N-2} + 2h_{N-1,N-1}) \\ &= 2\left(\frac{N}{2} - Nh_{N-1,N-1} + \frac{N}{2}h_{N-2,N-2}h_{N-1,N-1}\right) + 2Nh_{N-1,N-1}h_{N-2,N-2} \\ &\quad - N(1 - 2h_{N-1} + 3h_{N-2,N-2}h_{N-1,N-1}) \\ &= 0. \end{aligned}$$

This shows the value of $h_{N-1,N-2}$ characterized by (58) is the unique solution solving the equations (57). The expression determining $h_{N-1,N-2}$ is clearly continuous in $h_{N-2,N-2}, h_{N-1,N-1}$.

Now suppose $k < N-2$. We first consider the following system of equations (without $s_{N-1,k} = 0$ and $s_{N,k} = 0$):

$$\begin{aligned} s_{k,k} &= \lambda_{N,k}\left(\sum_{i=k+1}^{N-1} 2h_{i,k} + 2h_{k,k} - q_k\right) = 0 \\ s_{k+1,k} &= 2\lambda_{k+1,k}(1 - h_{k,k}) + 2\lambda_{k+2,k+1}h_{k+1,k} + 2\lambda_{N,k}\sum_{i=k+1}^{N-1} h_{i,k+1} + 2\lambda_{N,k+1}\sum_{i=k+1}^{N-1} h_{i,k} = 0 \\ s_{\ell,k} &= -2\lambda_{\ell,\ell-1}h_{\ell-1,k} + 2\lambda_{\ell+1,\ell}h_{\ell,k} + 2\lambda_{N,k}\sum_{i=\ell}^{N-1} h_{i,\ell} + 2\lambda_{N,\ell}\sum_{i=\ell}^{N-1} h_{i,k} = 0 \quad (\ell = k+2, \dots, N-2) \end{aligned} \tag{59}$$

where the first identity has been reorganized in a simpler form using Proposition E.2. We show that the system (59) form an invertible linear system in the variables $h_{k+1,k}, \dots, h_{N-1,k}$, and thus have unique solutions. Then we show that these solutions are consistent with the remaining two equations to complete the proof.

Claim 1. The system of equations (59) is uniquely solvable.

In (59), the non-diagonal $h_{i,j}$'s with $j = k+1, \dots, N-2$ are involved only through the summation $\sum_{i=j}^{N-1} h_{i,j}$, which can be replaced with $q_j/2$ by the assumption (56). Now for each $\ell = k+1, \dots, N-2$ we subtract a multiple of $s_{k,k}$ from

each $s_{\ell,k}$ to eliminate the variables $h_{\ell,k}, \dots, h_{N-1,k}$ while retaining an equivalent system of linear equations:

$$\begin{aligned}
s'_{k+1,k} &:= s_{k+1,k} - \frac{\lambda_{N,k+1}}{\lambda_{N,k}} s_{k,k} \\
&= 2\lambda_{k+2,k+1} h_{k+1,k} + \underbrace{\lambda_{N,k} q_{k+1} + 2\lambda_{k+1,k}(1 - h_{k,k}) + \lambda_{N,k+1}(q_k - 2h_{k,k})}_{:= t_{k+1,k} = \text{Terms not depending on } h_{k+1,k}, \dots, h_{N-1,k}} \\
s'_{\ell,k} &:= s_{\ell,k} - \frac{\lambda_{N,\ell}}{\lambda_{N,k}} s_{k,k} \\
&= -2\lambda_{\ell,\ell-1} h_{\ell-1,k} + 2\lambda_{\ell+1,\ell} h_{\ell,k} - 2\lambda_{N,\ell} \sum_{i=k+1}^{\ell-1} h_{i,k} + \underbrace{\lambda_{N,\ell}(q_k - 2h_{k,k}) + \lambda_{N,k} q_\ell}_{\substack{:= t_{\ell,k} \\ (\text{Terms not depending on } h_{k+1,k}, \dots, h_{N-1,k})}} \quad (\ell = k+2, \dots, N-2).
\end{aligned}$$

For the sake of conciseness, we separately define the terms within $s'_{\ell,k}$ not depending on $h_{k+1,k}, \dots, h_{N-1,k}$:

$$t_{\ell,k} = \begin{cases} \lambda_{N,k} q_{k+1} + 2\lambda_{k+1,k}(1 - h_{k,k}) + \lambda_{N,k+1}(q_k - 2h_{k,k}) & \text{if } \ell = k+1 \\ \lambda_{N,\ell}(q_k - 2h_{k,k}) + \lambda_{N,k} q_\ell & \text{if } \ell = k+2, \dots, N-2 \end{cases} \tag{60}$$

so that we can write

$$\begin{aligned}
s'_{k+1,k} &= 2\lambda_{k+2,k+1} h_{k+1,k} + t_{k+1,k} \\
s'_{\ell,k} &= -2\lambda_{\ell,\ell-1} h_{\ell-1,k} + 2\lambda_{\ell+1,\ell} h_{\ell,k} - 2\lambda_{N,\ell} \sum_{i=k+1}^{\ell-1} h_{i,k} + t_{\ell,k}.
\end{aligned}$$

Next, we express the system of equations $\frac{s'_{\ell,k}}{2} = 0$ ($\ell = k+1, \dots, N-1$) in matrix form. To do this, we define the vectors within $\mathbb{R}^{N-k-1}$, holding the coefficients attached to the variables $h_{k+1,k}, \dots, h_{N-1,k}$ within each $\frac{s'_{\ell,k}}{2}$:

$$\begin{array}{rcccccccc}
\mathbf{a}_1 & = \big[& \lambda_{k+2,k+1} & 0 & 0 & \cdots & 0 & 0 & 0 \big]^\intercal \\
\mathbf{a}_2 & = \big[& -\lambda_{N,k+2} - \lambda_{k+2,k+1} & \lambda_{k+3,k+2} & 0 & \cdots & 0 & 0 & 0 \big]^\intercal \\
\mathbf{a}_3 & = \big[& -\lambda_{N,k+3} & -\lambda_{N,k+3} - \lambda_{k+3,k+2} & \lambda_{k+4,k+3} & \cdots & 0 & 0 & 0 \big]^\intercal \\
\vdots & & & & \vdots & & & & \\
\mathbf{a}_{N-k-2} & = \big[& -\lambda_{N,N-2} & -\lambda_{N,N-2} & -\lambda_{N,N-2} & \cdots & -\lambda_{N,N-2} - \lambda_{N-2,N-3} & \lambda_{N-1,N-2} & 0 \big]^\intercal
\end{array}$$

so that if we define $\mathbf{h} = \begin{bmatrix} h_{k+1,k} & \cdots & h_{N-1,k} \end{bmatrix}^\intercal$, then

$$\frac{s'_{\ell,k}}{2} = \mathbf{a}_{\ell-k}^\intercal \mathbf{h} + \frac{t_{\ell,k}}{2} \tag{61}$$

for $\ell = k+1, \dots, N-2$. More precisely,

$$\begin{aligned}
\mathbf{a}_1 &= \lambda_{k+2,k+1} \mathbf{e}_1 \\
\mathbf{a}_{\ell-k} &= -\lambda_{N,\ell} \sum_{i=1}^{\ell-k-1} \mathbf{e}_i - \lambda_{\ell,\ell-1} \mathbf{e}_{\ell-k-1} + \lambda_{\ell+1,\ell} \mathbf{e}_{\ell-k}, \quad \ell = k+2, \dots, N-2
\end{aligned} \tag{62}$$

where $\mathbf{e}_1, \dots, \mathbf{e}_{N-k-2} \in \mathbb{R}^{N-k-1}$ are elementary basis vectors. Finally, define

$$\mathbf{a}_{N-k-1} = \sum_{i=1}^{N-k-1} \mathbf{e}_i = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$$

so that

$$\frac{s_{k,k}}{2\lambda_{N,k}} = \sum_{i=k+1}^{N-1} h_{i,k} + h_{k,k} - \frac{q_k}{2} = \mathbf{a}_{N-k-1}^\intercal \mathbf{h} + h_{k,k} - \frac{q_k}{2}. \tag{63}$$

Now, by defining the matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_{N-k-1} \end{bmatrix}^\intercal \in \mathbb{R}^{(N-k-1)\times(N-k-1)}$$

we can write

$$\begin{bmatrix} s_{k,k} \\ s_{k+1,k} \\ \vdots \\ s_{N-2,k} \end{bmatrix} = 0 \iff \frac{1}{2}\begin{bmatrix} s'_{k+1,k} \\ \vdots \\ s'_{N-2,k} \\ s_{k,k}/\lambda_{N,k} \end{bmatrix} = 0 \iff \mathbf{A}\mathbf{h} = \frac{1}{2}\begin{bmatrix} -t_{k+1,k} \\ \vdots \\ -t_{N-2,k} \\ q_k - 2h_{k,k} \end{bmatrix}.$$

Note that $\mathbf{A}$ is lower-triangular:

$$\mathbf{A} = \begin{bmatrix} \lambda_{k+2,k+1} & 0 & \cdots & 0 & 0 \\ -\lambda_{N,k+2} - \lambda_{k+2,k+1} & \lambda_{k+3,k+2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\lambda_{N,N-2} & -\lambda_{N,N-2} & \cdots & \lambda_{N-1,N-2} & 0 \\ 1 & 1 & \cdots & 1 & 1 \end{bmatrix}$$

Therefore, it is invertible provided that all diagonal entries $\lambda_{k+2,k+1}, \dots, \lambda_{N-1,N-2}$ are positive. Because all entries of $\mathbf{A}$ and $t_{k+1,k}, \dots, t_{N-2,k}$ are expressed as continuous functions of $h_{1,1}, \dots, h_{N-1,N-1}$ (because λ is defined by (52) and q_ℓ's are defined by (55)), applying Cramer's rule to the above system, we can write $h_{k+1,k}, \dots, h_{N-1,k}$ as continuous functions of $h_{k,k}, \dots, h_{N-1,N-1}$.

Claim 2. Assuming that the solutions of (59) also satisfy $s_{N-1,k} = 0$, we have $s_{N,k} = 0$.

Recall that

$$\begin{aligned} s_{N-1,k} &= 2\lambda_{N,k}h_{N-1,N-1} - 2\lambda_{N-1,N-2}h_{N-2,k} + 2\lambda_{N,N-1}h_{N-1,k} \\ s_{N,k} &= 2\lambda_{N,k} - \sum_{\ell=k}^{N-1}\left(1 + \sum_{j=1}^{\ell}\lambda_{N,j}\right)2h_{\ell,k}. \end{aligned} \tag{64}$$

Assume $s_{N-1,k} = 0$ (as well as $s_{\ell,k} = 0$ for $\ell = k, \dots, N-2$). Then

$$\begin{aligned} 0 = s'_{N-1,k} &:= s_{N-1,k} - \frac{\lambda_{N,N-1}}{\lambda_{N,k}}s_{k,k} \\ &= 2\lambda_{N,k}h_{N-1,N-1} + 2\lambda_{N,N-1}h_{N-1,k} - 2\lambda_{N-1,N-2}h_{N-2,k} - \lambda_{N,N-1}\left(\sum_{i=k+1}^{N-1} 2h_{i,k} + 2h_{k,k} - q_k\right) \\ &= -\lambda_{N,N-1}\sum_{i=k+1}^{N-3} 2h_{i,k} - (\lambda_{N,N-1} + \lambda_{N-1,N-2})2h_{N-2,k} + \underbrace{2\lambda_{N,k}h_{N-1,N-1} + \lambda_{N,N-1}(q_k - 2h_{k,k})}_{:= t_{N-1,k} = \text{ Terms not depending on } h_{k+1,k},\dots,h_{N-1,k}}. \end{aligned}$$

Now sum up all $s'_{\ell,k}$'s for $\ell = k+1, \dots, N-1$ (up to scalar multiplication by 2, the coefficients of $h_{k+1,k}, \dots, h_{N-1,k}$ correspond to $\mathbf{a}_1 + \cdots + \mathbf{a}_{N-k-2}$ (recall (62)) plus the vector $\begin{bmatrix} -\lambda_{N,N-1} & \cdots & -\lambda_{N,N-1} & -\lambda_{N,N-1} - \lambda_{N-1,N-2} & 0 \end{bmatrix}$) to obtain

$$\begin{aligned} 0 &= \sum_{\ell=k+1}^{N-1} s'_{\ell,k} \\ &= \sum_{\ell=k+1}^{N-1}\left(-\sum_{i=\ell+1}^{N-1}\lambda_{N,i}\right)2h_{\ell,k} + \sum_{\ell=k+1}^{N-1} t_{\ell,k}. \end{aligned}$$

Using the fact $1 + \sum_{i=1}^{N-1} \lambda_{N,i} = N$ (from Proposition E.1), if we add

$$\begin{aligned} 0 = \frac{N}{\lambda_{N,k}} s_{k,k} &= \left(1 + \sum_{i=1}^{N-1} \lambda_{N,i}\right) \left(\sum_{\ell=k+1}^{N-1} 2h_{\ell,k} + 2h_{k,k} - q_k\right) \\ &= \sum_{\ell=k+1}^{N-1} \left(1 + \sum_{i=1}^{N-1} \lambda_{N,i}\right) 2h_{\ell,k} + N(2h_{k,k} - q_k) \end{aligned}$$

to both sides, we get

$$\begin{aligned} 0 &= \sum_{\ell=k+1}^{N-1} \left(1 + \sum_{i=1}^{\ell} \lambda_{N,i}\right) 2h_{\ell,k} + N(2h_{k,k} - q_k) + \sum_{\ell=k+1}^{N-1} t_{\ell,k} \\ &\overset{(64)}{=} -s_{N,k} + 2\lambda_{N,k} - \left(1 + \sum_{i=1}^{k} \lambda_{N,i}\right) 2h_{k,k} + N(2h_{k,k} - q_k) + \sum_{\ell=k+1}^{N-1} t_{\ell,k} \\ &= -s_{N,k} + \underbrace{2\lambda_{N,k} + \left(\sum_{i=k+1}^{N-1} \lambda_{N,i}\right) 2h_{k,k} - Nq_k + \sum_{\ell=k+1}^{N-1} t_{\ell,k}}_{:= t_{N,k} = \text{Terms not depending on } h_{k+1,k}, \ldots, h_{N-1,k}}. \end{aligned} \tag{65}$$

We show that $t_{N,k} = 0$, which then implies $s_{N,k} = 0$. We expand and rearrange the summation $\sum_{\ell=k+1}^{N-1} t_{\ell,k}$ as following:

$$\begin{aligned} \sum_{\ell=k+1}^{N-1} t_{\ell,k} &= t_{k+1,k} + \sum_{\ell=k+2}^{N-2} t_{\ell,k} + t_{N-1,k} \\ &= \lambda_{N,k} q_{k+1} + 2\lambda_{k+1,k}(1 - h_{k,k}) + \lambda_{N,k+1}(q_k - 2h_{k,k}) + \sum_{\ell=k+2}^{N-2} (\lambda_{N,\ell}(q_k - 2h_{k,k}) + \lambda_{N,k} q_\ell) \\ &\quad + 2\lambda_{N,k} h_{N-1,N-1} + \lambda_{N,N-1}(q_k - 2h_{k,k}) \\ &= \lambda_{N,k} \left(\sum_{\ell=k+1}^{N-2} q_\ell + 2h_{N-1,N-1}\right) + 2\lambda_{k+1,k}(1 - h_{k,k}) + \left(\sum_{\ell=k+1}^{N-1} \lambda_{N,\ell}\right)(q_k - 2h_{k,k}). \end{aligned} \tag{66}$$

Because of the definition (55) of q_ℓ, their sum telescopes:

$$\begin{aligned} \sum_{\ell=k+1}^{N-2} q_\ell + 2h_{N-1,N-1} &= \left(\sum_{\ell=k+1}^{N-2} 1 - (N-\ell)p_{\ell+1} + (N-\ell+1)p_\ell\right) + 2h_{N-1,N-1} \\ &= (N-k-2) - 2p_{N-1} + (N-k)p_{k+1} + 2h_{N-1,N-1} \\ &= (N-k-2) + (N-k)p_{k+1} \end{aligned} \tag{67}$$

where the last line follows from $p_{N-1} = \prod_{\ell=N-1}^{N-1} h_{\ell,\ell} = h_{N-1,N-1}$. Now plugging in the expression (66) into (65) and applying simplification by (67) gives

$$t_{N,k} = \underbrace{\lambda_{N,k}\left((N-k) + (N-k)p_{k+1}\right)}_{\text{(I)}} + \underbrace{2\lambda_{k+1,k}(1 - h_{k,k})}_{\text{(II)}} + \underbrace{\left(\sum_{\ell=k+1}^{N-1} \lambda_{N,\ell} - N\right) q_k}_{\text{(III)}}.$$

We expand each term (I), (II), (III): first, plugging in the expression for $\lambda_{N,k}$ from (52), we obtain

$$\begin{aligned}
\text{(I)} &= (N-k)\lambda_{N,k}\,(1+p_{k+1}) \\
&= \left(\frac{N}{N-k-1} - \frac{N(N-k)}{N-k-1}p_{k+1} + Np_k\right)(1+p_{k+1}) \\
&= \frac{N}{N-k-1} - \frac{N(N-k)}{N-k-1}p_{k+1} + Np_k + \frac{N}{N-k-1}p_{k+1} - \frac{N(N-k)}{N-k-1}p_{k+1}^2 + Np_kp_{k+1} \\
&= \frac{N}{N-k-1} + Np_k - Np_{k+1} + Np_kp_{k+1} - \frac{N(N-k)}{N-k-1}p_{k+1}^2. \qquad (68)
\end{aligned}$$

Next, again from (52), plug in the expression for $\lambda_{k+1,k}$ into (II) to get

$$\begin{aligned}
\text{(II)} &= \frac{2N}{N-k-1}p_{k+1}\left((N-k)p_{k+1}-1\right)(1-h_{k,k}) \\
&= \frac{2N(N-k)}{N-k-1}p_{k+1}^2 - \frac{2N(N-k)}{N-k-1}p_{k+1}^2h_{k,k} - \frac{2N}{N-k-1}p_{k+1} + \frac{2N}{N-k-1}p_{k+1}h_{k,k} \\
&= \frac{2N}{N-k-1}p_k - \frac{2N}{N-k-1}p_{k+1} - \frac{2N(N-k)}{N-k-1}p_kp_{k+1} + \frac{2N(N-k)}{N-k-1}p_{k+1}^2 \qquad (69)
\end{aligned}$$

where in the last line, we use $p_k = \prod_{\ell=k}^{N-1} h_{\ell,\ell} = h_{k,k}\prod_{\ell=k+1}^{N-1} h_{\ell,\ell} = p_{k+1}h_{k,k}$. Finally, applying Proposition E.1, we rewrite (III) as

$$\begin{aligned}
\text{(III)} &= \left(\frac{N(N-k-2)}{N-k-1} + \frac{N}{N-k-1}p_{k+1} - N\right)q_k \\
&= \left(-\frac{N}{N-k-1} + \frac{N}{N-k-1}p_{k+1}\right)\left(1-(N-k)p_{k+1} + (N-k+1)p_k\right) \\
&= -\frac{N}{N-k-1} + \frac{N}{N-k-1}p_{k+1} + \frac{N(N-k)}{N-k-1}p_{k+1} - \frac{N(N-k)}{N-k-1}p_{k+1}^2 - \frac{N(N-k+1)}{N-k-1}p_k + \frac{N(N-k+1)}{N-k-1}p_kp_{k+1} \\
&= -\frac{N}{N-k-1} - \frac{N(N-k+1)}{N-k-1}p_k + \frac{N(N-k+1)}{N-k-1}p_{k+1} + \frac{N(N-k+1)}{N-k-1}p_kp_{k+1} - \frac{N(N-k)}{N-k-1}p_{k+1}^2. \qquad (70)
\end{aligned}$$

Summing up the expressions (68), (69), (70), the coefficients of $p_k, p_{k+1}, p_kp_{k+1}, p_{k+1}^2$ terms and the constant term all vanish, and we conclude

$$t_{N,k} = \text{(I)} + \text{(II)} + \text{(III)} = 0.$$

Claim 3. The solutions of (59) satisfy $s_{N-1,k} = 0$.

To prove this, we first have to match the coefficients of $h_{k+1,k}, \ldots, h_{N-1,k}$ within $s_{N-1,k}$ via linear combination of rows of $\mathbf{A}$ through Gaussian elimination. The following proposition and its proof outlines this process.

Proposition E.4. *The following identity holds:*

$$s_{N-1,k} - 2\lambda_{N,k}h_{N-1,N-1} = \lambda_{N,N-1}\left(\frac{1}{\lambda_{N,k}}s_{k,k} - 2h_{k,k} + q_k\right) - \sum_{\ell=k+1}^{N-2}\frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}}\left(s'_{\ell,k} - t_{\ell,k}\right) \qquad (71)$$

where the sequences $\{a_\ell, b_\ell\}_{\ell=k+1}^{N-1}$ are defined by $a_{N-1} = \lambda_{N,N-1} + \lambda_{N-1,N-2}, b_{N-1} = \lambda_{N,N-1}$ and the reverse recursion

$$\begin{aligned}
a_\ell &= b_{\ell+1} + \frac{\lambda_{N,\ell} + \lambda_{\ell,\ell-1}}{\lambda_{\ell+1,\ell}}a_{\ell+1} \\
b_\ell &= b_{\ell+1} + \frac{\lambda_{N,\ell}}{\lambda_{\ell+1,\ell}}a_{\ell+1} \qquad (72)
\end{aligned}$$

for $\ell = N-1, \ldots, k+1$.

Proof. Define $\mathbf{v} = \begin{bmatrix} 0 & \cdots & 0 & -\lambda_{N-1,N-2} & \lambda_{N,N-1} \end{bmatrix}^\intercal$, so that

$$s_{N-1,k} - 2\lambda_{N,k} h_{N-1,N-1} = -2\lambda_{N-1,N-2} h_{N-2,k} + 2\lambda_{N,N-1} h_{N-1,k} = 2\mathbf{v}^\intercal \mathbf{h}$$

Recall the vector identities (61), (63)

$$s'_{\ell,k} - t_{\ell,k} = 2\mathbf{a}_{\ell-k}^\intercal \mathbf{h}$$
$$\frac{1}{\lambda_{N,k}} s_{k,k} - 2h_{k,k} + q_k = 2\mathbf{a}_{N-k-1}^\intercal \mathbf{h}$$

where $\mathbf{a}_1^\intercal, \dots, \mathbf{a}_{N-k-1}^\intercal$ are row vectors of $\mathbf{A}$. Because $\mathbf{A}$ is lower-triangular, we can express $\mathbf{v}$ as a linear combination of $\mathbf{a}_j$ by matching the entries from backwards. We start with

$$\mathbf{v}_{N-1} := \mathbf{v} - \lambda_{N,N-1}\mathbf{a}_{N-k-1} = \begin{bmatrix} -\lambda_{N,N-1} & -\lambda_{N,N-1} & \cdots & -\lambda_{N,N-1} & -\lambda_{N,N-1} - \lambda_{N-1,N-2} & 0 \end{bmatrix}^\intercal.$$

We recursively define a sequence of vectors $\mathbf{v}_\ell$ ($\ell = N-2, \dots, k+1$) by adding a constant multiple of $\mathbf{a}_{\ell-k}$ from $\mathbf{v}_{\ell+1}$ which eliminates the $(\ell-k)$th entry of $\mathbf{v}_{\ell+1}$. We denote the value of the $(\ell-k)$th entry of $\mathbf{v}_{\ell+1}$ by $-a_{\ell+1}$, and the common value of the remaining entries by $-b_{\ell+1}$, so:

$$a_{N-1} = \lambda_{N,N-1} + \lambda_{N-1,N-2}$$
$$b_{N-1} = \lambda_{N,N-1}.$$

Note that for $\ell = N-2, \dots, k+1$, the last nonzero entry of $\mathbf{a}_{\ell-k}$ is its $(\ell-k)$th entry $\lambda_{\ell+1,\ell}$. Hence, to cancel out the $(\ell-k)$th entry $-a_{\ell+1}$ of $\mathbf{v}_{\ell+1}$, we proceed as:

$$\begin{aligned}
\mathbf{v}_\ell &= \mathbf{v}_{\ell+1} + \frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}} \mathbf{a}_{\ell-k} \\
&= \begin{bmatrix} -b_{\ell+1} & \cdots & -b_{\ell+1} & -b_{\ell+1} & -a_{\ell+1} & 0 & \cdots & 0 \end{bmatrix}^\intercal \\
&\quad + \frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}} \begin{bmatrix} -\lambda_{N,\ell} & \cdots & -\lambda_{N,\ell} & -\lambda_{N,\ell} - \lambda_{\ell,\ell-1} & \lambda_{\ell+1,\ell} & 0 & \cdots & 0 \end{bmatrix}^\intercal \\
&= \begin{bmatrix} -b_{\ell+1} - \frac{\lambda_{N,\ell}}{\lambda_{\ell+1,\ell}} a_{\ell+1} & \cdots & -b_{\ell+1} - \frac{\lambda_{N,\ell}}{\lambda_{\ell+1,\ell}} a_{\ell+1} & -b_{\ell+1} - \frac{\lambda_{N,\ell}+\lambda_{\ell,\ell-1}}{\lambda_{\ell+1,\ell}} a_{\ell+1} & 0 & 0 & \cdots & 0 \end{bmatrix}^\intercal.
\end{aligned}$$

Because $\mathbf{v}_\ell = \begin{bmatrix} -b_\ell & \cdots & -b_\ell & -a_\ell & 0 & 0 & \cdots & 0 \end{bmatrix}^\intercal$, we obtain the recursion

$$a_\ell = b_{\ell+1} + \frac{\lambda_{N,\ell} + \lambda_{\ell,\ell-1}}{\lambda_{\ell+1,\ell}} a_{\ell+1}, \quad b_\ell = b_{\ell+1} + \frac{\lambda_{N,\ell}}{\lambda_{\ell+1,\ell}} a_{\ell+1} \quad (\ell = N-1, \dots, k+2)$$

as defined in the statement of Proposition E.4. Repeating this process until we reach $\ell = k+1$, we arrive at the identity

$$\mathbf{v}_{k+1} = \mathbf{v} - \lambda_{N,N-1}\mathbf{a}_{N-k-1} + \sum_{\ell=k+1}^{N-2} \frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}} \mathbf{a}_{\ell-k} = 0.$$

This implies

$$\begin{aligned}
s_{N-1,k} - 2\lambda_{N,k} h_{N-1,N-1} &= 2\mathbf{v}^\intercal \mathbf{h} \\
&= \lambda_{N,N-1} \cdot 2\mathbf{a}_{N-k-1}^\intercal \mathbf{h} - \sum_{\ell=k+1}^{N-2} \frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}} 2\mathbf{a}_{\ell-k}^\intercal \mathbf{h} \\
&= \lambda_{N,N-1} \left(\frac{1}{\lambda_{N,k}} s_{k,k} - 2h_{k,k} + q_k \right) - \sum_{\ell=k+1}^{N-2} \frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}} \left(s'_{\ell,k} - t_{\ell,k} \right)
\end{aligned}$$

which is the desired conclusion. □

Proposition E.5. *The sequences a_ℓ, b_ℓ defined in Proposition E.4 have the closed-form expressions*

$$a_\ell = \frac{(N-\ell-1)! \cdot N(N-\ell+1) p_\ell \prod_{i=\ell}^{N-1} p_i}{\prod_{j=\ell+1}^{N-1} \left((N-j+1)p_j - 1 \right)}, \quad b_\ell = \frac{(N-\ell-1)! \cdot N \prod_{i=\ell}^{N-1} p_i}{\prod_{j=\ell+1}^{N-1} \left((N-j+1)p_j - 1 \right)} \tag{73}$$

for $\ell = N-1, \dots, k+1$.

Proof. We use induction on $\ell = N-1, \dots, k+1$. In the case $\ell = N-1$,

$$\begin{aligned} b_{N-1} &= \lambda_{N,N-1} = N h_{N-1,N-1} = N p_{N-1} \\ a_{N-1} &= \lambda_{N,N-1} + \lambda_{N-1,N-2} \\ &= N h_{N-1,N-1} + N p_{N-1}(2p_{N-1} - 1) \\ &= N p_{N-1} + N p_{N-1}(2p_{N-1} - 1) \\ &= 2N p_{N-1}^2 \end{aligned}$$

which is consistent with (73) (with the vacuous product in the denominators being 1). Now let $k+1 \le \ell \le N-2$ and assume that the formula (73) holds for $\ell+1$. Recall that

$$\begin{aligned} \lambda_{N,\ell} &= \frac{N}{(N-\ell)(N-\ell-1)} - \frac{N}{N-\ell-1} p_{\ell+1} + \frac{N}{N-\ell} p_\ell \\ \lambda_{\ell+1,\ell} &= \frac{N}{N-\ell-1} p_{\ell+1} \left((N-\ell)p_{\ell+1} - 1\right) \\ \lambda_{\ell,\ell-1} &= \frac{N}{N-\ell} p_\ell \left((N-\ell+1)p_\ell - 1\right). \end{aligned}$$

Rewriting

$$\lambda_{N,\ell} = -\frac{N}{(N-\ell)(N-\ell-1)} \left((N-\ell)p_{\ell+1} - 1\right) + \frac{N}{N-\ell} p_\ell,$$

we see that

$$\begin{aligned} \frac{\lambda_{N,\ell}}{\lambda_{\ell+1,\ell}} &= \frac{1}{\frac{N}{N-\ell-1} p_{\ell+1} \left((N-\ell)p_{\ell+1} - 1\right)} \left(-\frac{N}{(N-\ell)(N-\ell-1)} \left((N-\ell)p_{\ell+1} - 1\right) + \frac{N}{N-\ell} p_\ell \right) \\ &= -\frac{1}{(N-\ell)p_{\ell+1}} + \frac{(N-\ell-1)p_\ell}{(N-\ell)p_{\ell+1}\left((N-\ell)p_{\ell+1} - 1\right)}. \end{aligned}$$

Using the above formula and the induction hypothesis (73), we compute the term $b_\ell - b_{\ell+1} = \frac{\lambda_{N,\ell}}{\lambda_{\ell+1,\ell}} a_{\ell+1}$ which comes from the recursion (72):

$$\begin{aligned} \frac{\lambda_{N,\ell}}{\lambda_{\ell+1,\ell}} a_{\ell+1} &= \left(-\frac{1}{(N-\ell)p_{\ell+1}} + \frac{(N-\ell-1)p_\ell}{(N-\ell)p_{\ell+1}\left((N-\ell)p_{\ell+1} - 1\right)} \right) \frac{(N-\ell-2)! \cdot N(N-\ell)p_{\ell+1} \prod_{i=\ell+1}^{N-1} p_i}{\prod_{j=\ell+2}^{N-1} \left((N-j+1)p_j - 1\right)} \\ &= -\frac{(N-\ell-2)! \cdot N \prod_{i=\ell+1}^{N-1} p_i}{\prod_{j=\ell+2}^{N-1} \left((N-j+1)p_j - 1\right)} + \frac{(N-\ell-1)! \cdot N \prod_{i=\ell}^{N-1} p_i}{\prod_{j=\ell+1}^{N-1} \left((N-j+1)p_j - 1\right)} \\ &= -b_{\ell+1} + \frac{(N-\ell-1)! \cdot N \prod_{i=\ell}^{N-1} p_i}{\prod_{j=\ell+1}^{N-1} \left((N-j+1)p_j - 1\right)}. \end{aligned}$$

This proves the identity (73) for b_ℓ.

Similarly, we compute

$$\begin{aligned} \lambda_{N,\ell} + \lambda_{\ell,\ell-1} &= -\frac{N}{(N-\ell)(N-\ell-1)} \left((N-\ell)p_{\ell+1} - 1\right) + \frac{N}{N-\ell} p_\ell + \frac{N}{N-\ell} p_\ell ((N-\ell+1)p_\ell - 1) \\ &= -\frac{N}{(N-\ell)(N-\ell-1)} \left((N-\ell)p_{\ell+1} - 1\right) + \frac{N(N-\ell+1)}{N-\ell} p_\ell^2 \end{aligned}$$

so

$$\begin{aligned} \frac{\lambda_{N,\ell} + \lambda_{\ell,\ell-1}}{\lambda_{\ell+1,\ell}} &= \frac{1}{\frac{N}{N-\ell-1} p_{\ell+1} \left((N-\ell)p_{\ell+1} - 1\right)} \left(-\frac{N}{(N-\ell)(N-\ell-1)} \left((N-\ell)p_{\ell+1} - 1\right) + \frac{N(N-\ell+1)}{N-\ell} p_\ell^2 \right) \\ &= -\frac{1}{(N-\ell)p_{\ell+1}} + \frac{(N-\ell+1)(N-\ell-1)p_\ell^2}{(N-\ell)p_{\ell+1}\left((N-\ell)p_{\ell+1} - 1\right)}. \end{aligned}$$

Plugging this into the right hand side of the recursion $a_\ell - b_{\ell+1} = \frac{\lambda_{N,\ell}+\lambda_{\ell,\ell-1}}{\lambda_{\ell+1,\ell}} a_{\ell+1}$ we see that

$$\begin{aligned}
&\frac{\lambda_{N,\ell}+\lambda_{\ell,\ell-1}}{\lambda_{\ell+1,\ell}} a_{\ell+1} \\
&= \left(-\frac{1}{(N-\ell)p_{\ell+1}} + \frac{(N-\ell+1)(N-\ell-1)p_\ell^2}{(N-\ell)p_{\ell+1}\left((N-\ell)p_{\ell+1}-1\right)} \right) \frac{(N-\ell-2)!\cdot N(N-\ell)p_{\ell+1}\prod_{i=\ell+1}^{N-1} p_i}{\prod_{j=\ell+2}^{N-1}\left((N-j+1)p_j-1\right)} \\
&= -b_{\ell+1} + \frac{(N-\ell-1)!\cdot N(N-\ell+1)p_\ell \prod_{i=\ell}^{N-1} p_i}{\prod_{j=\ell+1}^{N-1}\left((N-j+1)p_j-1\right)}
\end{aligned}$$

which proves the identity (73) for a_ℓ. This completes the induction. □

From (71) we have

$$\begin{aligned}
s_{N-1,k} = \frac{\lambda_{N,N-1}}{\lambda_{N,k}} s_{k,k} - \sum_{\ell=k+1}^{N-2} \frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}} s'_{\ell,k} \\
\underbrace{+\, 2\lambda_{N,k} h_{N-1,N-1} + \lambda_{N,N-1}(q_k - 2h_{k,k}) + \sum_{\ell=k+1}^{N-2} \frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}} t_{\ell,k}}_{:= r_{N-1,k}} .
\end{aligned}$$

Because $s_{\ell,k} = 0$ for $\ell = k+1, \ldots, N-2$, if we prove $r_{N-1,k} = 0$ then we are done. Plugging the definition (60) of $t_{\ell,k}$'s into $r_{N-1,k}$ we get

$$\begin{aligned}
&r_{N-1,k} \\
&= 2\lambda_{N,k} h_{N-1,N-1} + \lambda_{N,N-1}(q_k - 2h_{k,k}) + \frac{a_{k+2}}{\lambda_{k+2,k+1}} \left(\lambda_{N,k} q_{k+1} + 2\lambda_{k+1,k}(1-h_{k,k}) + \lambda_{N,k+1}(q_k - 2h_{k,k}) \right) \\
&\quad + \sum_{\ell=k+2}^{N-2} \frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}} \left(\lambda_{N,\ell}(q_k - 2h_{k,k}) + \lambda_{N,k} q_\ell \right) \\
&= 2\lambda_{N,k} h_{N-1,N-1} + \left(\lambda_{N,N-1} + \sum_{\ell=k+1}^{N-2} \frac{\lambda_{N,\ell}}{\lambda_{\ell+1,\ell}} a_{\ell+1} \right) (q_k - 2h_{k,k}) + \frac{2\lambda_{k+1,k}}{\lambda_{k+2,k+1}} a_{k+2}(1-h_{k,k}) \\
&\quad + \lambda_{N,k} \sum_{\ell=k+1}^{N-2} \frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}} q_\ell .
\end{aligned} \tag{74}$$

From the proof of Proposition E.5, we have

$$\frac{\lambda_{N,\ell}}{\lambda_{\ell+1,\ell}} a_{\ell+1} = b_\ell - b_{\ell+1},$$

for $\ell = k+1, \ldots, N-2$, so by telescoping, we can simplify:

$$\lambda_{N,N-1} + \sum_{\ell=k+1}^{N-2} \frac{\lambda_{N,\ell}}{\lambda_{\ell+1,\ell}} a_{\ell+1} = \lambda_{N,N-1} + b_{k+1} - b_{N-1} = b_{k+1} \tag{75}$$

Next, observe that

$$q_\ell = -((N-\ell)p_{\ell+1} - 1) + (N-\ell+1)p_\ell$$

so that

$$\begin{aligned}
\frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}}q_\ell &= \left[\frac{1}{\frac{N}{N-\ell-1}p_{\ell+1}\left((N-\ell)p_{\ell+1}-1\right)}\left(-((N-\ell)p_{\ell+1}-1)+(N-\ell+1)p_\ell\right)\right]a_{\ell+1} \\
&= \left[-\frac{N-\ell-1}{Np_{\ell+1}}+\frac{(N-\ell+1)(N-\ell-1)p_\ell}{Np_{\ell+1}((N-\ell)p_{\ell+1}-1)}\right]\frac{(N-\ell-2)!\cdot N(N-\ell)p_{\ell+1}\prod_{i=\ell+1}^{N-1}p_i}{\prod_{j=\ell+2}^{N-1}\left((N-j+1)p_j-1\right)} \\
&= -\frac{(N-\ell)!\prod_{i=\ell+1}^{N-1}p_i}{\prod_{j=\ell+2}^{N-1}\left((N-j+1)p_j-1\right)}+\frac{(N-\ell+1)!\prod_{i=\ell}^{N-1}p_i}{\prod_{j=\ell+1}^{N-1}\left((N-j+1)p_j-1\right)} \\
&= -\frac{(N-\ell)(N-\ell-1)b_{\ell+1}}{N}+\frac{(N-\ell+1)(N-\ell)b_\ell}{N}
\end{aligned}$$

which allows to simplify the summation via telescoping:

$$\sum_{\ell=k+1}^{N-2}\frac{a_{\ell+1}}{\lambda_{\ell+1,\ell}}q_\ell = -2h_{N-1,N-1}+\frac{(N-k)(N-k-1)b_{k+1}}{N} \tag{76}$$

Plugging (75) and (76) into (74), we obtain

$$\begin{aligned}
r_{N-1,k} = 2\lambda_{N,k}h_{N-1,N-1}+b_{k+1}(q_k-2h_{k,k})+\frac{2\lambda_{k+1,k}}{\lambda_{k+2,k+1}}a_{k+2}(1-h_{k,k}) \\
+\lambda_{N,k}\left(-2h_{N-1,N-1}+\frac{(N-k)(N-k-1)b_{k+1}}{N}\right)
\end{aligned} \tag{77}$$

The terms $\pm 2\lambda_{N,k}h_{N-1,N-1}$ in (77) cancel out each other. Now we plug in the identity

$$\begin{aligned}
\frac{\lambda_{k+1,k}}{\lambda_{k+2,k+1}}a_{k+2} &= \frac{\frac{N}{N-k-1}p_{k+1}((N-k)p_{k+1}-1)}{\frac{N}{N-k-2}p_{k+2}((N-k-1)p_{k+2}-1)}\frac{(N-k-3)!\cdot N(N-k-1)p_{k+2}\prod_{i=k+2}^{N-1}p_i}{\prod_{j=k+3}^{N-1}((N-j+1)p_j-1)} \\
&= \frac{(N-k-2)!\cdot N\prod_{i=k+1}^{N-1}p_i}{\prod_{j=k+2}^{N-1}((N-j+1)p_j-1)}((N-k)p_{k+1}-1) \\
&= b_{k+1}((N-k)p_{k+1}-1)
\end{aligned}$$

into (77) to obtain

$$\begin{aligned}
r_{N-1,k} &= b_{k+1}(q_k-2h_{k,k})+2b_{k+1}((N-k)p_{k+1}-1)(1-h_{k,k})+\lambda_{N,k}\frac{(N-k)(N-k-1)b_{k+1}}{N} \\
&= b_{k+1}\left[q_k-2h_{k,k}+2((N-k)p_{k+1}-1)(1-h_{k,k})+\frac{(N-k)(N-k-1)\lambda_{N,k}}{N}\right].
\end{aligned}$$

Finally, from the identity

$$\begin{aligned}
&q_k-2h_{k,k}+2((N-k)p_{k+1}-1)(1-h_{k,k}) \\
&= q_k-2h_{k,k}+2(N-k)p_{k+1}-2(N-k)p_{k+1}h_{k,k}-2+2h_{k,k} \\
&= 1-(N-k)p_{k+1}+(N-k+1)p_k+2(N-k)p_{k+1}-2(N-k)p_k-2 \\
&= -1+(N-k)p_{k+1}-(N-k-1)p_k \\
&= -\frac{(N-k)(N-k-1)}{N}\lambda_{N,k}
\end{aligned}$$

we conclude that $r_{N-1,k}=0$. This completes the proof of Proposition E.3.

□

Proposition E.3, together with Proposition E.2, implies that we can inductively solve the systems $s_{\ell,k}=0$ $(\ell=k,\dots,N)$ from $k=N-2$ to $k=1$ to determine the solution $h^*_{i,k}$'s with $N-1\geq i\geq k$ as continuous functions of $h_{1,1},\dots,h_{N-1,N-1}$. This completes Step 4 within the proof outline of Appendix E.1, and together with the remaining arguments from Appendix E.1, finally proves Theorem 4.1.

F. Remaining proofs and details for Section 4.2

F.1. Proof of Proposition 4.2

For simplicity, write $(H_{\text{OHM}})_{k,j} = h^{(0)}_{k,j}$ and $(H_{\text{Dual-OHM}})_{k,j} = h^{(1)}_{k,j}$ throughout this section. In the proof of Theorem 4.1 in Appendix E, we have seen that

$$p^{(0)}_k = \prod_{\ell=k}^{N-1} h^{(0)}_{\ell,\ell} = \frac{k}{N}, \quad p^{(1)}_k = \prod_{\ell=k}^{N-1} h^{(1)}_{\ell,\ell} = \frac{1}{N-k+1}$$

so that

$$u := \left(p^{(0)}_2, \dots, p^{(0)}_{N-1}\right) \in \partial C, \quad v := \left(p^{(1)}_2, \dots, p^{(1)}_{N-1}\right) \in \partial C.$$

Recall that for $w = (p_2, \dots, p_{N-1}) \in C$,

$$\Phi(w) = \begin{bmatrix} \frac{1}{Np_2} & 0 & \cdots & 0 \\ h^*_{2,1}\left(\frac{1}{Np_2}, \frac{p_2}{p_3}, \dots, p_{N-1}\right) & \frac{p_2}{p_3} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h^*_{N-1,1}\left(\frac{1}{Np_2}, \frac{p_2}{p_3}, \dots, p_{N-1}\right) & h^*_{N-1,2}\left(\frac{1}{Np_2}, \frac{p_2}{p_3}, \dots, p_{N-1}\right) & \cdots & p_{N-1} \end{bmatrix}.$$

Clearly, we can directly extend the definition of Φ to u, v for diagonal entries while preserving continuity. For non-diagonal entries, however, we have to check whether $h^*_{i,k}$'s with $i > k$, which are defined as solutions of linear systems, converge to $h^{(m)}_{i,k}$ as $w \to \left(p^{(m)}_2, \dots, p^{(m)}_{N-1}\right)$, for $m = 0, 1$. In Appendix E, we first characterize $h^*_{i,k}$ as solutions to the set of equations $s_{\ell,k} = 0$ ($\ell = k, \dots, N-2$) which is equivalent to a linear system determined by the invertible lower-triangular matrix $\mathbf{A}$ defined therein, and then show that those solutions also satisfy $s_{N-1,k} = 0 = s_{N,k}$. However, within the process of rewriting the system $s_{\ell,k} = 0$ ($\ell = k, \dots, N-2$) in terms of $\mathbf{A}$ we use division by $\lambda_{N,k}$ (so we are implicitly assuming $\lambda_{N,k} > 0$), and the resulting matrix $\mathbf{A}$ is not invertible if one of $\lambda_{j+1,j}$ ($j = k+1, \dots, N-2$) is zero. Because $\lambda_{N,k} = 0$ ($k = 1, \dots, N-2$) for OHM and $\lambda_{j+1,j} = 0$ ($j = 1, \dots, N-2$) for Dual-OHM, the same argument is not directly applicable to these cases.

Instead, we retreat to the original set of equations and consider the system $s_{\ell,k} = 0$ ($k = k+1, \dots, N-1$), which has the matrix form

$$\underbrace{\begin{bmatrix} \lambda_{k+2,k+1} + \lambda_{N,k+1} & \lambda_{N,k+1} & \cdots & \lambda_{N,k+1} & \lambda_{N,k+1} \\ -\lambda_{k+2,k+1} & \lambda_{k+3,k+2} + \lambda_{N,k+2} & \cdots & \lambda_{N,k+2} & \lambda_{N,k+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{N-1,N-2} + \lambda_{N,N-2} & \lambda_{N,N-2} \\ 0 & 0 & \cdots & -\lambda_{N-1,N-2} & \lambda_{N,N-1} \end{bmatrix}}_{:=\mathbf{B}} \underbrace{\begin{bmatrix} h_{k+1,k} \\ h_{k+2,k} \\ \vdots \\ h_{N-2,k} \\ h_{N-1,k} \end{bmatrix}}_{\mathbf{h}}$$

$$= \underbrace{\frac{1}{2}\begin{bmatrix} -2\lambda_{k+1,k}(1 - h_{k,k}) - \lambda_{N,k}(q_k - 2h_{k,k}) \\ -\lambda_{N,k} q_{k+2} \\ \vdots \\ -\lambda_{N,k} q_{N-2} \\ -2\lambda_{N,k} h_{N-1,N-1} \end{bmatrix}}_{:=\mathbf{b}}.$$

As we take $w \to u$ (u is the $(p_2, \dots, p_{N-1})$-coordinate for OHM), all $\lambda_{N,j}$ with $j = k+1, \dots, N-2$ vanishes, and $\mathbf{B}$ converges to the lower bidiagonal matrix

$$\mathbf{B}^{(0)} = \begin{bmatrix} \lambda^{(0)}_{k+2,k+1} & & & \\ -\lambda^{(0)}_{k+2,k+1} & \lambda^{(0)}_{k+3,k+2} & & \\ & \ddots & \ddots & \\ & & -\lambda^{(0)}_{N-1,N-2} & \lambda^{(0)}_{N,N-1} \end{bmatrix}$$

where $\lambda^{(0)}_{j+1,j} = \frac{j(j+1)}{N}$ denotes the limit of $\lambda_{j+1,j}$ as $w \to u$ (up to a constant, these are the weights multiplied to the monotonicity inequalities in the convergence proof of OHM). It is clear that $\mathbf{B}^{(0)}$ is invertible, and because λ's depend continuously on $p_2, \dots, p_{N-1}$, the matrix $\mathbf{B}$ is invertible in a neighborhood of u. Within this neighborhood, the solutions $h^*_{k+1,k}, \dots, h^*_{N-1,k}$ are determined via Cramer's rule applied to the system $\mathbf{Bh} = \mathbf{b}$. As all entries of $\mathbf{b}$ depend continuously on $p_2, \dots, p_{N-1}$ (because λ's and q_j's are), this proves that $h^*_{i,k} \to h^{(0)}_{i,k}$ as $w \to u$ for each $i = k+1, \dots, N-1$. (Note that $h^{(0)}_{i,k}$ must be the unique solutions satisfying $\mathbf{B}^{(0)}\mathbf{h} = \mathbf{b}^{(0)} = \lim_{w\to u} \mathbf{b}$, as OHM also satisfies the identity (41).)

Similarly, for the case of Dual-OHM, we observe that all $\lambda_{j+1,j}$ with $j = k+1, \dots, N-2$ vanish in the limit $w \to v$, so $\mathbf{B}$ converges to the upper triangular matrix

$$\mathbf{B}^{(1)} = \begin{bmatrix} \lambda^{(1)}_{N,k+1} & \lambda^{(1)}_{N,k+1} & \cdots & \lambda^{(1)}_{N,k+1} \\ & \lambda^{(1)}_{N,k+2} & \cdots & \lambda^{(1)}_{N,k+2} \\ & & \ddots & \vdots \\ & & & \lambda^{(1)}_{N,N-1} \end{bmatrix}$$

where $\lambda^{(1)}_{N,j} = \frac{N}{(N-j)(N-j+1)} = \lim_{w\to v} \lambda_{N,j}$ (as before, these are the weights multiplied to the monotonicity inequalities in the convergence proof of Dual-OHM up to a constant). Again, $\mathbf{B}^{(1)}$ is invertible and thus $h^*_{i,k}$ are determined via Cramer's rule applied to the system $\mathbf{Bh} = \mathbf{b}$ in a neighborhood of v, which shows that $h^*_{i,k} \to h^{(1)}_{i,k}$ as $w \to v$.

The above arguments altogether show that

$$\begin{aligned} \Phi(w) &\to H_{\text{OHM}} \quad \text{as} \quad w \to u \\ \Phi(w) &\to H_{\text{Dual-OHM}} \quad \text{as} \quad w \to v \end{aligned}$$

and thus concludes the proof of Proposition 4.2.

F.2. Explicit characterization of optimal algorithm family for the case $N = 3$

For a nonexpansive operator $\mathbb{T}$, we let $\mathbb{A} = 2(\mathbb{I} + \mathbb{T})^{-1} - \mathbb{I}$, $x_{k+1} = \mathbb{J}_{\mathbb{A}}(y_k)$ and $\tilde{\mathbb{A}}x_{k+1} = y_k - x_{k+1}$ as usual. Consider the algorithm defined by

$$\begin{aligned} y_1 &= y_0 - h_{1,1}(y_0 - \mathbb{T}y_0) \\ &= y_0 - 2h_{1,1}\tilde{\mathbb{A}}x_1 \\ y_2 &= y_1 - h_{2,1}(y_0 - \mathbb{T}y_0) - h_{2,2}(y_1 - \mathbb{T}y_1) \\ &= y_1 - 2h_{2,1}\tilde{\mathbb{A}}x_1 - 2h_{2,2}\tilde{\mathbb{A}}x_2 \end{aligned}$$

where $h_{1,1}, h_{2,2} \in \left[\frac{1}{2}, \frac{2}{3}\right]$, $h_{1,1}h_{2,2} = \frac{1}{3}$ and $h_{1,1} + h_{2,1} + h_{2,2} = 1$. Let

$$\lambda_{3,2} = 3h_{2,2}, \quad \lambda_{3,1} = 2 - 3h_{2,2}, \quad \lambda_{2,1} = 3h_{2,2}(2h_{2,2} - 1)$$

as defined in (52) (we substitute $N = 3$ and $p_2 = h_{2,2}$). Clearly, we have $\lambda_{3,2}, \lambda_{3,1}, \lambda_{2,1} \geq 0$ because $\frac{1}{2} \leq h_{2,2} \leq \frac{2}{3}$.

Now, let $g_k = \tilde{\mathbb{A}}x_k$ for $k = 1, 2, 3$. Then

$$\begin{aligned} x_2 - x_1 &= (y_1 - g_2) - (y_0 - g_1) = (y_1 - y_0) - (g_2 - g_1) = (1 - 2h_{1,1})g_1 - g_2 \\ x_3 - x_2 &= (y_2 - g_3) - (y_1 - g_2) = (y_2 - y_1) - (g_3 - g_2) = -2h_{2,1}g_1 + (1 - 2h_{2,2})g_2 - g_3 \\ x_3 - x_1 &= (x_3 - x_2) + (x_2 - x_1) = (1 - 2h_{1,1} - 2h_{2,1})g_1 - 2h_{2,2}g_2 - g_3 \\ x_3 - y_0 &= (x_3 - x_1) - (y_0 - x_1) = x_3 - x_1 - g_1 = -(2h_{1,1} + 2h_{2,1})g_1 - 2h_{2,2}g_2 - g_3 \end{aligned}$$

so we can expand

$$\begin{aligned}
&\left\langle \tilde{\mathbf{A}}x_3, x_3 - y_0\right\rangle + 3\left\|\tilde{\mathbf{A}}x_3\right\|^2 + \lambda_{2,1}\left\langle x_2 - x_1, \tilde{\mathbf{A}}x_2 - \tilde{\mathbf{A}}x_1\right\rangle + \lambda_{3,1}\left\langle x_3 - x_1, \tilde{\mathbf{A}}x_3 - \tilde{\mathbf{A}}x_1\right\rangle + \lambda_{3,2}\left\langle x_3 - x_2, \tilde{\mathbf{A}}x_3 - \tilde{\mathbf{A}}x_2\right\rangle \\
&= \left\langle g_3, x_3 - y_0\right\rangle + 3\left\|g_3\right\|^2 + \lambda_{2,1}\left\langle x_2 - x_1, g_2 - g_1\right\rangle + \lambda_{3,1}\left\langle x_3 - x_1, g_3 - g_1\right\rangle + \lambda_{3,2}\left\langle x_3 - x_2, g_3 - g_2\right\rangle \\
&= \left\langle g_3, -(2h_{1,1} + 2h_{2,1})g_1 - 2h_{2,2}g_2 - g_3\right\rangle + 3\left\|g_3\right\|^2 + \lambda_{2,1}\left\langle (1 - 2h_{1,1})g_1 - g_2, g_2 - g_1\right\rangle \\
&\quad + \lambda_{3,1}\left\langle (1 - 2h_{1,1} - 2h_{2,1})g_1 - 2h_{2,2}g_2 - g_3, g_3 - g_1\right\rangle + \lambda_{3,2}\left\langle -2h_{2,1}g_1 + (1 - 2h_{2,2})g_2 - g_3, g_3 - g_2\right\rangle \\
&= (\lambda_{2,1}(2h_{1,1} - 1) - \lambda_{3,1}(1 - 2h_{1,1} - 2h_{2,1}))\left\|g_1\right\|^2 + (\lambda_{2,1}(2 - 2h_{1,1}) + 2\lambda_{3,1}h_{2,2} + 2\lambda_{3,2}h_{2,1})\left\langle g_2, g_1\right\rangle \\
&\quad + (-(2h_{1,1} + 2h_{2,1}) + \lambda_{3,1}(2 - 2h_{1,1} - 2h_{2,1}) - 2\lambda_{3,2}h_{2,1})\left\langle g_3, g_1\right\rangle + (-\lambda_{2,1} + \lambda_{3,2}(2h_{2,2} - 1))\left\|g_2\right\|^2 \\
&\quad + (-2h_{2,2} - 2\lambda_{3,1}h_{2,2} + \lambda_{3,2}(2 - 2h_{2,2}))\left\langle g_3, g_2\right\rangle + (-1 + 3 - \lambda_{3,1} - \lambda_{3,2})\left\|g_3\right\|^2 \\
&= s_{1,1}\left\|g_1\right\|^2 + s_{2,1}\left\langle g_2, g_1\right\rangle + s_{3,1}\left\langle g_3, g_1\right\rangle + s_{2,2}\left\|g_2\right\|^2 + s_{3,2}\left\langle g_3, g_2\right\rangle + s_{3,3}\left\|g_3\right\|^2.
\end{aligned}$$

We verify:

$$\begin{aligned}
s_{3,3} &= 2 - \lambda_{3,1} - \lambda_{3,2} = 0 \\
s_{3,2} &= -2h_{2,2} - 2\lambda_{3,1}h_{2,2} + \lambda_{3,2}(2 - 2h_{2,2}) \\
&= -2h_{2,2} - 2(2 - 3h_{2,2})h_{2,2} + 3h_{2,2}(2 - 2h_{2,2}) = 0 \\
s_{2,2} &= -\lambda_{2,1} + \lambda_{3,2}(2h_{2,2} - 1) \\
&= -3h_{2,2}(2h_{2,2} - 1) + 3h_{2,2}(2h_{2,2} - 1) = 0.
\end{aligned}$$

Next,

$$\begin{aligned}
s_{3,1} &= -(2h_{1,1} + 2h_{2,1}) + \lambda_{3,1}(2 - 2h_{1,1} - 2h_{2,1}) - 2\lambda_{3,2}h_{2,1} \\
&= 2(h_{2,2} - 1) + (2 - 3h_{2,2})2h_{2,2} - 6h_{2,2}\left(1 - \frac{1}{3h_{2,2}} - h_{2,2}\right) = 0
\end{aligned}$$

where we plug in the definitions of $\lambda_{3,2}$ and eliminate $h_{2,1}, h_{1,1}$ using the conditions $h_{1,1}h_{2,2} = \frac{1}{3}$ and $h_{1,1} + h_{2,1} + h_{2,2} = 1$ to obtain the second equality. Similarly,

$$\begin{aligned}
s_{2,1} &= \lambda_{2,1}(2 - 2h_{1,1}) + 2\lambda_{3,1}h_{2,2} + 2\lambda_{3,2}h_{2,1} \\
&= 3h_{2,2}(2h_{2,2} - 1)\left(2 - \frac{2}{3h_{2,2}}\right) + 2(2 - 3h_{2,2})h_{2,2} + 6h_{2,2}\left(1 - \frac{1}{3h_{2,2}} - h_{2,2}\right) \\
&= (2h_{2,2} - 1)(6h_{2,2} - 2) + 2h_{2,2}(2 - 3h_{2,2}) + 6h_{2,2} - 2 - 6h_{2,2}^2 \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
s_{1,1} &= \lambda_{2,1}(2h_{1,1} - 1) - \lambda_{3,1}(1 - 2h_{1,1} - 2h_{2,1}) \\
&= 3h_{2,2}(2h_{2,2} - 1)\left(\frac{2}{3h_{2,2}} - 1\right) - (2 - 3h_{2,2})(2h_{2,2} - 1) \\
&= (2h_{2,2} - 1)(2 - 3h_{2,2}) - (2 - 3h_{2,2})(2h_{2,2} - 1) \\
&= 0.
\end{aligned}$$

This shows that

$$\begin{aligned}
0 &= \left\langle \tilde{\mathbf{A}}x_3, x_3 - y_0\right\rangle + 3\left\|\tilde{\mathbf{A}}x_3\right\|^2 + \lambda_{2,1}\left\langle x_2 - x_1, \tilde{\mathbf{A}}x_2 - \tilde{\mathbf{A}}x_1\right\rangle \\
&\quad + \lambda_{3,1}\left\langle x_3 - x_1, \tilde{\mathbf{A}}x_3 - \tilde{\mathbf{A}}x_1\right\rangle + \lambda_{3,2}\left\langle x_3 - x_2, \tilde{\mathbf{A}}x_3 - \tilde{\mathbf{A}}x_2\right\rangle \\
&\geq \left\langle \tilde{\mathbf{A}}x_3, x_3 - y_0\right\rangle + 3\left\|\tilde{\mathbf{A}}x_3\right\|^2,
\end{aligned}$$

which, together with Lemma 3.2, proves $\left\|\tilde{\mathbf{A}}x_3\right\|^2 \leq \frac{\|y_0 - y_\star\|^2}{9}$.

Note that our choice of $h_{2,1}$ is consistent with the formula (53) for $k = 1$, i.e., $2(h_{1,1} + h_{2,1}) = 2 - 2h_{2,2} = 1 - 2p_2 + 3p_1$. Indeed, the calculation above is just a concrete demonstration of the more complicated series of computations presented in Appendix E.

F.3. Examples of optimal algorithms not covered by Theorem 4.1

Let $N > 3$, and fix $2 \leq N' \leq N-1$. Consider the following algorithm that first runs N' steps of Dual-OHM (with $y_{N'-1}$ as terminal iterate) and then $N - N'$ steps of OHM (as if $y_{N'-1}$ was generated by usual OHM and then continuing on using the update rule of OHM):

$$\begin{aligned} y_{k+1} &= y_k + \frac{N'-k-1}{N'-k}(\mathbb{T}y_k - \mathbb{T}y_{k-1}) \quad \text{if } \; k = 0, \dots, N'-2 \\ y_{k+1} &= \frac{k+1}{k+2}\mathbb{T}y_k + \frac{1}{k+2}y_0 \qquad\qquad\quad \text{if } \; k = N'-1, \dots, N-2. \end{aligned} \tag{78}$$

We show that this algorithm also has the exact optimal rate

$$\|y_{N-1} - \mathbb{T}y_{N-1}\|^2 \leq \frac{4\|y_0 - y_\star\|^2}{N^2}.$$

First, because $y_{N'-1}$ is generated via Dual-OHM for $N'-1$ iterations, we have

$$\begin{aligned} &\left\langle \tilde{\mathbb{A}}x_{N'}, x_{N'} - y_0 \right\rangle + N' \left\|\tilde{\mathbb{A}}x_{N'}\right\|^2 \\ &\leq \left\langle \tilde{\mathbb{A}}x_{N'}, x_{N'} - y_0 \right\rangle + N' \left\|\tilde{\mathbb{A}}x_{N'}\right\|^2 + \sum_{j=1}^{N'-1} \frac{N'}{(N'-j)(N'-j+1)} \left\langle x_{N'} - x_j, \tilde{\mathbb{A}}x_{N'} - \tilde{\mathbb{A}}x_j \right\rangle \\ &= 0 \end{aligned}$$

where as usual, $\mathbb{A} = 2(\mathbb{I}+\mathbb{T})^{-1} - \mathbb{I}$, $x_{k+1} = \mathbb{J}_{\mathbb{A}}(y_k)$ and $\tilde{\mathbb{A}}x_{k+1} = y_k - x_{k+1}$. Now from the analysis of OHM, the identity

$$\underbrace{\left((k+1)^2 \left\|\tilde{\mathbb{A}}x_{k+1}\right\|^2 + (k+1)\left\langle \tilde{\mathbb{A}}x_{k+1}, x_{k+1} - y_0 \right\rangle\right)}_{U_{k+1}} - \underbrace{\left((k+2)^2 \left\|\tilde{\mathbb{A}}x_{k+2}\right\|^2 + (k+2)\left\langle \tilde{\mathbb{A}}x_{k+2}, x_{k+2} - y_0 \right\rangle\right)}_{U_{k+2}}$$
$$= (k+1)(k+2)\left\langle x_{k+2} - x_{k+1}, \tilde{\mathbb{A}}x_{k+2} - x_{k+1} \right\rangle$$

holds if the relationship

$$y_{k+1} = \frac{k+1}{k+2}\mathbb{T}y_k + \frac{1}{k+2}y_0 = \frac{k+1}{k+2}(y_k - 2\tilde{\mathbb{A}}x_{k+1}) + \frac{1}{k+2}y_0$$

holds (regardless of whether y_k is an iterate generated by OHM or not); see, e.g., Ryu & Yin (2022). Therefore, we have

$$\begin{aligned} &N^2 \left\|\tilde{\mathbb{A}}x_N\right\|^2 + N\left\langle \tilde{\mathbb{A}}x_N, x_N - y_0 \right\rangle = U_N \\ &\leq U_{N-1} \leq \cdots \leq U_{N'} = (N')^2 \left\|\tilde{\mathbb{A}}x_{N'}\right\|^2 + N'\left\langle \tilde{\mathbb{A}}x_{N'}, x_{N'} - y_0 \right\rangle \leq 0 \end{aligned}$$

and dividing both sides by N and applying Lemma 3.2 we obtain $\left\|\tilde{\mathbb{A}}x_N\right\|^2 \leq \frac{\|y_0 - y_\star\|^2}{N^2}$.

However, one can expect that the algorithm (78) will not belong to the optimal algorithm family in Theorem 4.1, because its convergence proof above uses the set of inequalities

$$\lambda_{i,j}\left\langle x_i - x_j, \tilde{\mathbb{A}}x_i - \tilde{\mathbb{A}}x_j \right\rangle$$

with $(i,j) \in \{(N',1), \dots, (N',N'-1)\} \cup \{(N'+1,N'), \dots, (N,N-1)\}$, which is different from the set $\{(N,1), \dots, (N,N-1)\} \cup \{(2,1), \dots, (N,N-1)\}$ used in Theorem 4.1. Indeed, consider the algorithm (78) in the case $N' = N-1$. Then the upper $(N-2)\times(N-2)$ submatrix of its H-matrix equals the H-matrix of Dual-OHM for with terminal iterate y_{N-2}. Next, it can be checked that the sum of each column in the H-matrix of Dual-OHM is always $\frac{1}{2}$

(regardless of the total iteration count), so that $y_{N-2} = y_0 - \tilde{\mathbb{A}}x_1 - \cdots - \tilde{\mathbb{A}}x_{N-2}$. Then we see that

$$\begin{aligned} y_{N-1} &= \frac{N-1}{N}\mathbb{T}y_{N-2} + \frac{1}{N}y_0 \\ &= \frac{N-1}{N}(y_{N-2} - 2\tilde{\mathbb{A}}x_{N-1}) + \frac{1}{N}y_0 \\ &= y_0 - \frac{N-1}{N}\left(\tilde{\mathbb{A}}x_1 + \cdots + \tilde{\mathbb{A}}x_{N-2}\right) - \frac{2(N-1)}{N}\tilde{\mathbb{A}}x_{N-1} \\ &= y_{N-2} + \frac{1}{N}\left(\tilde{\mathbb{A}}x_1 + \cdots + \tilde{\mathbb{A}}x_{N-2}\right) - \frac{2(N-1)}{N}\tilde{\mathbb{A}}x_{N-1}, \end{aligned}$$

and thus, the H-matrix of (78) with $N' = N-1$ is

$$H = \begin{bmatrix} \frac{N-2}{N-1} & & & \\ \vdots & \ddots & & \\ -\frac{1}{(N-1)(N-2)} & \cdots & \frac{1}{2} & \\ -\frac{1}{2N} & \cdots & -\frac{1}{2N} & \frac{N-1}{N} \end{bmatrix}.$$

For this H-matrix, we have

$$p_{N-1} = \frac{N-1}{N}, \quad p_{N-2} = \frac{N-1}{2N}, \quad \ldots, \quad p_2 = \frac{N-1}{(N-2)N}$$

and it can be checked that $w = (p_2, \ldots, p_{N-1}) \in C$. However, if $H = \Phi(w)$, then by construction of Φ (see (55), (56)) we must have

$$2(h_{N-1,N-2} + h_{N-2,N-2}) = 1 - 2p_{N-1} + 3p_{N-2}.$$

However, the left hand side is

$$2\left(\frac{1}{2} - \frac{1}{2N}\right) = \frac{N-1}{N},$$

while the right hand side is

$$1 - \frac{2(N-1)}{N} + \frac{3(N-1)}{2N} = \frac{N+1}{2N},$$

which is a contradiction. Therefore, we conclude that $H \neq \Phi(w)$.

G. Omitted details from Section 6

G.1. Lyapunov analysis of Dual-FEG

In this section, we denote $\mathbb{A}x = \nabla_{\pm}\mathbf{L}(x)$ for simplicity. Recall the update rules of Dual-FEG:

$$\begin{aligned} x_{k+1/2} &= x_k - \alpha z_k - \alpha \mathbb{A}x_k \\ x_{k+1} &= x_{k+1/2} - \frac{N-k-1}{N-k}\alpha\left(\mathbb{A}x_{k+1/2} - \mathbb{A}x_k\right) \\ z_{k+1} &= \frac{N-k-1}{N-k}z_k - \frac{1}{N-k}\mathbb{A}x_{k+1/2}, \end{aligned} \tag{Dual-FEG}$$

and the Lyapunov function

$$V_k = -\alpha\left\|z_k + \mathbb{A}x_N\right\|^2 + \frac{2}{N-k}\left\langle z_k + \mathbb{A}x_N, x_k - x_N\right\rangle$$

for $k = 1, \ldots, N-1$. We prove that

$$\begin{aligned} V_k - V_{k+1} = \frac{2}{(N-k)(N-k-1)}\underbrace{\left\langle \mathbb{A}x_N - \mathbb{A}x_{k+1/2}, x_N - x_{k+1/2}\right\rangle}_{\mathrm{MI}_k} \\ + \frac{1}{\alpha(N-k)^2}\underbrace{\left(\|x_{k+1/2} - x_k\|^2 - \alpha^2\|\mathbb{A}x_{k+1/2} - \mathbb{A}x_k\|^2\right)}_{\mathrm{LI}_k} \end{aligned}$$

for $k = 1, \ldots, N-2$. This implies $V_k \geq V_{k+1}$ because $\mathrm{MI}_k \geq 0$ by monotonicity of $\mathbb{A}$ and $\mathrm{LI}_k \geq 0$ from L-Lipschitz continuity of $\mathbb{A}$ (L-smoothness of $\mathbf{L}$) and $0 < \alpha \leq \frac{1}{L}$.

We first decompose V_k as following:

$$V_k = -\alpha\left\|z_k\right\|^2 - \alpha\left\|\mathbb{A}x_N\right\|^2 + \underbrace{\left(\frac{2}{N-k}\left\langle \mathbb{A}x_N, x_k - x_N\right\rangle - 2\alpha\left\langle \mathbb{A}x_N, z_k\right\rangle\right)}_{:=V_k^{(1)}} + \underbrace{\frac{2}{N-k}\left\langle z_k, x_k - x_N\right\rangle}_{:=V_k^{(2)}}.$$

Observe that $x_{k+1/2}$ can be written in the following two ways:

$$\begin{aligned} x_{k+1/2} &= x_k - \alpha z_k - \alpha \mathbb{A}x_k \\ x_{k+1/2} &= x_{k+1} - \frac{N-k-1}{N-k}\alpha(\mathbb{A}x_k - \mathbb{A}x_{k+1/2}). \end{aligned} \tag{79}$$

Appropriately plugging (79) into MI_k we can rewrite it as

$$\begin{aligned} \mathrm{MI}_k &= (N-k)\left\langle \mathbb{A}x_N, x_N - x_{k+1/2}\right\rangle - (N-k-1)\left\langle \mathbb{A}x_N, x_N - x_{k+1/2}\right\rangle - \left\langle \mathbb{A}x_{k+1/2}, x_N - x_{k+1/2}\right\rangle \\ &= (N-k)\left\langle \mathbb{A}x_N, x_N - x_{k+1} + \frac{N-k-1}{N-k}\alpha(\mathbb{A}x_k - \mathbb{A}x_{k+1/2})\right\rangle \\ &\quad - (N-k-1)\left\langle \mathbb{A}x_N, x_N - x_k + \alpha z_k + \alpha \mathbb{A}x_k\right\rangle - \left\langle \mathbb{A}x_{k+1/2}, x_N - x_{k+1/2}\right\rangle \\ &= (N-k)\left\langle \mathbb{A}x_N, x_N - x_{k+1}\right\rangle - (N-k-1)\left\langle \mathbb{A}x_N, x_N - x_k\right\rangle \\ &\quad - \alpha(N-k-1)\left\langle \mathbb{A}x_N, \mathbb{A}x_{k+1/2} + z_k\right\rangle - \left\langle \mathbb{A}x_{k+1/2}, x_N - x_{k+1/2}\right\rangle. \end{aligned}$$

Now multiply $\frac{2}{(N-k)(N-k-1)}$ to MI_k:

$$\begin{aligned} &\frac{2}{(N-k)(N-k-1)}\mathrm{MI}_k \\ &= \frac{2}{N-k-1}\left\langle \mathbb{A}x_N, x_N - x_{k+1}\right\rangle - \frac{2}{N-k}\left\langle \mathbb{A}x_N, x_N - x_k\right\rangle \\ &\quad - 2\alpha\left\langle \mathbb{A}x_N, \frac{1}{N-k}\mathbb{A}x_{k+1/2} + \frac{1}{N-k}z_k\right\rangle - \frac{2}{N-k-1}\left\langle \frac{1}{N-k}\mathbb{A}x_{k+1/2}, x_N - x_{k+1/2}\right\rangle \end{aligned} \tag{80}$$

and plug the following identity (which is the third line of Dual-FEG)

$$\frac{1}{N-k}\mathbb{A}x_{k+1/2} = \frac{N-k-1}{N-k}z_k - z_{k+1} \tag{81}$$

into (80) to obtain

$$\begin{aligned}
&\frac{2}{(N-k)(N-k-1)}\mathrm{MI}_k \\
&= \frac{2}{N-k-1}\left\langle \mathbb{A}x_N, x_N - x_{k+1}\right\rangle - \frac{2}{N-k}\left\langle \mathbb{A}x_N, x_N - x_k\right\rangle \\
&\quad - 2\alpha\left\langle \mathbb{A}x_N, z_k - z_{k+1}\right\rangle - \frac{2}{N-k-1}\left\langle \frac{N-k-1}{N-k}z_k - z_{k+1}, x_N - x_{k+1/2}\right\rangle \\
&= \frac{2}{N-k}\left\langle \mathbb{A}x_N, x_k - x_N\right\rangle - 2\alpha\left\langle \mathbb{A}x_N, z_k\right\rangle - \left(\frac{2}{N-k-1}\left\langle \mathbb{A}x_N, x_{k+1} - x_N\right\rangle - 2\alpha\left\langle \mathbb{A}x_N, z_{k+1}\right\rangle\right) \\
&\quad \underbrace{+ 2\left\langle \frac{1}{N-k}z_k - \frac{1}{N-k-1}z_{k+1}, x_{k+1/2} - x_N\right\rangle}_{:=R_k} \\
&= V_k^{(1)} - V_{k+1}^{(1)} + R_k
\end{aligned} \tag{82}$$

We rewrite R_k as following, using (79):

$$\begin{aligned}
R_k &= 2\left\langle \frac{1}{N-k}z_k, x_{k+1/2} - x_N\right\rangle - 2\left\langle \frac{1}{N-k-1}z_{k+1}, x_{k+1/2} - x_N\right\rangle \\
&= 2\left\langle \frac{1}{N-k}z_k, x_k - x_N - \alpha z_k - \alpha\mathbb{A}x_k\right\rangle - 2\left\langle \frac{1}{N-k-1}z_{k+1}, x_{k+1} - x_N - \frac{N-k-1}{N-k}\alpha(\mathbb{A}x_k - \mathbb{A}x_{k+1/2})\right\rangle \\
&= \frac{2}{N-k}\left\langle z_k, x_k - x_N\right\rangle - \frac{2}{N-k-1}\left\langle z_{k+1}, x_{k+1} - x_N\right\rangle \\
&\quad + \frac{2\alpha}{N-k}\left\langle \mathbb{A}x_k, z_{k+1} - z_k\right\rangle - \frac{2\alpha}{N-k}\|z_k\|^2 - 2\alpha\left\langle z_{k+1}, \frac{1}{N-k}\mathbb{A}x_{k+1/2}\right\rangle \\
&= V_k^{(2)} - V_{k+1}^{(2)} + \alpha\left(\frac{2}{N-k}\left\langle \mathbb{A}x_k, z_{k+1} - z_k\right\rangle - \frac{2}{N-k}\|z_k\|^2 - \frac{2(N-k-1)}{N-k}\left\langle z_{k+1}, z_k\right\rangle + 2\|z_{k+1}\|^2\right)
\end{aligned} \tag{83}$$

where the last equality uses (81). Now multiplying $\frac{1}{\alpha(N-k)^2}$ to LI_k and applying the identities (79), (81) we obtain

$$\begin{aligned}
&\frac{1}{\alpha(N-k)^2}\left(\mathrm{LI}_k\right) \\
&= \frac{1}{\alpha(N-k)^2}\left(\|x_{k+1/2} - x_k\|^2 - \alpha^2\|\mathbb{A}x_{k+1/2} - \mathbb{A}x_k\|^2\right) \\
&= \frac{1}{(N-k)^2}\left(\alpha\|z_k + \mathbb{A}x_k\|^2 - \alpha\|(N-k-1)z_k - (N-k)z_{k+1} - \mathbb{A}x_k\|^2\right) \\
&= \alpha\left(\left\|\frac{1}{N-k}z_k + \frac{1}{N-k}\mathbb{A}x_k\right\|^2 - \left\|\frac{1}{N-k}z_k + \frac{1}{N-k}\mathbb{A}x_k + (z_{k+1} - z_k)\right\|^2\right) \\
&= -\alpha\left\langle z_{k+1} - \left(1 - \frac{2}{N-k}\right)z_k + \frac{2}{N-k}\mathbb{A}x_k, z_{k+1} - z_k\right\rangle \\
&= -\alpha\left(\|z_{k+1}\|^2 + \left(1 - \frac{2}{N-k}\right)\|z_k\|^2 - \frac{2(N-k-1)}{N-k}\left\langle z_k, z_{k+1}\right\rangle + \frac{2}{N-k}\left\langle \mathbb{A}x_k, z_{k+1} - z_k\right\rangle\right)
\end{aligned} \tag{84}$$

holds. Now, we add (82) with (84), plug in (83) and simplify to obtain:

$$\begin{aligned}
\frac{2}{(N-k)(N-k-1)}\left(\mathrm{MI}_k\right) + \frac{1}{\alpha(N-k)^2}\left(\mathrm{LI}_k\right) &= (V_k^{(1)} + V_k^{(2)} - \alpha\|z_k\|^2) - (V_{k+1}^{(1)} + V_{k+1}^{(2)} - \alpha\|z_{k+1}\|^2) \\
&= V_k - V_{k+1}.
\end{aligned}$$

It remains to show $V_{N-1} \geq 0$. When $k = N-1$, we have

$$x_N = x_{N-1/2} = x_{N-1} - \alpha z_{N-1} - \alpha \mathbb{A} x_{N-1} \quad \Longleftrightarrow \quad z_{N-1} + \mathbb{A} x_{N-1} = -\frac{1}{\alpha}\left(x_N - x_{N-1}\right).$$

Therefore,

$$\begin{aligned}
V_{N-1} &= -\alpha \left\| z_{N-1} + \mathbb{A} x_N \right\|^2 + 2\left\langle z_{N-1} + \mathbb{A} x_N, x_{N-1} - x_N \right\rangle \\
&= -\alpha \left\| z_{N-1} + \mathbb{A} x_{N-1} + (\mathbb{A} x_N - \mathbb{A} x_{N-1}) \right\|^2 + 2\left\langle z_{N-1} + \mathbb{A} x_{N-1}, x_{N-1} - x_N \right\rangle \\
&= -\alpha \left\| -\frac{1}{\alpha}\left(x_N - x_{N-1}\right) + (\mathbb{A} x_N - \mathbb{A} x_{N-1}) \right\|^2 + \frac{2}{\alpha}\left\| x_N - x_{N-1} \right\|^2 \\
&= \frac{1}{\alpha}\left(\left\| x_N - x_{N-1} \right\|^2 - \alpha^2 \left\| \mathbb{A} x_N - \mathbb{A} x_{N-1} \right\|^2 \right) + 2\left\langle x_N - x_{N-1}, \mathbb{A} x_N - \mathbb{A} x_{N-1} \right\rangle \geq 0,
\end{aligned}$$

which concludes the proof.

G.2. Proof of Proposition 6.2

In Appendix B, we derive the H-matrix forms of FEG and Dual-FEG, respectively (21) and (24). It remains to check that the matrices are indeed in the anti-transpose relationship. Recall the definition

$$(H_{\text{Dual-FEG}})_{\ell,i} = \begin{cases} h_{\ell/2,(i-1)/2} & \text{if } i \leq \ell \\ 0 & \text{if } i > \ell. \end{cases}$$

For comparison, write

$$(H_{\text{FEG}})_{\ell,i} = \begin{cases} \hat{h}_{\ell/2,(i-1)/2} & \text{if } i \leq \ell \\ 0 & \text{if } i > \ell. \end{cases}$$

We check that $h_{\ell/2,(i-1)/2} = (H_{\text{Dual-FEG}})_{\ell,i} = (H^{\mathsf{A}}_{\text{FEG}})_{\ell,i} = (H_{\text{FEG}})_{2N-i+1,2N-\ell+1} = \hat{h}_{N-(i-1)/2,N-\ell/2}$ by carefully dividing cases, which completes the proof.

Case 1. For $\ell = i = 2k+1$ $(k = 0, \ldots, N-1)$,

$$h_{\ell/2,(i-1)/2} = h_{k+1/2,k} = \alpha L = \hat{h}_{N-k,N-k-1/2} = \hat{h}_{N-(i-1)/2,N-\ell/2}.$$

Case 2. For $\ell = 2k+1$ $(k = 1, \ldots, N-1)$, $i = 2j+2$ $(j = 0, \ldots, k-1)$,

$$h_{\ell/2,(i-1)/2} = h_{k+1/2,j+1/2} = -\frac{N-k}{(N-j-1)(N-j)}\alpha L = \hat{h}_{N-j-1/2,N-k-1/2} = \hat{h}_{N-(i-1)/2,N-\ell/2}$$

(note that $N-j > N-k$).

Case 3. For $\ell = 2k+1$ $(k = 1, \ldots, N-1)$, $i = 2j+1$ $(j = 0, \ldots, k-1)$,

$$h_{\ell/2,(i-1)/2} = h_{k+1/2,j} = 0 = \hat{h}_{N-j,N-k-1/2} = \hat{h}_{N-(i-1)/2,N-\ell/2}$$

(note that $N-j > N-k$).

Case 4. For $\ell = i = 2k+2$ $(k = 0, \ldots, N-1)$,

$$h_{\ell/2,(i-1)/2} = h_{k+1,k+1/2} = \frac{N-k-1}{N-k}\alpha L = \hat{h}_{N-k-1/2,N-k-1} = \hat{h}_{N-(i-1)/2,N-\ell/2}.$$

Case 5. For $\ell = 2k+2$, $i = 2k+1$ $(k = 0, \ldots, N-1)$,

$$h_{\ell/2,(i-1)/2} = h_{k+1,k} = -\frac{N-k-1}{N-k}\alpha L = \hat{h}_{N-k,N-k-1} = \hat{h}_{N-(i-1)/2,N-\ell/2}.$$

Case 6. For $\ell = 2k+2$, $i = 1, \ldots, 2k$ ($k = 0, \ldots, N-1$), because $\hat{h}_{k',j'} = 0$ for any $j' = 0, 1, \ldots$ and $k' \geq j' + 3/2$,

$$h_{\ell/2,(i-1)/2} = h_{k+1,(i-1)/2} = 0 = \hat{h}_{N-(i-1)/2,N-k-1} = \hat{h}_{N-(i-1)/2,N-\ell/2}.$$

□

H. Omitted details from Section 7

H.1. Derivation of ODE as continuous-time limit

In this section, we derive the Dual-Anchor ODE as continuous-time limit of Dual-OHM and Dual-FEG. The derivation is informal in the sense that we do not rigorously derive the sup-norm convergence

$$\lim_{\alpha\to 0^+} \sup_{t\in[0,T]} \left\|X(t) - x_{\lfloor t/\alpha\rfloor}\right\| = 0,$$

but we clarify all essential correspondence between discrete and continuous-time quantities.

H.1.1. Dual-OHM to Dual-Anchor ODE

Assume $\mathbb{A}$ is continuous monotone operator and let $\alpha > 0$. Consider Dual-OHM with $\mathbb{T} = 2\mathbb{J}_{\frac{\alpha}{2}\mathbb{A}} - \mathbb{I}$ (note that this differs from the "usual identification" $\mathbb{T} = 2\mathbb{J}_{\mathbb{A}} - \mathbb{I}$ by scale, because here we need a step-size α with respect to which we take the limit). Let $x_{k+1} = \mathbb{J}_{\frac{\alpha}{2}\mathbb{A}} y_k$, so that $x_{k+1} + \frac{\alpha}{2}\mathbb{A}x_{k+1} = y_k$. Then

$$\mathbb{T}y_k = 2x_{k+1} - y_k = y_k - \alpha\mathbb{A}x_{k+1},$$

and we can rewrite the z-form of Dual-OHM (3) as

$$\begin{aligned}
\frac{\frac{z_{k+1}}{\alpha} - \frac{z_k}{\alpha}}{\alpha} &= -\frac{1}{\alpha(N-k)}\frac{z_k}{\alpha} - \frac{1}{\alpha(N-k)}\mathbb{A}x_{k+1}\\
\frac{y_{k+1} - y_k}{\alpha} &= -\frac{z_{k+1}}{\alpha} - \mathbb{A}x_{k+1}.
\end{aligned}$$

Identifying $\alpha k = t$, $\alpha N = T$, $y_k = Y(t)$, $\frac{z_k}{\alpha} = Z(t)$ and $x_k = X(t)$, we obtain

$$\begin{aligned}
\dot{Z}(t) &= -\frac{1}{T-t}Z(t) - \frac{1}{T-t}\mathbb{A}(X(t+\alpha))\\
\dot{Y}(t) &= -Z(t+\alpha) - \mathbb{A}(X(t+\alpha)).
\end{aligned}$$

Assuming differentiability of X, Y, Z and taking the limit $\alpha \to 0$ gives

$$\begin{aligned}
\dot{Z}(t) &= -\frac{1}{T-t}Z(t) - \frac{1}{T-t}\mathbb{A}(X(t))\\
\dot{Y}(t) &= -Z(t) - \mathbb{A}(X(t)).
\end{aligned} \tag{85}$$

Once we show $\dot{Y}(t) = \dot{X}(t)$ in the limit $\alpha \to 0$, (85) becomes the Dual-Anchor ODE (5). From $x_{k+1} + \frac{\alpha}{2}\mathbb{A}x_{k+1} = y_k$, we have

$$\dot{X}(t) + \mathcal{O}(\alpha) = \frac{x_{k+1} - x_k}{\alpha} = \frac{y_k - y_{k-1}}{\alpha} - \frac{1}{2}\left(\mathbb{A}x_{k+1} - \mathbb{A}x_k\right) = \dot{Y}(t) + \mathcal{O}(\alpha) - \frac{1}{2}\left(\mathbb{A}(X(t+\alpha)) - \mathbb{A}(X(t))\right)$$

which shows that indeed, $\dot{X}(t) = \dot{Y}(t)$ in the limit $\alpha \to 0$.

H.1.2. Dual-FEG to Dual-Anchor ODE

From the definition of Dual-FEG, we have

$$\begin{aligned}
x_{k+1} &= x_{k+1/2} - \frac{N-k-1}{N-k}\alpha\left(\mathbb{A}x_{k+1/2} - \mathbb{A}x_k\right)\\
&= x_k - \alpha z_k - \alpha\mathbb{A}x_k - \frac{N-k-1}{N-k}\alpha\left(\mathbb{A}x_{k+1/2} - \mathbb{A}x_k\right)
\end{aligned} \tag{86}$$

where $\mathbb{A} = \nabla_{\pm}\mathbf{L}$. Therefore, we can write

$$\begin{aligned}
\frac{x_{k+1} - x_k}{\alpha} &= -z_k - \mathbb{A}x_k - \frac{\alpha(N-k-1)}{\alpha(N-k)}\left(\mathbb{A}x_{k+1/2} - \mathbb{A}x_k\right)\\
\frac{z_{k+1} - z_k}{\alpha} &= -\frac{1}{\alpha(N-k)}z_k - \frac{1}{\alpha(N-k)}\mathbb{A}x_{k+1/2}.
\end{aligned}$$

Now identify $\alpha k = t$, $\alpha N = T$, $x_k = X(t)$ and $z_k = Z(t)$ to obtain

$$\dot{X}(t) = -Z(t) - \mathbb{A}(X(t)) - \frac{T-t-\alpha}{T-t}\left(\mathbb{A}(X(t)+\mathcal{O}(\alpha)) - \mathbb{A}(X(t))\right)$$
$$\dot{Z}(t) = -\frac{1}{T-t}Z(t) - \frac{1}{T-t}\mathbb{A}(X(t)+\mathcal{O}(\alpha))$$

where we use $x_{k+1/2} = x_k + \mathcal{O}(\alpha)$. Thus, under the limit $\alpha \to 0$, we obtain

$$\dot{X}(t) = -Z(t) - \mathbb{A}(X(t))$$
$$\dot{Z}(t) = -\frac{1}{T-t}Z(t) - \frac{1}{T-t}\mathbb{A}(X(t))$$

which is the Dual-Anchor ODE (5).

H.2. Proof of Proposition 7.2

The Anchor ODE (4) (Ryu et al., 2019; Suh et al., 2023) is

$$\dot{X}(t) = -\mathbb{A}(X(t)) - \frac{1}{t}(X(t) - X_0) \tag{87}$$

where $X(0) = X_0$. We first rewrite (87) in the H-kernel form

$$\dot{X}(t) = -\int_0^t H(t,s)\mathbb{A}(X(s))ds.$$

Multiply t to both sides of (87) and reorganize to get

$$\frac{d}{dt}\left(t(X(t) - X_0)\right) = t\dot{X}(t) + (X(t) - X_0) = -t\mathbb{A}(X(t)).$$

Integrating from 0 to t and dividing both sides by t we obtain

$$X(t) - X_0 = -\frac{1}{t}\int_0^t s\mathbb{A}(X(s))ds.$$

Differentiating both sides gives

$$\dot{X}(t) = \frac{1}{t^2}\int_0^t s\mathbb{A}(X(s))ds - \mathbb{A}(X(t)) = \int_0^t \left(\frac{s}{t^2} - \delta(s-t)\right)\mathbb{A}(X(s))ds = -\int_0^t H(t,s)\mathbb{A}(X(s))ds$$

with

$$H(t,s) = -\frac{s}{t^2} + \delta(s-t)$$

and $\delta(\cdot)$ is the Dirac delta function. The continuous-time anti-diagonal transpose of this H-kernel is:

$$H^{\mathrm{A}}(t,s) = H(T-s, T-t) = -\frac{T-t}{(T-s)^2} + \delta(s-t),$$

Therefore, the H-dual of the Anchor ODE is

$$\dot{X}(t) = -\int_0^t H^{\mathrm{A}}(t,s)\mathbb{A}(X(s))ds = \int_0^t \left(\frac{T-t}{(T-s)^2} - \delta(s-t)\right)\mathbb{A}(X(s))ds = \int_0^t \frac{T-t}{(T-s)^2}\mathbb{A}(X(s))ds - \mathbb{A}(X(t)).$$

Now define

$$Z(t) = -\dot{X}(t) - \mathbb{A}(X(t)) = -(T-t)\int_0^t \frac{1}{(T-s)^2}\mathbb{A}(X(s))ds.$$

Differentiating both sides of the above equation gives

$$\dot{Z}(t) = \int_0^t \frac{1}{(T-s)^2}\mathbb{A}(X(s))ds - \frac{1}{T-t}\mathbb{A}(X(t)) = -\frac{1}{T-t}Z(t) - \frac{1}{T-t}\mathbb{A}(X(t)).$$

The last equation, together $\dot{X}(t) = -Z(t) - \mathbb{A}(X(t))$ which follows directly from definition of Z, becomes (5), which is the Dual-Anchor ODE. This shows that Dual-Anchor ODE and Anchor ODE are H-duals of each other.

H.3. Existence, uniqueness and Lyapunov analysis of Dual Anchor ODE

H.3.1. Well-posedness of the dynamics

Theorem H.1. *Suppose $\mathbb{A}\colon \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous and let $T > 0$. Consider the differential equation*

$$\begin{pmatrix} \dot{X}(t) \\ \dot{Z}(t) \end{pmatrix} = -\begin{pmatrix} Z(t) + \mathbb{A}(X(t)) \\ \frac{1}{T-t} Z(t) + \frac{1}{T-t}\mathbb{A}(X(t)) \end{pmatrix} \tag{88}$$

for $t \in (0, T)$, with initial conditions $X(0) = X_0$ and $Z(0) = 0$. Then there is a unique solution $\begin{pmatrix} X \\ Z \end{pmatrix} \in \mathcal{C}^1([0,T), \mathbb{R}^{2d})$ that satisfies (88) for all $t \in (0, T)$.

Proof. The right hand side of (88), as a function of t and $\begin{pmatrix} X \\ Z \end{pmatrix}$, can be rewritten as

$$F\left(t, \begin{pmatrix} X \\ Z \end{pmatrix}\right) = -\begin{pmatrix} Z + \mathbb{A}(X) \\ \frac{1}{T-t} Z + \frac{1}{T-t}\mathbb{A}(X) \end{pmatrix} = -\begin{pmatrix} I & 0 \\ 0 & \frac{1}{T-t} I \end{pmatrix} \begin{pmatrix} Z + \mathbb{A}(X) \\ Z + \mathbb{A}(X) \end{pmatrix}.$$

Let $L > 0$ be the Lipschitz continuity parameter for $\mathbb{A}$. Fix $\bar{t} \in (0, T)$; then for $t \in [0, \bar{t}]$, F is Lipschitz continuous with parameter

$$\sqrt{2\left(1 + \frac{1}{(T-\bar{t})^2}\right)} \max\{1, L\}$$

in $\begin{pmatrix} X \\ Z \end{pmatrix}$. By classical ODE theory, Lipschitz continuity of F on $[0, \bar{t}]$ implies that the solution $\begin{pmatrix} X \\ Z \end{pmatrix} \in \mathcal{C}^1([0, \bar{t}], \mathbb{R}^{2d})$ uniquely exists. Since $\bar{t} \in (0, T)$ is arbitrary, the solution uniquely exists on $[0, T)$ and the proof is complete. □

Corollary H.2. *Suppose $\mathbb{A}\colon \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous and let $T > 0$. The solution $\begin{pmatrix} X \\ Z \end{pmatrix} : [0, T) \to \mathbb{R}^{2d}$ to (88) satisfies the differential equation*

$$\ddot{X}(t) + \frac{1}{T-t}\dot{X}(t) + \frac{d}{dt}\mathbb{A}(X(t)) = 0 \tag{89}$$

almost everywhere in $t \in [0, T)$. Conversely, if $X \in \mathcal{C}^1([0,T), \mathbb{R}^d)$ satisfies (89) almost everywhere in $t \in [0, T)$, then with $Z\colon [0, T) \to \mathbb{R}^d$ defined as $Z(t) = -\dot{X}(t) - \mathbb{A}(X(t))$, $\begin{pmatrix} X \\ Z \end{pmatrix}$ is the solution of (88).

Proof. Let L be the Lipschitz continuity parameter of $\mathbb{A}$ and take $\bar{t} \in (0, T)$. As $t \mapsto \dot{X}(t)$ is continuous on the compact interval $[0, \bar{t}]$, we have $M = \max_{t \in [0,\bar{t}]} \left\|\dot{X}(t)\right\| < \infty$. Therefore $t \mapsto X(t)$ is M-Lipschitz, which implies that $t \mapsto \mathbb{A}(X(t))$ is LM-Lipschitz; hence $\mathbb{A}(X(t))$ is differentiable almost everywhere in t. Then $\dot{X}(t) = -Z(t) - \mathbb{A}(X(t))$ is absolutely continuous in t (being the sum of $\mathcal{C}^1$ function Z and Lipschitz continuous $\mathbb{A}(X(t))$) and hence differentiable almost everywhere. Now the equivalence between (88) and (89) can be proved via same arguments as in Appendix B.6, with all equalities holding almost everywhere in t. □

H.3.2. Regularity at the terminal time $t = T$

Next, we show the solution on $[0, T)$ can be continuously extended to the terminal time $t = T$ due to its favorable properties. First, we need:

Lemma H.3. *Suppose $\mathbb{A}\colon \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous and monotone. Let $\begin{pmatrix} X \\ Z \end{pmatrix}$ be the solution to (88) on $[0, T)$. Then*

$$\Psi(t) = \frac{1}{(T-t)^2}\left\|\dot{X}(t)\right\|^2$$

is nonincreasing in $t \in [0, T)$. In particular, this implies

$$\left\|\dot{Z}(t)\right\|^2 = \frac{1}{(T-t)^2}\left\|\dot{X}(t)\right\|^2 \le \frac{1}{T^2}\left\|\mathbb{A}(X_0)\right\|^2.$$

Proof. Take the inner product with $\frac{1}{(T-t)^2}\dot{X}(t)$ to both sides of (89), we have for almost every $t \in [0, T)$,

$$\begin{aligned} 0 &= \frac{1}{(T-t)^2}\left\langle \ddot{X}(t), \dot{X}(t)\right\rangle + \frac{1}{(T-t)^3}\left\|\dot{X}(t)\right\|^2 + \frac{1}{(T-t)^2}\left\langle \frac{d}{dt}\mathbb{A}(X(t)), \dot{X}(t)\right\rangle \\ &= \frac{d}{dt}\left(\frac{1}{2(T-t)^2}\left\|\dot{X}(t)\right\|^2\right) + \frac{1}{(T-t)^2}\left\langle \frac{d}{dt}\mathbb{A}(X(t)), \dot{X}(t)\right\rangle. \end{aligned}$$

Note that because $\dot{X}(t)$ is absolutely continuous, $\Psi(t)$ is also absolutely continuous. Integrating from 0 to t and reorganizing, we obtain the following "conservation law":

$$E \equiv \frac{1}{2T^2}\left\|\dot{X}(0)\right\|^2 = \underbrace{\frac{1}{2(T-t)^2}\left\|\dot{X}(t)\right\|^2}_{\frac{\Psi(t)}{2}} + \int_0^t \frac{1}{(T-s)^2}\left\langle \frac{d}{ds}\mathbb{A}(X(s)), \dot{X}(s)\right\rangle ds$$

(E is a constant). Note that by monotonicity of $\mathbb{A}$,

$$\left\langle \frac{d}{dt}\mathbb{A}(X(t)), \dot{X}(t)\right\rangle = \lim_{h\to 0}\frac{\langle \mathbb{A}(X(t+h)) - \mathbb{A}(X(t)), X(t+h) - X(t)\rangle}{h^2} \ge 0.$$

Therefore for almost every $t \in (0, T)$,

$$\dot{\Psi}(t) = -\frac{2}{(T-t)^2}\left\langle \frac{d}{dt}A(X(t)), \dot{X}(t)\right\rangle \le 0,$$

from which we conclude that $\Psi(t)$ is nonincreasing. Finally, (88) gives $\dot{Z}(t) = \frac{1}{T-t}\dot{X}(t)$ and $\dot{X}(0) = -Z(0) - \mathbb{A}(X_0) = -\mathbb{A}(X_0)$, so we conclude that

$$\left\|\dot{Z}(t)\right\|^2 = \frac{1}{(T-t)^2}\left\|\dot{X}(t)\right\|^2 = \Psi(t) \le \Psi(0) = \frac{1}{T^2}\left\|\mathbb{A}(X_0)\right\|^2.$$

□

Corollary H.4. *Suppose $\mathbb{A}\colon \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous and monotone. Then, given the solution $\begin{pmatrix} X \\ Z \end{pmatrix}$ to (88) on $[0, T)$, one can continuously extend $X(t)$ to $t = T$, and the extension $X\colon [0, T] \to \mathbb{R}^d$ satisfies*

$$\lim_{t\to T^-}\dot{X}(t) = \lim_{t\to T^-}\frac{X(t) - X(T)}{t - T} = 0.$$

Proof. Let $t \in [0, T)$ and let $h > 0$ satisfy $t + h < T$. Then by Lemma H.3,

$$\left\|X(t+h) - X(t)\right\| = \left\|\int_t^{t+h}\dot{X}(s)ds\right\| \le \int_t^{t+h}\left\|\dot{X}(s)\right\|ds \le \int_t^{t+h}\frac{T-s}{T}\left\|\mathbb{A}(X_0)\right\|ds \le h\left\|\mathbb{A}(X_0)\right\|.$$

This shows that $X\colon [0, T) \to \mathbb{R}^d$ is Lipschitz continuous, which implies any sequence $\{X(t_j)\}$ with $0 < t_j \nearrow T$ is a Cauchy sequence, and such sequential limit is unique. Setting $X(T)$ to this unique limit, we see that $X\colon [0, T] \to \mathbb{R}^d$ is a (unique) continuous function extending X. Additionally, Lemma H.3 gives

$$\lim_{t\to T^-}\left\|\dot{X}(t)\right\| \le \lim_{t\to T^-}\frac{T-t}{T}\left\|\mathbb{A}(X_0)\right\| = 0.$$

Therefore,

$$\lim_{t\to T^-}\left\|\frac{X(t) - X(T)}{t - T}\right\| = \lim_{t\to T^-}\frac{1}{T-t}\left\|\int_t^T \dot{X}(s)ds\right\| \le \lim_{t\to T^-}\frac{1}{T-t}\int_t^T\left\|\dot{X}(s)\right\|ds = \lim_{t\to T^-}\left\|\dot{X}(t)\right\| = 0$$

where the second-to-last equality uses L'Hôpital's rule. □

H.3.3. Proof of Theorem 7.1

Recall that we define $V\colon [0,T) \to \mathbb{R}$ by

$$V(t) = -\left\|Z(t) + \mathbb{A}(X(T))\right\|^2 + 2\left\langle Z(t) + \mathbb{A}(X(T)), \frac{1}{T-t}\left(X(t) - X(T)\right)\right\rangle.$$

Differentiate V and apply the substitutions $Z = -(\dot{X} + \mathbb{A}(X))$ and $\dot{Z} = \frac{1}{T-t}\dot{X}$ to obtain

$$\begin{aligned}
\dot{V}(t) &= -2\left\langle \dot{Z}(t), Z(t) + \mathbb{A}(X(T))\right\rangle + 2\left\langle \dot{Z}(t), \frac{1}{T-t}\left(X(t) - X(T)\right)\right\rangle \\
&\quad + 2\left\langle Z(t) + \mathbb{A}(X(T)), \frac{d}{dt}\left(\frac{1}{T-t}\left(X(t) - X(T)\right)\right)\right\rangle \\
&= 2\left\langle \frac{1}{T-t}\dot{X}(t), \dot{X}(t) + \mathbb{A}(X(t)) - \mathbb{A}(X(T)) + \frac{1}{T-t}\left(X(t) - X(T)\right)\right\rangle \\
&\quad - 2\left\langle \dot{X}(t) + \mathbb{A}(X(t)) - \mathbb{A}(X(T)), \frac{1}{(T-t)^2}\left(X(t) - X(T)\right) + \frac{1}{T-t}\dot{X}(t)\right\rangle \\
&= -\frac{2}{(T-t)^2}\left\langle X(t) - X(T), \mathbb{A}(X(t)) - \mathbb{A}(X(T))\right\rangle \\
&\le 0.
\end{aligned}$$

On the other hand, by Corollary H.4, we have $\lim_{t\to T^-}\dot{X}(t) = 0$, which implies

$$\lim_{t\to T^-} Z(t) = \lim_{t\to T^-}\left(-\dot{X}(t) - \mathbb{A}(X(t))\right) = -\mathbb{A}(X(T))$$

and thus,

$$\lim_{t\to T^-} V(t) = \lim_{t\to T^-}\left(-\left\|Z(t) + \mathbb{A}(X(T))\right\|^2\right) + \lim_{t\to T^-} 2\left\langle Z(t) + \mathbb{A}(X(T)), \frac{X(t) - X(T)}{T-t}\right\rangle = 0$$

where the last equality holds because both $Z(t) + \mathbb{A}(X(T))$ and $\frac{X(t)-X(T)}{t-T}$ converges to 0 as $t \to T^-$. Therefore,

$$0 = \lim_{t\to T^-} V(t) \le V(0) = -\left\|\mathbb{A}(X(T))\right\|^2 - \frac{2}{T}\left\langle \mathbb{A}(X(T)), X(T) - X_0\right\rangle$$

where the last equality uses $Z(0) = 0$. Now multiplying $-\frac{T}{2}$ to both sides and applying Lemma 3.2, we conclude

$$\left\|\mathbb{A}(X(T))\right\|^2 \le \frac{4\left\|X_0 - X_\star\right\|^2}{T^2}.$$

H.3.4. Correspondence of continuous-time analysis with discrete-time analysis

We overview how the continuous-time analysis presented above corresponds to respective analyses of Dual-OHM and Dual-FEG. Precisely, we verify that under the identifcation of terms from Appendix H.1, the following holds true:

(i) The discrete Lyapunov function V_k corresponds to $V(t)$

(ii) The consecutive difference $V_{k+1} - V_k$ corresponds to $\dot{V}(t)$

- Dual-OHM

  In Appendix H.1.1 we consider Dual-OHM with $\mathbb{T} = 2\mathbb{J}_{\frac{\alpha}{2}\mathbb{A}} - \mathbb{I}$ for continuous monotone $\mathbb{A}$ to derive the continuous-time limit, and we have the identities $x_{k+1} = \mathbb{J}_{\frac{\alpha}{2}\mathbb{A}} y_k$ and $x_{k+1} + \frac{\alpha}{2}\mathbb{A}x_{k+1} = y_k$. Therefore, the corresponding Lyapunov analysis of Dual-OHM should use $g_k = \frac{\alpha}{2}\mathbb{A}x_k$ (instead of $g_k = \tilde{\mathbb{A}}x_k$). Recall that we identify $\alpha k = t$, $\alpha N = T$, $x_k = X(t)$, $y_k = Y(t)$ and $\frac{z_k}{\alpha} = Z(t)$.

(i) $\frac{V_k}{\alpha^2} \longleftrightarrow V(t)$.

Replacing $g_N = \tilde{\mathbb{A}} x_N$ with $\frac{\alpha}{2}\mathbb{A}x_N$ in the Lyapunov function of Dual-OHM gives

$$V_k = -\frac{N-k-1}{N-k}\left\|z_k + \alpha\mathbb{A}x_N\right\|^2 + \frac{2}{N-k}\left\langle z_k + \alpha\mathbb{A}x_N, y_k - y_{N-1}\right\rangle.$$

Therefore

$$\begin{aligned}\frac{V_k}{\alpha^2} &= -\left(1 - \frac{\alpha}{\alpha(N-k)}\right)\left\|\frac{z_k}{\alpha} + \mathbb{A}x_N\right\|^2 + \frac{2}{\alpha(N-k)}\left\langle \frac{z_k}{\alpha} + \mathbb{A}x_N, y_k - y_{N-1}\right\rangle \\ &= -\left(1 + \mathcal{O}(\alpha)\right)\left\|Z(t) + \mathbb{A}(X(T))\right\|^2 + \frac{2}{T-t}\left\langle Z(t) + \mathbb{A}(X(T)), Y(t) - Y(T-\alpha)\right\rangle.\end{aligned}$$

From the identity $x_{k+1} + \frac{\alpha}{2}\mathbb{A}x_{k+1} = y_k$ we have $Y(t) = X(t+\alpha) + \mathcal{O}(\alpha)$, so in the limit $\alpha \to 0$, $Y(t) = X(t)$. Then the above equation establishes the desired correspondence, as

$$\text{(RHS)} = -\left\|Z(t) + \mathbb{A}(X(T))\right\|^2 + \frac{2}{T-t}\left\langle Z(t) + \mathbb{A}(X(T)), X(t) - X(T)\right\rangle.$$

(ii) $\frac{\frac{1}{\alpha^2}V_{k+1} - \frac{1}{\alpha^2}V_k}{\alpha} \longleftrightarrow \dot{V}(t)$.

The Lyapunov analysis of Dual-OHM establishes

$$V_{k+1} - V_k = -\frac{4}{(N-k)(N-k-1)}\left\langle x_{k+1} - x_N, g_{k+1} - g_N\right\rangle.$$

Replacing g_k with $\frac{\alpha}{2}\mathbb{A}x_k$, dividing both sides by α^3 and applying the identifications, we obtain

$$\begin{aligned}\frac{\frac{1}{\alpha^2}V_{k+1} - \frac{1}{\alpha^2}V_k}{\alpha} &= -\frac{4}{\alpha^3(N-k)(N-k-1)}\left\langle x_{k+1} - x_N, \frac{\alpha}{2}\mathbb{A}x_{k+1} - \frac{\alpha}{2}\mathbb{A}x_N\right\rangle \\ &= -\frac{2}{(\alpha N - \alpha k)(\alpha N - \alpha k - \alpha)}\left\langle x_{k+1} - x_N, \mathbb{A}(X(t+\alpha) - \mathbb{A}(X(T))\right\rangle \\ &= -\frac{2}{(T-t)^2 + \mathcal{O}(\alpha)}\left\langle X(t+\alpha) - X(T), \mathbb{A}(X(t+\alpha) - \mathbb{A}(X(T))\right\rangle.\end{aligned}$$

Taking $\alpha \to 0$, we have

$$\text{(RHS)} = -\frac{2}{(T-t)^2}\left\langle X(t) - X(T), \mathbb{A}(X(t)) - \mathbb{A}(X(T))\right\rangle = \dot{V}(t)$$

which gives the desired correspondence.

- Dual-FEG

As above, we use the identification $\alpha k = t$, $\alpha N = T$, $x_k = X(t)$ and $z_k = Z(t)$.

(i) $\frac{V_k}{\alpha} \longleftrightarrow V(t)$.

The Lyapunov function of Dual-FEG is

$$V_k = -\alpha\left\|z_k + \mathbb{A}x_N\right\|^2 + \frac{2}{N-k}\left\langle z_k + \mathbb{A}x_N, x_k - x_N\right\rangle.$$

Dividing both sides by α and applying the identifcations, we have

$$\begin{aligned}\frac{V_k}{\alpha} &= -\left\|z_k + \mathbb{A}x_N\right\|^2 + \frac{2}{\alpha(N-k)}\left\langle z_k + \mathbb{A}x_N, x_k - x_N\right\rangle \\ &= -\left\|Z(t) + \mathbb{A}(X(T))\right\|^2 + \frac{2}{T-t}\left\langle Z(t) + \mathbb{A}(X(T)), X(t) - X(T)\right\rangle,\end{aligned}$$

we get the desired correspondence.

(ii) $\frac{\frac{1}{\alpha}V_{k+1}-\frac{1}{\alpha}V_k}{\alpha} \longleftrightarrow \dot{V}(t)$.

The Lyapunov analysis of Dual-FEG establishes

$$\begin{aligned}V_{k+1}-V_k = &-\frac{2}{(N-k)(N-k-1)}\left\langle \mathbb{A}x_N - \mathbb{A}x_{k+1/2}, x_N - x_{k+1/2}\right\rangle \\ &- \frac{1}{\alpha(N-k)^2}\Big(\|x_{k+1/2}-x_k\|^2 - \alpha^2\|\mathbb{A}x_{k+1/2}-\mathbb{A}x_k\|^2\Big)\end{aligned}$$

Because $x_{k+1/2} = x_k + \mathcal{O}(\alpha)$, dividing the both sides by α^2, we obtain

$$\begin{aligned}\frac{\frac{1}{\alpha}V_{k+1}-\frac{1}{\alpha}V_k}{\alpha} &= -\frac{2}{\alpha^2(N-k)(N-k-1)}\left\langle \mathbb{A}x_N - \mathbb{A}x_{k+1/2}, x_N - x_{k+1/2}\right\rangle \\ &\quad - \frac{1}{\alpha^2(N-k)^2}\Big(\frac{1}{\alpha}\|x_{k+1/2}-x_k\|^2 - \alpha\|\mathbb{A}x_{k+1/2}-\mathbb{A}x_k\|^2\Big) \\ &= -\frac{2}{(T-t)^2+\mathcal{O}(\alpha)}\left\langle \mathbb{A}(X(T)) - \mathbb{A}(X(t)+\mathcal{O}(\alpha)), X(T)-X(t)+\mathcal{O}(\alpha)\right\rangle \\ &\quad \underbrace{-\frac{1}{(T-t)^2}\Big(\frac{1}{\alpha}\mathcal{O}(\alpha^2) - \alpha\|\mathbb{A}(X(t)+\mathcal{O}(\alpha)) - \mathbb{A}(X(t))\|^2\Big)}_{\mathcal{O}(\alpha)}.\end{aligned}$$

Taking $\alpha \to 0$, we have

$$\text{(RHS)} = -\frac{2}{(T-t)^2}\left\langle X(t)-X(T), \mathbb{A}(X(t)) - \mathbb{A}(X(T))\right\rangle = \dot{V}(t)$$

which gives the desired correspondence.

H.4. Generalization to differential inclusion with maximally monotone $\mathbb{A}$

So far, we have analyzed the Dual-Anchor ODE with respect to Lipschitz continuous $\mathbb{A}$. In this section, we deal with its extension to differential inclusion, which is a generalized form of continuous-time model, covering the case of general (possibly set-valued) maximally monotone operators:

$$\begin{pmatrix}\dot{X}(t)\\ \dot{Z}(t)\end{pmatrix} \in -\begin{pmatrix}Z(t)+\mathbb{A}(X(t))\\ \frac{1}{T-t}Z(t)+\frac{1}{T-t}\mathbb{A}(X(t))\end{pmatrix}. \tag{90}$$

We say that $(X(t), Z(t))$ is a solution to this differential inclusion if it satisfies (90) almost everywhere in t. (So unlike in ODEs, we do not require X, Z to be differentiable everywhere, but only require absolute continuity, which implies differentiability almost everywhere.) Although (90) is technically not an ODE, with a slight abuse of notation, we will often refer to it as (generalized) Dual-Anchor ODE throughout the section.

H.4.1. Existence of a solution to the generalized Dual-Anchor ODE

Theorem H.5. *Let $\mathbb{A}\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be maximally monotone and let $T > 0$. Given the initial conditions $X(0) = X_0 \in \operatorname{dom}\mathbb{A}$ and $Z(0) = 0$, there exists a solution to the generalized Dual-Anchor ODE, i.e., an absolutely continuous curve $\begin{pmatrix}X\\ Z\end{pmatrix}\colon [0,T) \to \mathbb{R}^{2d}$ that satisfies (90) for $t \in (0,T)$ almost everywhere.*

We proceed with similar ideas as in (Aubin & Cellina, 1984, Chpater 3) and (Suh et al., 2023, Appendix B.2), i.e., we construct a sequence $\{(X_\delta(t), Z_\delta(t))\}_{\delta>0}$ of solutions to ODEs approximating (90), with $\mathbb{A}$ replaced by Yosida approximations $\mathbb{A}_\delta$ (which are much better-behaved, being single-valued and Lipschitz continuous). Then we show that the limit of $(X_\delta(t), Z_\delta(t))$ in $\delta \to 0$ is a solution to the original inclusion (90).

Below we present some well-known facts needed to rigorously establish the approximation argument.

Lemma H.6. *Let $\mathbb{A}\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be maximally monotone. Then for $\delta > 0$, the Yosida approximation operators*

$$\mathbb{A}_\delta = \frac{1}{\delta}\left(\mathbb{I} - \mathbb{J}_{\delta\mathbb{A}}\right) = \frac{1}{\delta}\left(\mathbb{I} - (\mathbb{I} + \delta\mathbb{A})^{-1}\right)$$

satisfy the followings:

(i) $\forall x \in \mathbb{R}^d$, $\mathbb{A}_\delta(x) \in \mathbb{A}(\mathbb{J}_{\delta\mathbb{A}}x)$.

(ii) $\mathbb{A}_\delta\colon \mathbb{R}^d \to \mathbb{R}^d$ *is singled-valued,* $\frac{1}{\delta}$*-Lipschitz continuous and maximally monotone.*

(iii) $\|\mathbb{A}_\delta(x)\| \le \|m(\mathbb{A}(x))\|$.

(iv) $\lim_{\delta\to 0+} \mathbb{A}_\delta(x) = m(\mathbb{A}(x)) := \operatorname{argmin}_{u\in\mathbb{A}(x)} \|u\|$. *(The minimum norm element of the set $\mathbb{A}(x)$ is well-defined because $\mathbb{A}(x)$ is a closed and convex set due to maximal monotonicity of $\mathbb{A}$.)*

Proof. See (Aubin & Cellina, 1984, Chpater 3.1, Theorem 2). □

Lemma H.7. *Let $\{x_n(\cdot)\}_n$ be a sequence of absolutely continuous functions $x_n\colon I \to \mathbb{R}^d$, where $I \subset \mathbb{R}$ is an interval of finite length. Suppose that*

(i) $\forall t \in I$, *the set $\{x_n(t)\}_n$ is bounded.*

(ii) There exists $M > 0$ such that $\|\dot{x}_n(t)\| \le M$ almost everywhere in $t \in I$.

Then there exists a subsequence $x_{n_k}(\cdot)$ such that

(1) $x_{n_k}(\cdot)$ uniformly converges to $x(\cdot)$ over compact subsets of I.

(2) $x(\cdot)$ is absolutely continuous, so $\dot{x}$ exists almost everywhere.

(3) $\dot{x}_{n_k}(\cdot)$ converges weakly to $\dot{x}(\cdot)$ in $L^2(I, \mathbb{R}^d)$.

Proof. This is a simplified version of (Aubin & Cellina, 1984, Chapter 0.3, Theorem 4). For completeness, we outline the proof below.

The family $\{x_n(\cdot)\}_n$ is pointwisely bounded (condition (i)) and equicontinuous as

$$\|x_n(t) - x_n(s)\| = \left\|\int_s^t \dot{x}_n(\tau)\, d\tau\right\| \le M|t-s|, \quad \forall s, t \in I.$$

By Arzelà-Ascoli theorem, there exists a subsequence $x_{n_k}(\cdot)$ such that $x_{n_k} \to x$ uniformly over compact subsets of I, where $x\colon I \to \mathbb{R}^d$ is continuous (so far we have not shown absolute continuity). Now observe that $\{\dot{x}_{n_k}(\cdot)\}_k$ is a bounded set within $L^\infty(I, \mathbb{R}^d) = L^1(I, \mathbb{R}^d)^*$. Therefore, by Banach-Alaoglu theorem, $\dot{x}_{n_k}(\cdot)$ again has a subsequence that converges in the weak-$*$ sense. By replacing n_k with this subsequence if necessary, assume without loss of generality that $\dot{x}_{n_k} \xrightarrow{\text{w}*} y \in L^\infty(I, \mathbb{R}^d)$. This means that for any $c \in L^1(I, \mathbb{R}^d)$,

$$\lim_{k\to\infty} \int_I \langle \dot{x}_{n_k}(\tau), c(\tau)\rangle\, d\tau = \int_I \langle y(\tau), c(\tau)\rangle\, d\tau. \tag{91}$$

For any $s, t \in I$ and $\mathbf{a} \in \mathbb{R}^d$, taking $c(\tau) = 1_{[s,t]}(\tau)\mathbf{a}$ in particular, we have

$$\lim_{k\to\infty} \int_s^t \langle \dot{x}_{n_k}(\tau), \mathbf{a}\rangle\, d\tau = \int_s^t \langle y(\tau), \mathbf{a}\rangle\, d\tau = \left\langle \int_s^t y(\tau)\, d\tau, \mathbf{a}\right\rangle.$$

On the other hand, the left hand side equals $\lim_{k\to\infty} \langle x_{n_k}(t) - x_{n_k}(s), \mathbf{a}\rangle = \langle x(t) - x(s), \mathbf{a}\rangle$. Since $\mathbf{a} \in \mathbb{R}^d$ is arbitrary, this shows

$$x(t) - x(s) = \int_s^t y(\tau)\, d\tau, \quad y \in L^\infty(I, \mathbb{R}^d)$$

so x is absolutely continuous (in fact, Lipschitz continuous) and $\dot{x} = y$ almost everywhere. Finally, note that $L^2(I, \mathbb{R}^d) \subset L^1(I, \mathbb{R}^d)$ (as I is of finite length), so we can take $c \in L^2(I, \mathbb{R}^d)$ in (91), showing that $\dot{x}_{n_k} \to y$ weakly in $L^2(I, \mathbb{R}^d)$. □

Using the above results, we can now prove Theorem H.5. For $\delta > 0$, the ODE

$$\begin{pmatrix} \dot{X}_\delta(t) \\ \dot{Z}_\delta(t) \end{pmatrix} = -\begin{pmatrix} Z_\delta(t) + \mathbb{A}_\delta(X_\delta(t)) \\ \frac{1}{T-t} Z_\delta(t) + \frac{1}{T-t}\mathbb{A}_\delta(X_\delta(t)) \end{pmatrix} \tag{92}$$

with initial conditions $X_\delta(0) = X_0 \in \operatorname{dom} \mathbb{A}$ and $Z_\delta(0) = 0$ has a unique $\mathcal{C}^1$ solution $(X_\delta(t), Z_\delta(t))$ for $t \in [0, T]$ by Theorem H.1 and Corollary H.4. Additionally by Lemma H.3, X_δ and Z_δ satisfy

$$\left\|\dot{X}_\delta(t)\right\| \le \frac{T-t}{T} \left\|m(\mathbb{A}(X_0))\right\| \le \left\|m(\mathbb{A}(X_0))\right\|, \qquad \left\|\dot{Z}_\delta(t)\right\| \le \frac{1}{T} \left\|m(\mathbb{A}(X_0))\right\|. \tag{93}$$

Therefore, we can apply Lemma H.7 to obtain a positive sequence $\delta_n \to 0$ such that

- $\begin{pmatrix} X_{\delta_n} \\ Z_{\delta_n} \end{pmatrix} \to \begin{pmatrix} X \\ Z \end{pmatrix}$ uniformly on $[0, T]$, where $\begin{pmatrix} X \\ Z \end{pmatrix}$ is absolutely continuous,
- $\begin{pmatrix} \dot{X}_{\delta_n} \\ \dot{Z}_{\delta_n} \end{pmatrix} \to \begin{pmatrix} \dot{X} \\ \dot{Z} \end{pmatrix}$ weakly in $L^2([0, T], \mathbb{R}^{2d})$.

Now, we wish to show that $(X(t), Z(t))$ is a solution to (90). Define $G_\delta, G \colon [0, T] \to \mathbb{R}^d$ as

$$G_\delta(t) = Z_\delta(t) + \dot{X}_\delta(t), \quad G(t) = Z(t) + \dot{X}(t).$$

By construction,

$$G_{\delta_n}(\cdot) \to G(\cdot) \quad \textit{weakly in } L^2([0, T], \mathbb{R}^d).$$

Note that because (X_δ, Z_δ) solves (92) and Lemma H.6 (i) holds, we have

$$G_{\delta_n}(t) = -\mathbb{A}_\delta(X_\delta(t)) \in -\mathbb{A}(\mathbb{J}_{\delta_n \mathbb{A}}(X_{\delta_n}(t))).$$

On the other hand, because $Z_{\delta_n}(0) = 0$, by (93) we have

$$\left\|Z_{\delta_n}(t)\right\| = \left\|\int_0^t \dot{Z}_{\delta_n}(s) ds\right\| \le \int_0^t \left\|\dot{Z}_{\delta_n}(s)\right\| ds \le \int_0^t \frac{\left\|\mathbb{A}_{\delta_n}(X_0)\right\|}{T} ds \le \left\|\mathbb{A}_{\delta_n}(X_0)\right\| \le \left\|m(\mathbb{A}(X_0))\right\|.$$

Together with the norm bound on $\dot{X}$ in (93), this implies that for all $t \in [0, T]$,

$$\begin{aligned} \left\|X_{\delta_n}(t) - \mathbb{J}_{\delta_n \mathbb{A}}(X_{\delta_n}(t))\right\| = \delta_n \left\|\mathbb{A}_{\delta_n}(X_{\delta_n}(t))\right\| &= \delta_n \left\|Z_{\delta_n}(t) + \dot{X}_{\delta_n}(t)\right\| \\ &\le \delta_n \left(\left\|Z_{\delta_n}(t)\right\| + \left\|\dot{X}_{\delta_n}(t)\right\|\right) \le 2\delta_n \left\|m(\mathbb{A}(X_0))\right\|. \end{aligned}$$

As $X_{\delta_n}(\cdot)$ converges uniformly to $X(\cdot)$, the above result shows that $\mathbb{J}_{\delta_n \mathbb{A}}(X_{\delta_n}(\cdot))$ uniformly converges to $X(\cdot)$ as well. In particular,

$$\mathbb{J}_{\delta_n \mathbb{A}}(X_{\delta_n}(\cdot)) \to X(\cdot) \quad \textit{strongly in } L^2([0, T], \mathbb{R}^d).$$

Now, define $\mathcal{A} \colon L^2([0, T], \mathbb{R}^d) \to L^2([0, T], \mathbb{R}^d)$ by

$$\mathcal{A}(y) = \left\{u \in L^2([0, T], \mathbb{R}^d) \,|\, u(t) \in \mathbb{A}(y(t)) \text{ for a.e. } t \in [0, T]\right\}.$$

$\mathcal{A}$ is monotone because $\mathbb{A}$ is monotone: if $u \in \mathcal{A}(y), v \in \mathcal{A}(z)$ then

$$\langle u - v, y - z\rangle_{L^2([0,T],\mathbb{R}^d)} = \int_0^T \langle u(t) - v(t), y(t) - z(t)\rangle \, dt \ge 0$$

since $u(t) \in \mathbb{A}(y(t)), v(t) \in \mathbb{A}(z(t))$ a.e. in $t \in [0, T]$. If $\mathcal{I} \colon L^2([0, T], \mathbb{R}^d) \to L^2([0, T], \mathbb{R}^d)$ is the identity operator, then $\mathcal{I} + \mathcal{A}$ is surjective: for any $u \in L^2([0, T], \mathbb{R}^d)$, we have $y(t) = \mathbb{J}_{\mathbb{A}}(u(t)) \in L^2([0, T], \mathbb{R}^d)$ because $\mathbb{J}_{\mathbb{A}}$ is nonexpansive,

and then $u \in \mathcal{A}(y)$ by construction. By Minty's surjectivity theorem (Minty, 1962), this implies that $\mathcal{A}$ is maximally monotone. Now because $\mathbb{J}_{\delta_n \mathbb{A}}(X_{\delta_n})$ converges to X strongly and $-G_{\delta_n} \in \mathcal{A}(\mathbb{J}_{\delta_n \mathbb{A}}(X_{\delta_n}))$ converges to $-G$ weakly, and the graph of a maximally monotone operator is closed under the strong-weak topology (Bauschke & Combettes, 2017, Proposition 20.38), we conclude that $-G \in \mathcal{A}(X)$, i.e.,

$$G(t) = Z(t) + \dot{X}(t) \in -\mathbb{A}(X(t)) \iff \dot{X}(t) \in -\left(Z(t) + \mathbb{A}(X(t))\right) \quad \text{a.e. in } t \in [0,T].$$

Finally, because $(T-t)\dot{Z}_\delta(t) = \dot{X}_\delta(t)$ for all $\delta > 0$ we have

$$(T-t)\dot{Z}_{\delta_n} = \dot{X}_{\delta_n} \xrightarrow{\text{w}} \dot{X} \quad \text{in } L^2([0,T], \mathbb{R}^d).$$

On the other hand, we had $\dot{Z}_{\delta_n} \xrightarrow{\text{w}} \dot{Z}$ in $L^2([0,T], \mathbb{R}^d)$, which implies

$$(T-t)\dot{Z}_{\delta_n} \xrightarrow{\text{w}} (T-t)\dot{Z}$$

because for any $c(t) \in L^2([0,T], \mathbb{R}^d)$, we have $(T-t)c(t) \in L^2([0,T], \mathbb{R}^d)$ as well. By uniqueness of the (weak) limit, we have $\dot{X} = (T-t)\dot{Z}$ a.e., so

$$\dot{Z}(t) = \frac{1}{T-t}\dot{X}(t) \in -\frac{1}{T-t}\left(Z(t) + \mathbb{A}(X(t))\right) \quad \text{a.e. in } t \in [0,T].$$

This shows that $\begin{pmatrix} X \\ Z \end{pmatrix}$ is indeed a solution of (90).

H.4.2. Behavior at the terminal time $t = T$

So far, we have successfully constructed an absolutely continuous solution $\begin{pmatrix} X \\ Z \end{pmatrix} : [0,T] \to \mathbb{R}^{2d}$. In this section we show that $X(t), Z(t)$ have two very favorable properties at the terminal time: X has left derivative 0, and $Z(T) \in -\mathbb{A}(X(T))$. To show this, we need:

Lemma H.8. *Let $\mathbb{A}\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be maximally monotone and let $\begin{pmatrix} X \\ Z \end{pmatrix}$ be the solution of (90) constructed in Theorem H.5. Then for almost every $t \in [0,T)$,*

$$\frac{1}{(T-t)^2}\left\|\dot{X}(t)\right\|^2 \le \frac{1}{T^2}\left\|m(\mathbb{A}(X_0))\right\|^2.$$

Proof. Let $\delta_n > 0$ be the sequence taken in the proof of Theorem H.5, for which the solutions $(X_{\delta_n}, Z_{\delta_n})$ to (92) with $\delta = \delta_n$ converge uniformly to (X, Z) and $(\dot{X}_{\delta_n}, \dot{Z}_{\delta_n})$ converge weakly to $(\dot{X}, \dot{Z})$. We have shown $\left\|\dot{X}_{\delta_n}(t)\right\| \le \frac{T-t}{T}\left\|m(\mathbb{A}(X_0))\right\|$ in Equation (93). Thus, the proof is done once we show

$$\left\|\dot{X}(t)\right\| \le \limsup_{n\to\infty}\left\|\dot{X}_{\delta_n}(t)\right\|$$

for almost every $t \in [0,T]$. (This would be straightforward if $\dot{X}$ were the pointwise limit of $\dot{X}_{\delta_n}$, but it is not the case; it is the weak limit. We need a careful argument, as presented below.)

Let D be any measurable subset of $[0,T]$. Since $\dot{X}_{\delta_{n_k}} \to \dot{X}$ weakly in $L^2([0,T], \mathbb{R}^d)$ and $\chi_D \dot{X} \in L^2([0,T], \mathbb{R}^d)$, we have

$$\begin{aligned}\int_D \left\|\dot{X}(t)\right\|^2 dt &= \int_0^T \left\langle \dot{X}(t), \chi_D(t)\dot{X}(t)\right\rangle dt = \lim_{n\to\infty}\int_0^T \left\langle \dot{X}_{\delta_n}(t), \chi_D(t)\dot{X}(t)\right\rangle dt \\ &= \lim_{n\to\infty}\int_D \left\langle \dot{X}_{\delta_n}(t), \dot{X}(t)\right\rangle dt \le \limsup_{n\to\infty}\int_D \left\|\dot{X}_{\delta_n}(t)\right\|\left\|\dot{X}(t)\right\| dt.\end{aligned} \tag{94}$$

Now from $\left\|\dot{X}_{\delta_n}(\cdot)\right\| \le \left\|m(\mathbb{A}(X_0))\right\|$ and $\left\|\dot{X}(\cdot)\right\| \in L^2([0,T], \mathbb{R}^d) \subset L^1([0,T], \mathbb{R}^d)$, we obtain

$$\left\|\dot{X}_{\delta_n}(\cdot)\right\|\left\|\dot{X}(\cdot)\right\| \le \left\|m(\mathbb{A}(X_0))\right\|\left\|\dot{X}(\cdot)\right\| \in L^1([0,T], \mathbb{R}^d).$$

Thus by reverse Fatou Lemma,

$$\limsup_{n\to\infty} \int_D \left\|\dot{X}_{\delta_n}(t)\right\| \left\|\dot{X}(t)\right\| dt \le \int_D \limsup_{n\to\infty} \left\|\dot{X}_{\delta_n}(t)\right\| \left\|\dot{X}(t)\right\| dt.$$

Combining the above inequality with (94) we obtain

$$\int_D \left(\limsup_{n\to\infty} \left\|\dot{X}_{\delta_n}(t)\right\| - \left\|\dot{X}(t)\right\|\right) \left\|\dot{X}(t)\right\| dt \ge 0.$$

As D was an arbitrary measurable subset of $[0,T]$, we conclude that for almost every $t \in [0,T]$,

$$\left(\limsup_{n\to\infty} \left\|\dot{X}_{\delta_n}(t)\right\| - \left\|\dot{X}(t)\right\|\right) \left\|\dot{X}(t)\right\| \ge 0 \quad \Longrightarrow \quad \left\|\dot{X}(t)\right\| \le \limsup_{n\to\infty} \left\|\dot{X}_{\delta_n}(t)\right\|.$$

□

Corollary H.9. *Let $\mathbb{A}\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be maximally monotone and let $\begin{pmatrix} X \\ Z \end{pmatrix}$ be the solution of (90) constructed in Theorem H.5. Then the following holds true:*

$$\lim_{t\to T^-} \frac{X(t) - X(T)}{t - T} = 0, \qquad -Z(T) \in \mathbb{A}(X(T)).$$

Proof. Using Lemma H.8, we obtain

$$\lim_{t\to T^-} \left\|\frac{X(t) - X(T)}{t - T}\right\| \le \lim_{t\to T^-} \int_t^T \frac{\left\|\dot{X}(s)\right\|}{T - t} ds \le \lim_{t\to T^-} \int_t^T \frac{\left\|\dot{X}(s)\right\|}{T - s} ds \le \lim_{t\to T^-} \int_t^T \frac{\|m(\mathbb{A}(X_0))\|}{T} ds = 0,$$

which proves the first property.

Next, because $(X(t), Z(t))$ satisfies (90) a.e. and $\|\dot{X}(t)\| \le \frac{T-t}{T} \|m(\mathbb{A}(X_0))\|$ a.e. (by Lemma H.8), we can take a sequence $t_k \in (0,T)$ such that $\lim_{k\to\infty} t_k = T$ and

$$-\left(\dot{X}(t_k) + Z(t_k)\right) \in \mathbb{A}(X(t_k)), \qquad \left\|\dot{X}(t_k)\right\| \le \frac{T - t_k}{T} \|\mathbb{A}(X_0)\|.$$

Then we have $\lim_{k\to\infty} \left\|\dot{X}(t_k)\right\| \le \lim_{k\to\infty} \frac{T-t_k}{T} \|\mathbb{A}(X_0)\| = 0$. On the other hand, $X(t_k) \to X(T)$ and $Z(t_k) \to Z(T)$ because X, Z are continuous. Finally, because the graph of $\mathbb{A}$ is closed in $\mathbb{R}^d \times \mathbb{R}^d$ (Bauschke & Combettes, 2017, Proposition 20.38), we conclude

$$-Z(T) = -\lim_{k\to\infty} \left(\dot{X}(t_k) + Z(t_k)\right) \in \mathbb{A}\left(\lim_{k\to\infty} X(t_k)\right) = \mathbb{A}(X(T)).$$

□

H.4.3. Convergence analysis

Based on the previous analyses, we can prove that the constructed solution has a Lyapunov function similar to that of Theorem 7.1 (the case of Lipschitz continuous operators).

Theorem H.10. *Let $\mathbb{A}\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be maximally monotone and $\operatorname{Zer} \mathbb{A} \neq \emptyset$. Let $\begin{pmatrix} X \\ Z \end{pmatrix}$ be the solution of (90) constructed in Theorem H.5. Then the function $V\colon [0,T) \to \mathbb{R}$ defined by*

$$V(t) = -\|Z(t) - Z(T)\|^2 + \frac{2}{T - t} \langle Z(t) - Z(T), X(t) - X(T)\rangle$$

is nonincreasing, and $\lim_{t\to T^-} V(t) = 0$. Furthermore, for $X_\star \in \operatorname{Zer} \mathbb{A}$,

$$\|m(\mathbb{A}(X(T)))\|^2 \le \|Z(T)\|^2 \le \frac{4\|X_0 - X_\star\|^2}{T^2}$$

where $m(\mathbb{A}(X(T)))$ is the minimum norm element of $\mathbb{A}(X(T))$.

Proof. Let $\delta_n > 0$ be the sequence taken in the proof of Theorem H.5. Let

$$V_{\delta_n}(t) = -\left\|Z_{\delta_n}(t) - Z_{\delta_n}(T)\right\|^2 + \frac{2}{T-t}\left\langle Z_{\delta_n}(t) - Z_{\delta_n}(T), X_{\delta_n}(t) - X_{\delta_n}(T)\right\rangle$$

for $t \in [0,T)$. As $\begin{pmatrix} X_{\delta_n} \\ Z_{\delta_n} \end{pmatrix}$ converges uniformly on $[0,T]$, for all $t \in [0,T)$ we have

$$V(t) = \lim_{n\to\infty} V_{\delta_n}(t).$$

Observe that $V_{\delta_n}(\cdot)$ is nonincreasing on $[0,T)$ for each n (by the result of Theorem 7.1 and $Z_{\delta_n}(T) = -\mathbb{A}(X_{\delta_n}(T))$, which follows from Corollary H.9 and uniqueness of the solution in the Lipschitz case). Therefore, for any $s,t \in [0,T)$ such that $s < t$, we have

$$V(s) = \lim_{n\to\infty} V_{\delta_n}(s) \geq \lim_{n\to\infty} V_{\delta_n}(t) = V(t),$$

which shows that V is nonincreasing. Furthermore, from $\lim_{t\to T^-} Z(t) = Z(T)$ (continuity of Z) and Corollary H.9, we have

$$\lim_{t\to T^-} V(t) = -\left\|\lim_{t\to T^-} Z(t) - Z(T)\right\|^2 - 2\left\langle \lim_{t\to T^-} Z(t) - Z(T), \lim_{t\to T^-} \frac{X(t) - X(T)}{t-T}\right\rangle = 0.$$

Therefore

$$0 = \lim_{t\to T^-} V(t) \leq V(0) = -\left\|Z(T)\right\|^2 - \frac{2}{T}\left\langle Z(T), X_0 - X(T)\right\rangle,$$

and Lemma 3.2 gives

$$\left\|Z(T)\right\|^2 \leq \frac{4\left\|X_0 - X_\star\right\|^2}{T^2}.$$

Finally, because $-Z(T) \in \mathbb{A}(X(T))$ by Corollary H.9, the left hand side is lower bounded by $\left\|m(\mathbb{A}(X(T)))\right\|^2$, which gives the desired convergence rate.

□

I. Omitted details from Section 8

I.1. Extragradient algorithm specification

In Figures 1a, 2a, we display comparison with Extragradient (EG) (Korpelevich, 1976). For completeness, we specify its definition here:

$$\begin{aligned} x_{k+1/2} &= x_k - \alpha \nabla_{\pm} \mathbf{L}(x_k) \\ x_{k+1} &= x_k - \alpha \nabla_{\pm} \mathbf{L}(x_{k+1/2}). \end{aligned}$$

I.2. Details of worst-case construction from Ouyang & Xu (2021)

For Figure 2a, we use the following construction due to Ouyang & Xu (2021), which has been used to establish complexity lower bounds:

$$A = \frac{1}{4}\begin{bmatrix} & & & & -1 & 1 \\ & & & \iddots & \iddots & \\ & & -1 & 1 & & \\ & -1 & 1 & & & \\ -1 & 1 & & & & \\ 1 & & & & & \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad b = \frac{1}{4}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \in \mathbb{R}^n, \quad g = \frac{1}{4}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^n,$$

and $G = 2A^\intercal A$. In Ouyang & Xu (2021), it is shown that $\|A\| \le \frac{1}{2}$ (so $\|G\| \le \frac{1}{2}$), so that the bilinear function

$$\mathbf{L}(u, v) = \frac{1}{2} u^\intercal G u - g^\intercal u - \langle Au - b, v\rangle$$

is 1-smooth.

I.3. H-dual algorithms produce identical terminal iterates for linear operators

In this section, we only consider explicit algorithms written in the form

$$x_{k+1} = x_k - \sum_{j=0}^{k} h_{k+1,j} \mathbb{A} x_j \tag{95}$$

where $\mathbb{A} : \mathbb{R}^d \to \mathbb{R}^d$ is a single-valued operator. If $\mathbb{A} = \nabla_{\pm}\mathbf{L}$, FEG and Dual-FEG are instances of this class if we identify $1/2$-indexed iterates with integer iterates in increasing order. If $\mathbb{A} = \nabla f$, constant step-size gradient descent or Nesterov's accelerated gradient method (Nesterov, 1983) are instances of this class. However, we do not need to restrict the discussion of this section to a specific operator class, because the following result does not rely on any particular property of the operator except for linearity.

Proposition I.1. *Suppose $\mathbb{A} : \mathbb{R}^d \to \mathbb{R}^d$ is linear, i.e., there exists a matrix $A \in \mathbb{R}^{d\times d}$ such that $\mathbb{A}x = Ax$ for all $x \in \mathbb{R}^d$. Let $N \ge 1$ and consider an algorithm defined by (95) for $k = 0, 1, \dots, N-1$. Then its H-dual algorithm, defined by*

$$\hat{x}_{k+1} = \hat{x}_k - \sum_{j=0}^{k} h_{N-j,N-k-1} \mathbb{A} x_j,$$

satisfies

$$x_N = \hat{x}_N$$

provided that $x_0 = \hat{x}_0$.

Proof. Because $\mathbb{A}$ is linear, we can explicitly describe x_k from the algorithm (95) via matrix polynomial in A and x_0. For $k = 1, 2, 3$, we have

$$\begin{aligned} x_1 &= x_0 - h_{1,0} A x_0 \\ x_2 &= x_1 - h_{2,0} A x_0 - h_{2,1} A x_1 \\ &= (x_0 - h_{1,0} A x_0) - h_{2,0} A x_0 - h_{2,1} A (x_0 - h_{1,0} A x_0) \\ &= x_0 - (h_{1,0} + h_{2,0} + h_{2,1}) A x_0 + h_{1,0} h_{2,1} A^2 x_0 \end{aligned}$$

and

$$\begin{aligned}
x_3 &= x_2 - h_{3,0}Ax_0 - h_{3,1}Ax_1 - h_{3,2}Ax_2 \\
&= (x_0 - (h_{1,0} + h_{2,0} + h_{2,1})Ax_0 + h_{1,0}h_{2,1}A^2x_0) - h_{3,0}Ax_0 - h_{3,1}A(x_0 - h_{1,0}Ax_0) \\
&\quad - h_{3,2}A(x_0 - (h_{1,0} + h_{2,0} + h_{2,1})Ax_0 + h_{1,0}h_{2,1}A^2x_0) \\
&= x_0 - (h_{1,0} + h_{2,0} + h_{3,0} + h_{2,1} + h_{3,1} + h_{3,2})Ax_0 \\
&\quad + (h_{1,0}h_{2,1} + h_{1,0}h_{3,1} + h_{1,0}h_{3,2} + h_{2,0}h_{3,2} + h_{2,1}h_{3,2})A^2x_0 - h_{1,0}h_{2,1}h_{3,2}A^3x_0.
\end{aligned}$$

The following lemma formally generalizes the pattern for these expressions:

Lemma I.2. *For linear operator $\mathbb{A}x = Ax$, the iterates of (95) satisfy*

$$x_k = \sum_{m=0}^{k} (-1)^m P(k,m) A^m x_0. \tag{96}$$

for $k = 1, \ldots, N$, where $P(k,0) = 1$ and for $m = 1, \ldots, k$,

$$P(k,m) = \sum_{\substack{i(1) \le j(2), \ldots, i(m-1) \le j(m) \\ i(m) \le k}} \prod_{\ell=1}^{m} h_{i(\ell),j(\ell)}. \tag{97}$$

Proof of Lemma I.2. We first clarify the definition of $P(k,m)$ by providing some examples. First, we have

$$P(1,1) = h_{1,0}, \qquad P(2,1) = h_{1,0} + h_{2,0} + h_{2,1}, \qquad P(3,1) = h_{1,0} + h_{2,0} + h_{3,0} + h_{2,1} + h_{3,1} + h_{3,2}.$$

This is because in (97), when $m = 1$, the constraints $i(1) \le j(2), \ldots, i(m-1) \le j(m)$ become vacuous so we add all $h_{i,j}$'s with $i \le k$ (note that always $j \le i - 1$). Next, we have

$$P(2,2) = h_{1,0}h_{2,1}, \qquad P(3,2) = h_{1,0}h_{2,1} + h_{1,0}h_{3,1} + h_{1,0}h_{3,2} + h_{2,0}h_{3,2} + h_{2,1}h_{3,2}.$$

Observe that by definition of P, we have to choose the (i,j) pairs within the product to satisfy the constraint $i(1) \le j(2)$, so if we choose $(i(1), j(1)) = (1,0)$ then we must have $j(2) \ge 1$, so $P(k,1)$ contains the products of $h_{1,0}$ with $h_{i,j}$'s satisfying $k \ge i > j \ge 1$. Similarly, if we choose $i(1) = 2$ then one must have $j(2) \ge 2$, and the only (i,j) pair satisfying $3 \ge i > j \ge 2$ is $(i,j) = (3,2)$, so $h_{2,0}h_{3,2}$ and $h_{2,1}h_{3,2}$ are the only terms within $P(3,2)$ that involve $h_{2,0}$ and $h_{2,1}$, respectively.

Because the constraint $j(\ell) < i(\ell)$ is implicit, the definition (97) is requiring the $(i(\ell), j(\ell))$ pairs consisting the product to satisfy

$$1 \le i(1) \le j(2) < \cdots < j(m-1) < i(m-1) \le j(m) < i(m) \le k.$$

Thus, when $m = k$, the only possible choice of indices is $i(\ell) = \ell$ and $j(\ell) = \ell - 1$, so

$$P(k,k) = h_{1,0}h_{2,1} \cdots h_{k,k-1}$$

and in particular, $P(3,3) = h_{1,0}h_{2,1}h_{3,2}$.

The above observations already show that (96) holds for $k = 1, 2, 3$. To prove the general statement, we use induction. Suppose (96) holds up to k, and consider

$$x_{k+1} = x_k - \sum_{n=0}^{k} h_{k+1,n} A x_n.$$

By induction hypothesis, we have

$$\begin{aligned}
x_{k+1} &= \sum_{m=0}^{k}(-1)^m P(k,m)A^m x_0 - \sum_{n=0}^{k} h_{k+1,n} A\left(\sum_{m=0}^{n}(-1)^m P(n,m)A^m x_0\right) \\
&= \sum_{m=0}^{k}(-1)^m P(k,m)A^m x_0 + \sum_{n=0}^{k}\sum_{m=0}^{n}(-1)^{m+1} h_{k+1,n} P(n,m)A^{m+1} x_0 \\
&= \sum_{m=0}^{k}(-1)^m P(k,m)A^m x_0 + \sum_{n=0}^{k}\sum_{m=1}^{n+1}(-1)^{m} h_{k+1,n} P(n,m-1)A^{m} x_0 \\
&= \sum_{m=0}^{k}(-1)^m P(k,m)A^m x_0 + \sum_{m=1}^{k+1}\sum_{n=m-1}^{k}(-1)^{m} h_{k+1,n} P(n,m-1)A^{m} x_0 \\
&= x_0 + h_{k+1,k}P(k,k)A^{k+1}x_0 + \sum_{m=1}^{k}(-1)^m\left(P(k,m) + \sum_{n=m-1}^{k} h_{k+1,n}P(n,m-1)\right)A^m x_0.
\end{aligned}$$

Observe that $h_{k+1,k}P(k,k) = h_{k+1,k}\left(h_{1,0}h_{2,1}\cdots h_{k,k-1}\right) = h_{1,0}h_{2,1}\cdots h_{k+1,k} = P(k+1,k+1)$. It remains to show that

$$P(k,m) + \sum_{n=m-1}^{k} h_{k+1,n}P(n,m-1) = P(k+1,m)$$

for $m = 1,\ldots,k$. Rewrite the right hand side:

$$\begin{aligned}
P(k+1,m) &= \sum_{\substack{i(1)\le j(2),\ldots,i(m-1)\le j(m)\\ i(m)\le k+1}} \prod_{\ell=1}^{m} h_{i(\ell),j(\ell)} \\
&= \sum_{\substack{i(1)\le j(2),\ldots,i(m-1)\le j(m)\\ i(m)\le k}} \prod_{\ell=1}^{m} h_{i(\ell),j(\ell)} + \sum_{\substack{i(1)\le j(2),\ldots,i(m-1)\le j(m)\\ i(m)= k+1}} \prod_{\ell=1}^{m} h_{i(\ell),j(\ell)} \\
&= P(k,m) + \sum_{n=1}^{k} \sum_{i(1)\le j(2),\ldots,i(m-1)\le j(m)=n} h_{k+1,n} \prod_{\ell=1}^{m-1} h_{i(\ell),j(\ell)} \\
&= P(k,m) + \sum_{n=1}^{k} h_{k+1,n} \sum_{i(1)\le j(2),\ldots,i(m-1)\le j(m)=n} \prod_{\ell=1}^{m-1} h_{i(\ell),j(\ell)} \\
&= P(k,m) + \sum_{n=m-1}^{k} h_{k+1,n} \sum_{\substack{i(1)\le j(2),\ldots,i(m-2)\le j(m-1)\\ i(m)\le n}} \prod_{\ell=1}^{m-1} h_{i(\ell),j(\ell)} \qquad (98)
\end{aligned}$$

where the last equality holds because for $n < m-1$, the inner summation with respect to $i(1) \le j(2),\ldots,i(m-2) \le j(m-1)$ is vacuous as it is impossible to choose $i(\ell)$'s satisfying $1 \le i(1) < i(2) < \cdots i(m-1) \le n$. Now replace the inner summation in (98) using the definition of $P(n,m-1)$:

$$P(k+1,m) = \sum_{n=m-1}^{k} h_{k+1,n}P(n,m-1).$$

This completes the induction. □

Lemma I.2 explicitly characterizes x_k in terms of matrix polynomial. Now the proof is done once we show that the explicit expression for x_N is invariant under anti-diagonal transposing, i.e., the replacement $h_{k,j} \mapsto h_{N-j,N-k}$. More precisely, we

have to show that $P(N,m) = \hat{P}(N,m)$, where

$$\hat{P}(N,m) = \sum_{\substack{i(1)\le j(2),\ldots,i(m-1)\le j(m)\\ i(m)\le N}} \prod_{\ell=1}^{m} h_{N-j(\ell),N-i(\ell)} \tag{99}$$

so that

$$\hat{x}_N = \sum_{m=0}^{N} (-1)^m \hat{P}(N,m) A^m \hat{x}_0.$$

Once this is done, provided that $\hat{x}_0 = x_0$, we can conclude

$$\hat{x}_N = \sum_{m=0}^{N} (-1)^m \hat{P}(N,m) A^m \hat{x}_0 = \sum_{m=0}^{N} (-1)^m P(N,m) A^m x_0 = x_N.$$

In the product within (99), we make the substitution $i'(\ell) = N - j(\ell)$ and $j'(\ell) = N - i(\ell)$. Then for $\ell = 1,\ldots,m-1$,

$$i(\ell) \le j(\ell+1) \iff N - j(\ell+1) \le N - i(\ell) \iff i'(\ell+1) \le j'(\ell)$$

and the condition $i(m) \le N$ is vacuous (can be dropped) because we are considering $h_{i,j}$ only for $i = 1,\ldots,N$. Thus,

$$\hat{P}(N,m) = \sum_{i'(m)\le j'(m-1),\ldots,i'(2)\le j'(1)} \prod_{\ell=1}^{m} h_{i'(\ell),j'(\ell)}.$$

Finally, reversing the order of $(i'(1),j'(1)),\ldots,(i'(m),j'(m))$ via substitution $i(\ell) = i'(m+1-\ell), j(\ell) = j'(m+1-\ell)$ we obtain

$$\hat{P}(N,m) = \sum_{i(1)\le j(2),\ldots,i(m-1)\le j(m)} \prod_{\ell=1}^{m} h_{i(\ell),j(\ell)} = P(N,m)$$

which concludes the proof.

□

I.4. Faster convergence of the Dual-Anchor ODE with strongly monotone operators

In this section, we show that the Dual-Anchor ODE converges much more rapidly than the primal Anchor ODE for the subclass of strongly monotone operators, which suggests a new potential value of the dual algorithms. While the Dual-Anchor ODE can only guarantee $\|\mathbb{A}(X(T))\| = \mathcal{O}(\epsilon)$ at the terminal time $T = \Omega(\frac{1}{\epsilon})$ when $\mathbb{A}$ is merely monotone, when $\mathbb{A}$ is strongly monotone and Lipschitz, the Dual-Anchor ODE achieves $\|X(t) - X(T)\| = \mathcal{O}(\epsilon)$ and $\|\mathbb{A}(X(t))\| = \mathcal{O}(\epsilon)$ at $t = \mathcal{O}(\log\frac{1}{\epsilon}) \ll T$ and makes negligible progress thereafter (similar to the pattern shown in Figure 2b). This allows one to apply *early stopping*; instead of waiting until the terminal time, we stop and return $X(t)$. On the other hand, this is not the case for primal Anchor ODE which, even when $\mathbb{A}$ is strongly monotone, behaves conservatively and converges no faster than the $\|\mathbb{A}(X(t))\|^2 = \mathcal{O}(1/t^2)$ rate. Below we provide the details.

Fast decay of $\|\mathbb{A}(X(t))\|$ and $\|X(t) - X(T)\|$ under strong monotonicity. We established in Lemma H.3 that $\Psi(t) = \frac{1}{(T-t)^2}\left\|\dot{X}(t)\right\|^2$ is nonincreasing by deriving

$$\dot{\Psi}(t) = -\frac{2}{(T-t)^2}\left\langle \frac{d}{dt}\mathbb{A}(X(t)), \dot{X}(t)\right\rangle.$$

If $\mathbb{A}$ is μ-strongly monotone for some $\mu > 0$, then the above identity implies

$$\dot{\Psi}(t) \le -\frac{2\mu}{(T-t)^2}\left\|\dot{X}(t)\right\|^2 = -2\mu\Psi(t)$$

so by Grönwall's inequality,

$$\Psi(t) \le \Psi(0)e^{-2\mu t} = \frac{\|\mathbb{A}(X_0)\|^2}{T^2}e^{-2\mu t}$$

for any $t \in (0, T)$. Reorganizing, we obtain

$$\left\|\dot{X}(t)\right\| = \sqrt{(T-t)^2\Psi(t)} \le \frac{T-t}{T}e^{-\mu t}\|\mathbb{A}(X_0)\| \le e^{-\mu t}\|\mathbb{A}(X_0)\|. \tag{100}$$

Complexity analysis. Suppose the desired accuracy level is $\|\mathbb{A}(\cdot)\| \le \epsilon$, so that we choose $T = \Theta(\frac{1}{\epsilon})$ according to the worst-case guarantee $\|\mathbb{A}(X(T))\| = \mathcal{O}\left(\frac{1}{T}\right)$. Given that $\mathbb{A}$ is μ-strongly monotone and L-Lipschitz, one achieves

$$\left\|\dot{X}(t)\right\| = \mathcal{O}\left(\frac{1}{LT^2}\right) \quad \text{at } t = \Theta\left(\frac{1}{\mu}\log LT\right) = \Theta\left(\frac{1}{\mu}\log\frac{L}{\epsilon}\right)$$

by (100) and then

$$\|X(t) - X(T)\| \le \int_t^T \left\|\dot{X}(s)\right\| ds \le \int_t^T \frac{T-s}{T-t}\left\|\dot{X}(t)\right\| ds \le T\mathcal{O}\left(\frac{1}{LT^2}\right) = \mathcal{O}\left(\frac{1}{LT}\right) = \mathcal{O}\left(\frac{\epsilon}{L}\right),$$

where the second inequality uses $\Psi(s) \le \Psi(t)$ for $s \ge t$. Finally, this implies

$$\|\mathbb{A}(X(t))\| \le \|\mathbb{A}(X(T))\| + \mathcal{O}\left(L\|X(t) - X(T)\|\right) = \mathcal{O}(\epsilon).$$

Anchor ODE is no faster than $\mathcal{O}\left(\frac{1}{t}\right)$ under strong monotonicity. Suppose $\mathbb{A}$ is linear ($\mathbb{A}x = Ax$ for some matrix A) and assume A is invertible. In this case the solution to the Anchor ODE can be explicitly characterized (Suh et al., 2023):

$$X_{\text{anchor}}(t) = \frac{1}{t}A^{-1}\left(I - e^{-tA}\right)X_0.$$

Now if A is μ-strongly monotone, then $A - \mu I$ is monotone and $0 \in \text{Zer}\,(A - \mu I)$, so we have

$$\begin{aligned}
&\frac{d}{dt}\left\|e^{-t(A-\mu I)}X_0\right\|^2 = -2\left\langle (A-\mu I)e^{-t(A-\mu I)}X_0, e^{-t(A-\mu I)}X_0\right\rangle \le 0 \\
&\implies \|X_0\| \ge \left\|e^{-t(A-\mu I)}X_0\right\| = \left\|e^{\mu t I}e^{-tA}X_0\right\| = e^{\mu t}\left\|e^{-tA}X_0\right\|,
\end{aligned}$$

which implies

$$\|AX_{\text{anchor}}(t)\| = \frac{1}{t}\left\|\left(I - e^{-tA}\right)X_0\right\| \ge \frac{1}{t}(1 - e^{-\mu t})\|X_0\|.$$

Therefore, the convergence of the Anchor ODE is generally not faster than $\mathcal{O}\left(\frac{1}{t}\right)$ even with strongly monotone (and linear) operators.