Optimized methods for composite optimization: a reduction perspective

Jinho Bok Jason M. Altschuler

Department of Statistics and Data Science, University of Pennsylvania

July 1, 2025

Abstract

Recent advances in convex optimization have leveraged computer-assisted proofs to develop *optimized* first-order methods that improve over classical algorithms. However, each optimized method is specially tailored for a particular problem setting, and it is a well-documented challenge to extend optimized methods to other settings due to their highly bespoke design and analysis. We provide a general framework that derives optimized methods for composite optimization *directly* from those for unconstrained smooth optimization. The derived methods naturally extend the original methods, generalizing how proximal gradient descent extends gradient descent. The key to our result is certain algebraic identities that provide a unified and straightforward way of extending convergence analyses from unconstrained to composite settings. As concrete examples, we apply our framework to establish (1) the phenomenon of stepsize acceleration for proximal gradient descent; (2) a convergence rate for the proximal optimized gradient method [47] which is faster than FISTA [8]; (3) a new method that improves the state-of-the-art rate for minimizing gradient norm in the composite setting.

A preliminary version [9] of this paper was accepted at the 38th Conference on Learning Theory (COLT 2025). This paper develops a general framework that recovers the results of [9] as a special case. Email: `{jinhobok, alts}@upenn.edu`.

Contents

1 Introduction

First-order methods such as gradient descent (GD) and Nesterov's accelerated gradient descent (AGD) [38] are widely popular in modern large-scale optimization. Recently, the optimization community has devoted significant effort to developing *optimized* methods which have faster convergence rates than classical algorithms such as GD or AGD for fundamental settings such as unconstrained smooth convex optimization (i.e., $\min_x f(x)$ where f is convex and smooth). A celebrated example is the *optimized gradient method* (OGM) [31]: in terms of worst-case convergence rate, this method is not only faster than AGD (by a constant factor), but moreover exactly achieves the best possible rate among all black-box first-order methods [15]. Another notable example is *stepsize-accelerated GD*, which incorporates time-varying stepsizes (but no momentum) and attains a provably faster rate than the traditional constant-stepsize GD [1, 2, 3, 23, 25].

However, it remains largely open whether optimized methods can be extended beyond the settings they were originally designed for. A central challenge is that the design and analysis of optimized methods are heavily tailored to their particular settings and, as a result, often lack flexible structures or properties that are applicable to other settings. A related challenge is that the design and analysis of optimized methods are based on computer-assisted proofs which are typically difficult to interpret, let alone adapt to other settings. In contrast, simpler algorithms such as GD and AGD have been successfully extended to many settings [4, 12, 20, 21, 28, 29] due to versatile components of their design and analysis, for example the descent lemma [39] and momentum [8, 18, 44, 52]. These challenges raise a fundamental question:

Is there a unified way to extend optimized methods to other settings?

1.1 Contribution

We answer this question in the affirmative for the fundamental setting of *composite convex optimization*, i.e., $\min_x f(x) + h(x)$ where f is convex and smooth, and h is convex (but potentially nonsmooth) and can be accessed via its proximal operator $\text{prox}_{\alpha h}(x) = \text{argmin}_z\{\alpha h(z) + \frac{1}{2}\|z - x\|^2\}$. The composite setting is a strict generalization of unconstrained optimization (by letting $h = 0$) and constrained optimization (by letting $h = \iota_{\mathcal{K}}$ be the indicator of a constraint set $\mathcal{K}$), and captures regularization-based problems prevalent in machine learning, signal processing, and statistics [11].

Conceptually, we develop a reduction-style argument which shows that *optimized methods in the unconstrained setting are also optimized in the composite setting.* This allows us to establish results of the following type: if an algorithm $\mathcal{A}$ for unconstrained convex optimization has convergence rate

$$f(x_n) - f(x_*) \leqslant \tau_n \|x_0 - x_*\|^2 \,,$$

then there is a *composite extension* $\mathcal{T}(\mathcal{A})$ of this algorithm for composite optimization that has convergence rate

$$f(x_n) + h(x_n) - f(x_*) - h(x_*) \leqslant O(\tau_n)\|x_0 - x_*\|^2 \,.$$

This reduction similarly applies to the alternative performance metric of gradient norm. The new algorithm $\mathcal{T}(\mathcal{A})$ extends the original algorithm $\mathcal{A}$ in a simple and natural way that generalizes how proximal gradient descent extends gradient descent; see the technical overview section below.

We believe that this unified framework provides a useful viewpoint for studying optimized first-order methods, since it reduces the design and analysis of methods for one setting (composite) to another setting (unconstrained). This is in contrast to prior approaches, where the design and analysis of each optimized method are typically done in a case-by-case fashion for each particular setting; see the discussion of prior work in Section 1.3.

Our unified framework leads to convergence rates that are competitive with highly optimized algorithms and, in some cases, yields state-of-the-art complexity guarantees and answers open problems. As concrete examples, we apply our framework to establish the following:

- **Stepsize-based acceleration of proximal GD.** We answer the open questions of [3, 23] by showing that stepsize-based acceleration is possible for proximal GD. That is, we show that proximal GD can achieve accelerated rates with a judicious choice of stepsizes—without any other modifications to the algorithm (e.g., momentum). Previous results for stepsize-based acceleration were limited to the setting of GD for unconstrained smooth convex optimization, and it was unknown whether this phenomenon was possible in constrained or composite settings. See [9, Section 1] for a comprehensive discussion of this open problem and the challenges. We show a rate of $O(1/n^{\log_2(1+\sqrt{2})}) \approx O(1/n^{1.2716})$, which improves over the classical $O(1/n)$ guarantee that is tight for proximal GD with constant stepsizes [50]. This asymptotic rate is conjecturally optimal even for the simpler setting of vanilla GD for unconstrained optimization [2, 3, 24, 25]; hence, our rate is also conjecturally optimal. Appealingly, our framework enables us to use the same stepsizes that were developed for accelerating vanilla GD—an approach that is natural, yet was previously unclear how to analyze beyond the original setting.

- **Proximal OGM.** Applying our framework to OGM yields an accelerated rate of $O(1/n^2)$ with nearly-optimal constant factor. This rate is faster than all prior methods except OptISTA, an exactly optimal method with a computer-assisted design that was recently developed specifically for this setting [27].

Unconstrained setting

Algorithm \ Performance Metric	Objective function	Gradient norm
Stepsize-accelerated GD	$O(1/n^{\log_2(1+\sqrt{2})})$ [2]	$O(1/n^{\log_2(1+\sqrt{2})})$ [25]
OGM(-G)	$O(1/n^2)$ [31]	$O(1/n^2)$ [33]

$\Downarrow$ **Composite extension** (Definition 2.2)

Composite setting

Algorithm \ Performance Metric	Objective function	Gradient norm
Stepsize-accelerated proximal GD	$O(1/n^{\log_2(1+\sqrt{2})})$ (Theorem 4.2)	$O(1/n^{\log_2(1+\sqrt{2})})$ (Theorem 5.3)
Proximal OGM(-G)	$O(1/n^2)$ (Theorem 4.7)	$O(1/n^2)$ (Theorem 5.5)

Table 1: Summary of applications. We develop a unified approach for extending optimized methods from unconstrained to composite settings, extending the asymptotic rate for each combination of algorithm and peformance metric.

- **Proximal OGM-G.** Applying our framework to OGM-G, a "gradient norm" version of OGM [33], yields the state-of-the-art convergence rate for minimizing gradient norm in the composite setting. This rate improves over the previous best guarantee [34] by a factor of nearly 10.

See Table 1 for a summary, and see Sections 4 and 5 for formal statements and further details.

1.2 Overview of framework

Here we overview our reduction-style approach for extending first-order methods from unconstrained to composite settings. We focus on the main conceptual ideas here; see Section 2 for formal statements.

Design of composite extension. Let $\mathcal{A}$ be a first-order method for the unconstrained setting. We propose a simple, unified way of designing an algorithm $\mathcal{T}(\mathcal{A})$ for the composite setting. The derived algorithm $\mathcal{T}(\mathcal{A})$ extends $\mathcal{A}$ in a way that generalizes how proximal GD extends GD. To explain this, recall that proximal GD updates as $x_{t+1} = \text{prox}_{\alpha_t h}(x_t - \alpha_t \nabla f(x_t))$. By definition of the proximal operator, this is equivalent to

$$x_{t+1} = x_t - \alpha_t(\nabla f(x_t) + s_{t+1}), \text{ where } s_{t+1} \in \partial h(x_{t+1}) .$$

One can view this proximal GD update as GD, except with $\nabla f(x_t)$ replaced by $\nabla f(x_t) + s_{t+1}$. Our proposed extension $\mathcal{T}(\mathcal{A})$ generalizes this: whenever a gradient iterate $\nabla f(x_t)$ appears in the original method $\mathcal{A}$, we replace it with $\nabla f(x_t) + s_{t+1}$ for the composite setting.

Analysis of composite extension. While the *design* of these composite algorithms is simple and natural, its *analysis* is nontrivial. This is the main content of the paper. Our starting point is existing analyses for optimized methods in the unconstrained setting. These have been driven by a powerful technique known as the *performance estimation problem* (PEP) [16]; in this framework, a dual solution of a certain semidefinite program provides a proof of a given algorithm's convergence rate. Specifically, in order to establish a convergence rate of the form $f(x_n) - f(x_*) \leqslant \tau_n \|x_0 - x_*\|^2$ for an algorithm $\mathcal{A}$, a dual solution to the semidefinite program provides multipliers $\lambda_{ij} \geqslant 0$ and a sum-of-squares (SOS) quadratic polynomial $P \geqslant 0$ such that the following identity holds:

$$\tau_n \|x_0 - x_*\|^2 - (f(x_n) - f(x_*)) = \sum_{i,j} \lambda_{ij} Q_{ij} + P \,, \tag{1.1}$$

where $Q_{ij} \geqslant 0$ is a quadratic polynomial in the iterates x_i, x_j and the first-order information of f at these points. This identity certifies the desired convergence rate since the right hand side is nonnegative.

Our analysis is based on a crucial observation in this template (1.1): *optimized methods have simple solutions.* In particular, for unconstrained optimized algorithms, the sum-of-squares term is often just a single square—rather than the sum of multiple squares that cannot be collapsed into a single square. In other words, the corresponding quadratic form is of rank 1. This core property was observed in [45, Chapter 5] for many different types of optimized methods, and it is an interesting open question in itself to understand if this phenomenon hints at an underlying general theory for optimized methods; see also the discussion of future work in Section 6.

Surprisingly, this common structure alone is informative enough for us to characterize a (candidate) dual solution in the composite setting. In this setting, a PEP-based approach tries to find multipliers $\lambda_{ij} \geqslant 0, \mu_{ij} \geqslant 0$ and an SOS quadratic polynomial $P \geqslant 0$ such that the following identity holds:

$$\tau_n \|x_0 - x_*\|^2 - (f(x_n) + h(x_n) - f(x_*) - h(x_*)) = \sum_{i,j} \lambda_{ij} Q^f_{ij} + \sum_{i,j} \mu_{ij} Q^h_{ij} + P\,, \tag{1.2}$$

where now we have pairs of valid inequalities $Q^f_{ij} \geqslant 0$ and $Q^h_{ij} \geqslant 0$ for the first-order information at f and h, respectively. The key difficulty in this extended template (1.2) is identifying the solutions, i.e., the quantities λ_{ij}, μ_{ij}, and P. This is in large part because of the additional complexity from h, which appears in (1.2) but not in (1.1). Finding such PEP solutions is a well-documented challenge; for example, the pioneering work on PEP for the composite setting [47] writes:

> *"... algorithmic analyses using [PEP] are intrinsically limited by our ability to solve semidefinite problems, both numerically ... or analytically .... Therefore, any idea leading to (convex) programs that are easier to solve while maintaining reasonable guarantees would be very advantageous."*

We overcome this challenge by providing closed-form expressions for the quantities in the composite identity (1.2) in terms of the quantities in the unconstrained identity (1.1). This lets us avoid solving (1.2) from scratch, and instead we borrow solutions from (1.1). In particular, an important starting point is that we use the *same* λ_{ij} from the unconstrained setting. While this approach is seemingly straightforward, the corresponding analysis is rather subtle since the corresponding terms—namely, $\lambda_{ij} Q_{ij}$ in (1.1) and $\lambda_{ij} Q^f_{ij}$ in (1.2)—are *not* equal. Indeed, Q^f_{ij} has additional terms involving h in the composite setting (see Definition 2.4 for details). Our analysis establishes that these differences can be offset with a careful construction of the multipliers μ_{ij} and the sum-of-squares term, both of which are new and different from the multipliers λ_{ij} and the sum-of-squares term in the unconstrained setting.

Candidate solution. Altogether, this provides a general reduction-style approach of obtaining convergence guarantees, since the resulting formulae for λ_{ij}, μ_{ij}, P are explicit, satisfy the identity (1.2), and can be applied to any method with the aforementioned rank-1 property.

We refer to this as a "reduction-style approach" since it is not an end-to-end reduction. In particular, one still needs to check the feasibility constraints[1] that $\mu_{ij} \geqslant 0$ and that P is actually SOS. Verifying these two conditions must be done in an algorithm-specific manner but is conceptually simple because of the closed-form expressions. See Sections 4 and 5 for multiple examples. We emphasize that this is much simpler than solving (1.2) from scratch—since that requires solving for λ_{ij}, μ_{ij}, P; verifying the identity (1.2); and checking the feasibility constraints $\lambda_{ij} \geqslant 0$, $\mu_{ij} \geqslant 0$, and that P is SOS.

Sum-of-squares structure. We comment in particular on the SOS verification of P, since our framework provides a conceptually new approach for accomplishing this. Given λ_{ij} and μ_{ij}, P is defined as the residual in the identity (1.2). Verifying that P is SOS is in general a key challenge for PEP-based analyses [16, 27, 47], especially in the composite setting since then P is typically of *high rank*, i.e., any decomposition as a sum of squares (if one exists) requires $\Omega(n)$ squares. Finding such a decomposition is challenging, as that amounts to verifying positive semidefiniteness of the corresponding coefficient matrix, the entries of which have complicated algebraic expressions.

We provide a new approach for this verification: we combine certain aspects of the analysis for the unconstrained setting (1.1) and the proximal point method [47]. In the former, as mentioned, the matrix is rank-1. In the latter, while the matrix has high rank, it has a simple Laplacian structure which implies positive semidefiniteness. We show that modulo a Schur complement, the matrix in our analysis is a sum of

[1] Feasibility of $\lambda_{ij} \geqslant 0$ is automatically guaranteed from the re-use of λ_{ij} in our approach.

a rank-1 matrix and a Laplacian matrix. Ultimately, this reduces checking positive semidefiniteness of P to checking that an explicit rank-1 perturbation of a Laplacian matrix remains Laplacian; in our applications, this amounts to simply comparing only a couple entries. We remark that more generally, such a combination of structures from different base algorithms may be useful for PEP-type analyses.

Performance metrics. So far, we have focused on convergence rates that are measured in terms of suboptimality of the objective function. Our approach applies in a nearly identical manner when convergence rates are instead measured in terms of approximate stationarity, i.e., making the (sub)gradient norm small. We show results of the following type: if an algorithm $\mathcal{A}'$ for unconstrained convex optimization has convergence rate

$$\|\nabla f(x_n)\|^2 \leqslant \tau_n'(f(x_0) - f(x_*))\,,$$

then the composite extension $\mathcal{T}(\mathcal{A}')$ has convergence rate

$$\|\nabla f(x_n) + s_n\|^2 \leqslant O(\tau_n')(f(x_0) + h(x_0) - f(x_*) - h(x_*))$$

for $s_n \in \partial h(x_n)$. Note that we use the same composite extension $\mathcal{T}$ for these results. For this setting of gradient norm minimization, the two building blocks we use for the SOS verification have slightly different forms: for the unconstrained setting, the residual term is zero rather than a single square; and for the proximal point method, the corresponding matrix is diagonally dominant rather than Laplacian [26]. The rest of the analysis is conceptually identical.

1.3 Prior work

Performance estimation problem (PEP). The PEP framework, pioneered by [16], formulates the worst-case performance (i.e., convergence rate) of a given algorithm $\mathcal{A}$ as a semidefinite program. In this auxiliary optimization problem, the objective function corresponds to the performance metric for $\mathcal{A}$, and the constraints are on the function class and the iterates being updated by $\mathcal{A}$. The primal searches over worst-case problem instances, while the dual searches over proofs of the convergence rate as described in the previous section. In many settings of convex optimization, this formulation as a semidefinite program is *tight* in that any proof for convergence rate can be expressed as a solution of the dual [48].

Among numerous applications of PEP (see e.g., [18, 47]), there have been two main streams of work. One line of work establishes tight rates for existing algorithms, such as gradient descent [5, 16, 36, 42], proximal point/gradient method [47, 49], splitting methods [43], ADMM [56], and Chambolle-Pock [10], among others. Another line of work develops new optimized methods that are either asymptotically or exactly optimal in each setting, including: smooth (strongly) convex optimization [31, 33, 46], nonsmooth convex optimization [17, 22, 57], composite optimization [27], fixed-point iteration [41], minimax optimization [53], and monotone inclusion [30]. Our framework is situated at the interface of these two directions: it yields nearly tight rates by extending existing optimized methods and their proofs to a new setting. In particular, we provide explicit formulae for the PEP-based proofs (a challenge for the first line of work) and show how optimized methods can be naturally extended to a more general setting (a challenge for the second line of work).

An emerging area of study which uses PEP is stepsize-accelerated GD [1, 2, 3, 13, 19, 23, 24, 25, 51, 58, 59], which seeks to improve the rate of GD by only changing the stepsizes. Classically, this was only known to be possible for minimizing convex quadratics [55]. The recent line of work extends this improvement beyond the quadratic setting by using time-varying stepsize schedules that are nonmonotone and use exceedingly large steps. In particular, [2] showed the "silver convergence rate" $O(1/n^{\log_2(1+\sqrt{2})})$ which is conjectured to be asymptotically optimal among all possible stepsize schedules. See [9, Section 1.3] for a recent overview of this very active literature on stepsize-based acceleration.

Composite optimization. Composite optimization arises in many applications due to the flexibility afforded by the non-smooth component h—in particular as a regularization penalty. For example, in the common setting where h is the ℓ_1 norm, proximal GD is known as ISTA [14]. The seminal work of [8] introduced FISTA, a widely popular algorithm with $O(1/n^2)$ rate. Since then, different variants of FISTA have been proposed [7, 32]; see also [6, Chapter 10] for an overview. Notably, the recent paper [27] presented an optimized method OptISTA whose convergence rate is faster than FISTA and moreover exactly matches the lower bound for black-box first-order methods. Several works have also made progress in the task of gradient norm minimization in the composite setting [34, 37], achieving $O(1/n^2)$ rate; however, the optimal constant factor is unknown.

Our framework provides a general way to design and analyze methods in this setting. This leads to new algorithms as well as new rates for existing algorithms. Here, we further contextualize the applications of our framework. Our composite extension of OGM recovers POGM, a method that appeared in the original paper on PEP for composite optimization [47]. However, that paper only presents numerical bounds from PEP; we provide rigorous convergence analyses here. We remark that while the design and analysis are simpler for POGM than OptISTA, it does not achieve the exactly optimal rate. This is not just an artefact of our technique and it was numerically observed that the POGM is suboptimal, though only by a small multiplicative factor of roughly 1.12 [47]. Our theoretical result achieves a rate for POGM with a mild multiplicative factor of roughly 1.29 compared to the exactly optimal rate. Furthermore, our composite extension of OGM-G achieves the state-of-the-art rate, improving over the result of [34, Section D.4] by a factor of roughly 9.30. For stepsize-based acceleration, previous results were known only for GD (in the unconstrained setting), and it was an open problem if this phenomenon extends to projected GD (in the constrained setting) or proximal GD (in the composite setting). This was posed as an open problem in [3, 23]; see [9, Section 1] for a detailed discussion of the challenges. We resolve this question by showing that the silver stepsizes for vanilla GD enable the same asymptotic rates for proximal GD.[2]

Connections between algorithms. A recent line of PEP-related work has investigated relations between different algorithms. For example, [37] identified a common geometric structure in accelerated algorithms, and [54] identified a common property in accelerated minimax algorithms. [40] established a connection between AGD and OGM, along with certain extensions of OGM. [34, 35] developed the notions of H-duality and mirror duality, which are one-to-one correspondences between certain algorithms, one of which is designed for minimizing the objective function and the other for minimizing the gradient norm.

This paper shares some similarities with H-duality (and mirror duality) [34, 35], in that they also reduce the design and analysis of one method to another method. However, the details are quite different. Most importantly, our results are orthogonal—in fact complementary—to theirs as we relate algorithms for *different problem settings with the same performance metric*, whereas H-duality relates algorithms for *the same problem setting with different performance metrics.* There are also other differences. H-duality provides a fully-fledged reduction whereas our framework is not fully black-box; however, our framework is general enough to apply to both stepsize-accelerated GD and OGM, which have markedly different proof structures (in terms of the multipliers $\{\lambda_{ij}\}$), while H-duality only applies to methods with specific proof structures.[3]

2 Main results

2.1 Composite extension and preliminaries

Throughout, we consider first-order methods with iterates $x_0, x_1, \dots, x_n$, where n is the total number of iterations. We let x_* denote an optimal point for the relevant optimization problem. For a smooth convex function f and convex function h, we use the shorthands $F := f + h$, $f_i := f(x_i)$, $g_i := \nabla f(x_i)$, $h_i := h(x_i)$, $F_i := F(x_i)$, and we let s_i denote a subgradient of h at x_i, for $i \in \{0, 1, \dots, n, *\}$.[4] Without loss of generality (by normalization), the smoothness parameter of f is assumed to be 1 throughout. Vectors are always vertical, and we denote the sum of the entries of a vector v by $\sum v$.

Our result is stated for a general class of first-order methods. In the literature, this class is known as fixed-step first-order methods.

[2] Part of these results appeared in the preliminary conference version [9] of the present paper. The proofs and rates in [9] are mathematically equivalent to Theorem 4.2 here, but our proof here is much simpler and based on the general framework developed in this paper. This framework also lets us extend these results both to minimizing gradient norm (Theorem 5.3), as well to analyzing momentum-based methods (Theorems 4.7 and 5.5).

[3] A certain analog of H-duality for stepsize-accelerated GD recently appeared in a different work [24, Section 3.1]. However, it only applies to specific choices of stepsizes.

[4] Unless otherwise specified, $*$ is always included when referring to the set of "all indices". $*$ is never included when indices are compared by their values (e.g., $i \geqslant 0, i \leqslant n$), and indices $* - 1, * + 1$ are defined to be equal to $*$.

Definition 2.1 (Stepsize matrix and first-order methods)**.** *An* n-stepsize matrix H *is an upper triangular matrix indexed as*

$$H := \begin{bmatrix} \alpha_{1,0} & \cdots & \alpha_{n,0} \\ & \ddots & \vdots \\ & & \alpha_{n,n-1} \end{bmatrix},$$

where $\alpha_{k,k-1} \neq 0$ *for all* $1 \leqslant k \leqslant n$*. The* ($n$-step) first-order method with H *is defined as*

$$\begin{aligned} x_1 &= x_0 - \alpha_{1,0} g_0, \\ x_2 &= x_1 - \alpha_{2,0} g_0 - \alpha_{2,1} g_1, \\ &\vdots \\ x_n &= x_{n-1} - \alpha_{n,0} g_0 - \cdots - \alpha_{n,n-1} g_{n-1}. \end{aligned}$$

The condition $\alpha_{k,k-1} \neq 0$ ensures that the iterates are not redundant (i.e., not a linear combination of previous iterates) and H is invertible.

Given a first-order method for the unconstrained setting (i.e., a stepsize matrix H), we define its *composite extension* by replacing g_t with $g_t + s_{t+1}$.

Definition 2.2 (Composite extension)**.** *Consider a first-order method with* n*-stepsize matrix* H*. Its* composite extension *is*

$$\begin{aligned} x_1 &= x_0 - \alpha_{1,0}(g_0 + s_1), \\ x_2 &= x_1 - \alpha_{2,0}(g_0 + s_1) - \alpha_{2,1}(g_1 + s_2), \\ &\vdots \\ x_n &= x_{n-1} - \alpha_{n,0}(g_0 + s_1) - \cdots - \alpha_{n,n-1}(g_{n-1} + s_n). \end{aligned}$$

Remark 2.3 (Implementation)**.** *An equivalent, implementable form of the composite extension (Definition 2.2) that only involves gradient and proximal oracles is*

$$\begin{aligned} x_1 &= \mathrm{prox}_{\alpha_{1,0}h}(x_0 - \alpha_{1,0} g_0), \\ s_1 &= \frac{1}{\alpha_{1,0}}(x_0 - x_1) - g_0, \\ x_2 &= \mathrm{prox}_{\alpha_{2,1}h}(x_1 - \alpha_{2,0}(g_0 + s_1) - \alpha_{2,1} g_1), \\ s_2 &= \frac{1}{\alpha_{2,1}}(x_1 - x_2 - \alpha_{2,0}(g_0 + s_1)) - g_1, \\ &\vdots \\ x_n &= \mathrm{prox}_{\alpha_{n,n-1}h}(x_{n-1} - \alpha_{n,0}(g_0 + s_1) - \cdots - \alpha_{n,n-1} g_{n-1}). \end{aligned}$$

An issue for practical implementation is efficiency: as written, this algorithm requires storing all previous (sub)gradients. Notably, this issue is not specific to our approach and already exists in the unconstrained setting. All of the algorithms in our results can be implemented efficiently; see Sections 4 and 5.

Following PEP-based analyses, we make use of *co-coercivities*, which form a complete set of inequalities for certifying convergence rates [48]. In the unconstrained setting, we only have a single set of co-coercivities $\{Q_{ij}\}$, whereas in the composite setting we have two sets of co-coercivities $\{Q^f_{ij}\}$ and $\{Q^h_{ij}\}$, respectively corresponding to the first-order information of f and h.

Definition 2.4 (Co-coercivities)**.** *Let* H *be a stepsize matrix and let* $\{x_i\}$ *be the iterates of the first-order method with* H*. Then define*

$$Q_{ij} := f_i - f_j - \langle g_j, x_i - x_j \rangle - \frac{1}{2} \|g_i - g_j\|^2 .$$

For the iterates $\{x_i\}$ of the composite extension with H, define

$$Q^f_{ij} := f_i - f_j - \langle g_j, x_i - x_j \rangle - \frac{1}{2}\|g_i - g_j\|^2,$$
$$Q^h_{ij} := h_i - h_j - \langle s_j, x_i - x_j \rangle .$$

Note here that the iterates $\{x_i\}$ in Q_{ij} and Q^f_{ij} are different, since they are generated by the algorithms in Definition 2.1 (unconstrained) and Definition 2.2 (composite), respectively.

These co-coercivities are sometimes called "valid inequalities" since they are positive for any f and h.

Lemma 2.5 ([48, Theorem 4])**.** *Let f be convex and 1-smooth, and let h be convex. Then $Q_{ij} \geqslant 0, Q^f_{ij} \geqslant 0$ and $Q^h_{ij} \geqslant 0$ for all i, j.*

2.2 Algebraic recipes for composite extension

We state our main technical result, starting with objective function as the performance metric for the convergence rates. As overviewed in Section 1.2, our result begins with a certificate (i.e., a dual PEP solution) for the unconstrained setting (2.1), which has a single square residual. From this, we establish a structured algebraic identity (2.2) which provides a candidate solution for the dual PEP in the composite setting. Once it is checked that this candidate is indeed a valid solution via its explicit formulae (see the remark below), we directly obtain a formal proof for the convergence rate.

Theorem 2.6 (Composite extension for objective function minimization)**.** *For a first-order method with H and its corresponding cocoercivities $\{Q_{ij}\}$, assume that there exist $\lambda = \{\lambda_{ij} \geqslant 0 : i \in \{0, 1, \ldots, n, *\}, j \in \{0, 1, \ldots, n\}\}$ and $\gamma = [\gamma_0, \gamma_1, \ldots, \gamma_n]$ such that*

$$\sum_{i,j} \lambda_{ij} Q_{ij} + \frac{1}{2}\|x_0 - x_* - \sum_{i=0}^{n} \gamma_i g_i\|^2 = R_n(f_* - f_n) + \frac{1}{2}\|x_0 - x_*\|^2 . \tag{2.1}$$

Then for the composite extension with H and its corresponding cocoercivities $\{Q^f_{ij}\}$ and $\{Q^h_{ij}\}$, there exist $\mu = \{\mu_{ij} : i \in \{1, \ldots, n, *\}, j \in \{1, \ldots, n\}\}, \sigma, S$ (explicitly stated in Definition 3.3) such that

$$\sum_{i,j} \lambda_{ij} Q^f_{ij} + \sum_{i,j} \mu_{ij} Q^h_{ij} + \frac{1}{2}\|x_0 - x_* - u\|^2 + \frac{1}{2}\operatorname{Tr}(VSV^T) = R_n(F_* - F_n) + \frac{1}{2}(1 + \xi)\|x_0 - x_*\|^2 , \tag{2.2}$$

where $V := \left[x_0 - x_* | s_1 | \ldots | s_n | s_*\right]$ is the columnwise-concatenated matrix, $u := \sum_{i=0}^{n} \gamma_i (g_i + s_{i+1}) - \gamma_n s_{n+1} + \sum_{i \in \{1, \ldots, n, *\}} \sigma_i s_i$, and $S := \begin{bmatrix} \xi & v^T \\ v & L \end{bmatrix}$ with $\sum v = 0$ and L being Laplacian.[5] *In particular, if*

(i) $\mu_{ij} \geqslant 0$ for all i, j

(ii) S is positive semidefinite (e.g., implied if $\xi = v^T L^\dagger v$)

then the composite extension with H has the following convergence guarantee:

$$F_n - F_* \leqslant \frac{1 + \xi}{2R_n}\|x_0 - x_*\|^2 .$$

The main technical challenge in establishing this theorem is the formulae for the multipliers and sum-of-squares term. For ease of exposition, these formulae are provided later in Definition 3.3, since this requires additional notation. Given these expressions, the proof is based on matching the coefficients (for linear forms of $\{f_i\}, \{h_i\}$ and quadratic forms of $x_0 - x_*, \{g_i\}, \{s_i\}$) in (2.1) and (2.2). See Appendix A for details.

[5] A symmetric matrix is Laplacian if all nondiagonal entries are nonpositive and all row (column) sums are 0. Laplacian matrices are positive semidefinite, since $x^T L x = \sum_{i<j} (-L_{ij})(x_i - x_j)^2 \geqslant 0$.

Remark 2.7 (Interpretation and instantiation)**.** *As overviewed in Section 1.2, Theorem 2.6 implies that only two statements—items (i) and (ii)—need to be checked in order to obtain a convergence guarantee for the composite extension. This verification is tractable because μ, S are given in closed form.*

The expression for ξ in Theorem 2.6, while providing the tightest possible rate using our framework, requires computing the pseudoinverse of L. In our applications, we bypass this by instead choosing $\xi > 0$ as a small constant which only inflates the rate slightly.

For checking (i), we can often use structural properties of λ and H from (2.1). For checking (ii), the key point is that "most of S" is already positive semidefinite; note that S is of dimension $(n+2) \times (n+2)$, and L is a positive semidefinite submatrix of dimension $(n+1) \times (n+1)$. In this sense, it suffices to show that the Schur complement $L - \frac{1}{\xi} vv^T$ (i.e., a small perturbation of L) is positive semidefinite. A tractable approach for this is to show that it is Laplacian, as we already know that the row and column sums are 0 from $\sum v = 0$. See the applications in Section 4 for details and concrete examples of checking (i) and (ii).

We also show an analogous result when the performance metric is the (sub)gradient norm. Here the algebraic structure is simpler, as we begin with the sum-of-squares term being 0 in the unconstrained setting.

Theorem 2.8 (Composite extension for gradient norm minimization)**.** *For a first-order method with H and its corresponding cocoercivities $\{Q_{ij}\}$, assume that there exists $\lambda' = \{\lambda'_{ij} \geqslant 0 : i, j \in \{0, 1, \dots, n\}\}$ such that*

$$\sum_{i,j} \lambda'_{ij} Q_{ij} = -\frac{R'_n}{2} \|g_n\|^2 + f_0 - f_n \,. \tag{2.3}$$

Then for the composite extension with H and its corresponding cocoercivities $\{Q^f_{ij}\}$ and $\{Q^h_{ij}\}$, there exist $\mu' = \{\mu'_{ij} : i \in \{0, \dots, n\}, j \in \{1, \dots, n\}\}, S'$ (explicitly stated in Definition 3.4) such that

$$\sum_{i,j} \lambda'_{ij} Q^f_{ij} + \sum_{i,j} \mu'_{ij} Q^h_{ij} + \frac{1}{2} \operatorname{Tr}(V' S' (V')^T) = -\frac{R'_n(1-\xi')}{2} \|g_n + s_n\|^2 + F_0 - F_n \,, \tag{2.4}$$

where $V' := \big[g_n | s_1 | \dots | s_n\big]$. Furthermore, if

(i) $\mu'_{ij} \geqslant 0$ for all i, j

(ii) S' is positive semidefinite (e.g., implied if $\xi' = 1 - \frac{\lambda'_{n-1,n} + \lambda'_{n,n-1}}{R'_n}$)

then the composite extension with H has the following convergence guarantee:

$$\|g_n + s_n\|^2 \leqslant \frac{2}{R'_n(1-\xi')} (F_0 - F_n) \,.$$

The proof is similar to that of Theorem 2.6 and is provided in Appendix B. Note that whereas the positive semidefinite matrix S in Theorem 2.6 has Laplacian structure, the matrix S' in Theorem 2.8 has diagonally dominant structure.[6]

Remark 2.9 (Performance metric for gradient norm minimization)**.** *In the literature on gradient norm minimization, results are often stated with respect to $f_0 - f_*$ instead of $f_0 - f_n$ (or in the composite setting, $F_0 - F_*$ rather than $F_0 - F_n$). It is easy to see that bounds with respect to the latter imply corresponding bounds with respect to the former. Furthermore, often the converse is also true, meaning that these performance metrics are essentially equivalent.*[7] *We state our results with respect to $f_0 - f_n$ and $F_0 - F_n$ because this yields slightly simpler formulations and also applies to settings where no finite minimizer exists.*

[6] A symmetric matrix is diagonally dominant if for each row/column, the absolute value of the diagonal entry is at least the sum of the absolute values of the nondiagonal entries. For example, Laplacian matrices are diagonally dominant. Diagonally dominant matrices are positive semidefinite by the Gershgorin circle theorem.

[7] Details: combining (2.3) with the inequality $Q_{n*} = f_n - f_* - \frac{1}{2}\|g_n\|^2 \geqslant 0$ yields $-\frac{R'_n+1}{2}\|g_n\|^2 + f_0 - f_* \geqslant 0$; combining (2.4) with the inequality $Q^f_{n*} + Q^h_{n*} = F_n - F_* - \frac{1}{2}\|g_n + s_*\|^2 \geqslant 0$ yields $-\frac{R'_n(1-\xi')}{2}\|g_n + s_n\|^2 + F_0 - F_* \geqslant 0$. In many cases, the bounds with respect to $f_0 - f_*$ (or $F_0 - F_*$) are obtained precisely in this way. See also, for example, [34, equation 3].

Remark 2.10 (Corollary for related performance metric)**.** *A standard trick is that by running methods for minimizing objective function and minimizing gradient norm, each for $n/2$ iterations, one obtains a final convergence rate for n steps—comparing gradient norm to* initial distance *(both squared)—which is the product of the two constituent rates, see e.g., [37]. For example, by using the composite extensions of OGM and OGM-G,*

$$\begin{aligned}\|g_n + s_n\|^2 &\leqslant O\left(\frac{1}{(n/2)^2}\right)(f(x_{n/2}) + h(x_{n/2}) - f(x_*) - h(x_*))\\ &\leqslant O\left(\frac{1}{(n/2)^2}\right) \times O\left(\frac{1}{(n/2)^2}\right)\|x_0 - x_*\|^2 = O\left(\frac{1}{n^4}\right)\|x_0 - x_*\|^2 .\end{aligned}$$

Similarly, for stepsize-accelerated proximal GD, our results yield a rate of $O(1/n^{(\log_2(1+\sqrt{2}))^2}) \approx O(1/n^{1.6168})$ for the final gradient norm compared to the initial distance.

2.3 Notation

To express convergence rates in a simpler form, we often use the standard notation $a_n \sim b_n$ if $\lim_n(a_n/b_n) = 1$. We write $[A]_{i,j}$ to denote the (i, j) entry of the matrix A.

It is also convenient to introduce some notation for triangular matrices corresponding to stepsize parameters. We define $U(a_1, \ldots, a_n)$ to be the $n \times n$ upper triangular matrix whose (i, j) entry is a_i if $i = j$ and 1 if $i < j$. We denote $U(\mathbf{1}_n)$ as U_n; this is the $n \times n$ upper triangular matrix with 1 in every entry on or above the diagonal. For a n-stepsize matrix H, we define $\widetilde{H} := HU_n$, i.e.,

$$\widetilde{H} = \begin{bmatrix} \widetilde{\alpha}_{1,0} & \cdots & \widetilde{\alpha}_{n,0} \\ & \ddots & \vdots \\ & & \widetilde{\alpha}_{n,n-1} \end{bmatrix},$$

where $\widetilde{\alpha}_{i,j} := \sum_{k=j+1}^{i} \alpha_{k,j}$. Then it can be observed that the iterates $x_0, \ldots, x_n$ of the first-order method with H satisfy

$$\begin{aligned} x_1 &= x_0 - \widetilde{\alpha}_{1,0} g_0, \\ x_2 &= x_0 - \widetilde{\alpha}_{2,0} g_0 - \widetilde{\alpha}_{2,1} g_1, \\ &\vdots \\ x_n &= x_0 - \widetilde{\alpha}_{n,0} g_0 - \cdots - \widetilde{\alpha}_{n,n-1} g_{n-1}. \end{aligned}$$

3 Formulae for multipliers

In this section, we formally define the multipliers in Theorems 2.6 and 2.8. We begin by introducing notation for certain simple linear transformations of the coefficients λ and μ that frequently appear in our analysis. These definitions are stated for λ and μ, and apply analogously to λ' and μ'.

Definition 3.1 (Multipliers)**.** *For $\lambda = \{\lambda_{i,j} : i \in \{0, 1, \ldots, n, *\}, j \in \{0, 1, \ldots, n\}\}$ and $\mu = \{\mu_{i,j} : i \in \{1, \ldots, n, *\}, j \in \{1, \ldots, n\}\}$, define*

$$\lambda_{i,\bullet} := \sum_j \lambda_{i,j}, \quad \lambda_{\bullet,j} := \sum_i \lambda_{i,j}, \quad \mu_{\bullet,j} := \sum_i \mu_{i,j} ,$$

and

$$\widehat{\lambda} := \begin{bmatrix} -(\lambda_{\bullet,0} + \lambda_{0,\bullet}) & \dots & \lambda_{0,n-2} + \lambda_{n-2,0} & \lambda_{0,n-1} + \lambda_{n-1,0} \\ \vdots & \ddots & \vdots & \vdots \\ \lambda_{n-2,0} + \lambda_{0,n-2} & \dots & -(\lambda_{\bullet,n-2} + \lambda_{n-2,\bullet}) & \lambda_{n-2,n-1} + \lambda_{n-1,n-2} \\ \lambda_{n-1,0} + \lambda_{0,n-1} & \dots & \lambda_{n-1,n-2} + \lambda_{n-2,n-1} & -(\lambda_{\bullet,n-1} + \lambda_{n-1,\bullet}) \end{bmatrix} \in \mathbb{R}^{n \times n},$$

$$\widetilde{\lambda} := \begin{bmatrix} \lambda_{1,0} & -\lambda_{\bullet,1} & \lambda_{1,2} & \dots & \lambda_{1,n-1} \\ \lambda_{2,0} & \lambda_{2,1} & -\lambda_{\bullet,2} & \dots & \lambda_{2,n-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \lambda_{n-1,0} & \lambda_{n-1,1} & \dots & \lambda_{n-1,n-2} & -\lambda_{\bullet,n-1} \\ \lambda_{n,0} & \lambda_{n,1} & \dots & \lambda_{n,n-2} & \lambda_{n,n-1} \end{bmatrix} \in \mathbb{R}^{n \times n},$$

$$\widetilde{\mu} := \begin{bmatrix} -\mu_{\bullet,1} & \dots & \mu_{1,n} \\ \vdots & \ddots & \vdots \\ \mu_{n,1} & \dots & -\mu_{\bullet,n} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Similar notations are also defined with respect to $\lambda' = \{\lambda'_{i,j} : i, j \in \{0, 1, \dots, n\}\}$ and $\mu' = \{\mu'_{i,j} : i \in \{0, \dots, n\}, j \in \{1, \dots, n\}\}$.

Remark 3.2 (Reparameterization)**.** *Below, it is convenient to define $\widetilde{\mu}$ and $\widetilde{\mu}'$ respectively instead of μ and μ' (which are the actual multipliers in Theorems 2.6 and 2.8). This is equivalent due to the one-to-one correspondence: given $\widetilde{\mu}$, $\mu_{i,j} = e_i^T \widetilde{\mu} e_j \cdot \mathbf{1}_{\{i \neq j\}}$ uniquely defines μ; similarly, given $\widetilde{\mu}'$, $\mu'_{i,j} = e_i^T \widetilde{\mu}' e_j \cdot \mathbf{1}_{\{i \neq j\}}$ uniquely defines μ', where we use the shorthand $e_0 := -\mathbf{1}_n$ and $e_* := -\mathbf{1}_n$.*

3.1 Objective function minimization

In the setting of objective function minimization (Theorem 2.6), the new coefficients σ, μ, S for the composite setting are formally defined as follows.

Definition 3.3. *In the setting of Theorem 2.6, define σ as*

$$\sigma := [\sigma_1, \dots, \sigma_n, \sigma_*], \quad \sigma_i := \frac{\lambda_{i-1,n} + \lambda_{n,i-1}}{\gamma_n}. \tag{3.1}$$

For notational convenience, we also define $\widetilde{\gamma} := [\gamma_0, \dots, \gamma_{n-1}], \widetilde{\sigma} := [\sigma_1, \dots, \sigma_n]$. Then μ, S are defined as

$$\widetilde{\mu} := -\widetilde{H}^{-1}((\widetilde{H}\widetilde{\lambda})^T + \widetilde{\gamma}(\widetilde{\gamma} + \widetilde{\sigma})^T) = \widetilde{H}^{-1}(\widehat{\lambda} - \widetilde{\gamma}\widetilde{\sigma}^T) + \widetilde{\lambda}, \tag{3.2}$$

$$S := \begin{bmatrix} \xi & v^T \\ v & L \end{bmatrix}, \quad v := [v_1, \dots, v_n, v_*], \quad v_i := \sigma_i + \lambda_{*,i-1} - \mu_{*,i}, \quad L := -\begin{bmatrix} \widehat{\lambda} & \widetilde{\gamma} \\ \widetilde{\gamma}^T & -\lambda_{*,\bullet} \end{bmatrix} - \sigma\sigma^T. \tag{3.3}$$

Note that the equality in (3.2) is due to (2.1); for details, see (A.2).

3.2 Gradient norm minimization

In the setting of gradient norm minimization (Theorem 2.8), the new coefficients μ', S' for the composite setting are formally defined as follows.

Definition 3.4. *In the setting of Theorem 2.8, define μ', S' as*

$$\widetilde{\mu}' := -\widetilde{H}^{-1}(\widetilde{H}\widetilde{\lambda}')^T, \tag{3.4}$$

$$S' := \begin{bmatrix} R'_n & (v')^T \\ v' & -\widehat{\lambda}' \end{bmatrix} - R'_n(1 - \xi')(e_1 + e_{n+1})(e_1 + e_{n+1})^T, \quad v' := [v'_1, \dots, v'_n], \quad v'_i := \lambda'_{i-1,n} + \lambda'_{n,i-1}. \tag{3.5}$$

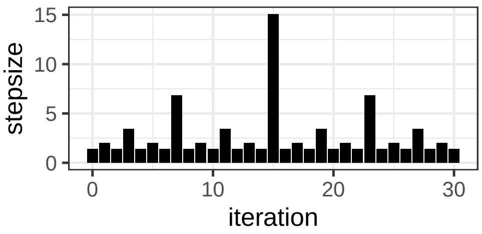

FIGURE 1: First 31 stepsizes of the silver stepsize schedule. The stepsizes are time-varying and fractal-like.

4 Applications to objective function minimization

Here we apply our main theorem for objective function minimization (Theorem 2.6) to the composite extensions of stepsize-accelerated gradient descent and optimized gradient method. For each result, we only need to verify items (i) and (ii) of the theorem. The following subsections do this by using the formulae for the multipliers in Definition 3.3.

4.1 Stepsize-accelerated proximal GD

Gradient descent (GD) is a first-order method with diagonal stepsize matrix H. Its composite extension is proximal GD with the same stepsizes.

For $k \in \mathbb{N}$ and $n = 2^k - 1$, the silver stepsize schedule $\pi^{(k)}$ (of length n) is defined as

$$\pi_i^{(k)} := 1 + \rho^{\nu(i)-1} \text{ for } 1 \leqslant i \leqslant n\,,$$

where $\rho := 1 + \sqrt{2}$ is the silver ratio and $\nu(i)$ is the largest integer j such that 2^j divides i. This can be equivalently defined in a recursive way as $\pi^{(1)} := [\sqrt{2}]$ and $\pi^{(k+1)} := [\pi^{(k)}, \rho^{k-1}+1, \pi^{(k)}]$. The silver stepsize schedule deviates qualitatively from mainstream stepsize schedules: it is time-varying, nonmonotone, fractal-like, and uses arbitrarily large stepsizes that are in particular larger than 2 (the threshold at which constant stepsize schedules make GD divergent). See Figure 1 for an illustration.

[2] introduced the silver stepsizes and proved the (partially) accelerated rate of $O(1/n^{\log_2 \rho})$. The following choice of λ and γ from [51, Theorem 5.2] certifies this asymptotic rate with tight constant factor. The recursive definition of λ, in essence, combines the proofs for the constitutent stepsize schedules in the recursive definition of π—a technique known as *recursive gluing* [2, 3].

Proposition 4.1 (Multipliers for unconstrained GD; [51, Theorem 5.2])**.** *For $k \in \mathbb{N}$ and $n = 2^k - 1$, GD with the silver stepsizes satisfies* (2.1) *with*

$$\begin{aligned} \lambda &= \lambda^{(k)}, \\ \gamma &= [\pi^{(k)}, \rho^k], \\ R_n &= 2\rho^k - 1. \end{aligned}$$

Here, $\lambda^{(k)}$ is recursively defined as

$$\lambda_{i,j}^{(k)} := \bar{\lambda}_{i,j}^{(k)} \mathbf{1}_{\{i \neq *\}} + \underline{\lambda}_j^{(k)} \mathbf{1}_{\{i = *\}}\,,$$

with $\underline{\lambda}^{(k)} := [\pi^{(k)}, \rho^k]$ and

$$\begin{aligned} \bar{\lambda}^{(1)} &:= \begin{array}{ccc} & 0 & 1 \\ 0 & 0 & \rho \\ 1 & 1 & 0 \end{array}, \\ \bar{\lambda}^{(k+1)} &:= \underbrace{\bar{\lambda}^{(k+1),\mathrm{rec}}}_{\textit{recursion}} + \underbrace{\bar{\lambda}^{(k+1),\mathrm{sp}}}_{\textit{sparse correction}} + \underbrace{\bar{\lambda}^{(k+1),\mathrm{lr}}}_{\textit{low-rank correction}}\,, \end{aligned}$$

where $\bar{\lambda}^{(k+1),\mathrm{rec}}, \bar{\lambda}^{(k+1),\mathrm{sp}}, \bar{\lambda}^{(k+1),\mathrm{lr}}$ *are defined as*

$$\begin{aligned}
\bar{\lambda}^{(k+1),\mathrm{rec}}_{i,j} &:= \bar{\lambda}^{(k)}_{i,j}\mathbf{1}_{\{0\leqslant i,j\leqslant n\}} + \rho^2\bar{\lambda}^{(k)}_{i-n-1,j-n-1}\mathbf{1}_{\{n+1\leqslant i,j\leqslant 2n+1\}}\,,\\
\bar{\lambda}^{(k+1),\mathrm{sp}}_{i,j} &:= \rho\mathbf{1}_{\{(i,j)=(n,2n+1)\}} + \rho^k\mathbf{1}_{\{(i,j)=(2n+1,n)\}}\,,\\
\bar{\lambda}^{(k+1),\mathrm{lr}}_{i,j} &:= \rho\pi^{(k)}_{j-n}\mathbf{1}_{\{i=n,n+1\leqslant j\leqslant 2n\}} + \rho\pi^{(k)}_{j-n}\mathbf{1}_{\{i=2n+1,n+1\leqslant j\leqslant 2n\}}\,.
\end{aligned}$$

Based on this, we show that proximal GD with the silver stepsizes also has an $O(1/n^{\log_2\rho})$ rate. This improves over the $O(1/n)$ classical rate of proximal GD [50, Theorem 7] and establishes that the phenomenon of stepsize-based acceleration extends to the composite setting.

Theorem 4.2 (Convergence rate of proximal GD)**.** *For $k \in \mathbb{N}$ and $n = 2^k - 1$, proximal GD with the silver stepsizes satisfies*

$$F(x_n) - F(x_*) \leqslant \frac{\rho}{\sqrt{2}(4\rho^k - 2)}\|x_0 - x_*\|^2 \sim \frac{\rho}{4\sqrt{2}n^{\log_2\rho}}\|x_0 - x_*\|^2\,.$$

Proof. The result follows from an application of Theorem 2.6 once we verify items (i) and (ii) there. For item (i), from $H = \mathrm{diag}(\widetilde{\gamma}) = \mathrm{diag}(\gamma_0, \ldots, \gamma_{n-1})$ we have

$$\widetilde{\mu} = -\widetilde{H}^{-1}\widetilde{\lambda}^T\widetilde{H}^T - \begin{bmatrix}\mathbf{0}_{(n-1)\times n}\\ (\widetilde{\gamma}+\widetilde{\sigma})^T\end{bmatrix}\,.$$

The main technical portion is to calculate the first summand. In general, the individual entry of such matrix (in the form $-\widetilde{H}^{-1}A^T\widetilde{H}^T$) can be expressed as a scaled partial sum as follows.

Lemma 4.3. *Let $H = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ where $\alpha_i \neq 0$ for all $1 \leqslant i \leqslant n$. Then*

$$[-\widetilde{H}^{-1}A^T\widetilde{H}^T]_{i,j} = \begin{cases}\alpha_j\left(\frac{\sum_{l\geqslant j}[A]_{l,i+1}}{\alpha_{i+1}} - \frac{\sum_{l\geqslant j}[A]_{l,i}}{\alpha_i}\right) & i \leqslant n-1\\ -\alpha_j\frac{\sum_{l\geqslant j}[A]_{l,n}}{\alpha_n} & i = n\,.\end{cases}$$

Proof. From $-\widetilde{H}^{-1}A^T\widetilde{H}^T = -U_n^{-1}\mathrm{diag}(1/\alpha_1, \ldots, 1/\alpha_n)A^TU_n^T\mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ and $[U_n^{-1}]_{i,j} = \begin{cases}1 & i=j\\ -1 & j=i+1\\ 0 & \text{else}\end{cases}$,

$$[-U_n^{-1}\mathrm{diag}(1/\alpha_1, \ldots, 1/\alpha_n)A^T]_{i,j} = \begin{cases}\frac{[A]_{j,i+1}}{\alpha_{i+1}} - \frac{[A]_{j,i}}{\alpha_i} & i \leqslant n-1\\ -\frac{[A]_{j,n}}{\alpha_n} & i = n\,.\end{cases}$$

Thus (post)multiplying $U_n^T\mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ to this matrix yields the result. □

In this sense, to establish nonnegativity we need to compare partial sums in adjacent columns of $\widetilde{\lambda}$. The key lemma here is that, fortunately, we can compare the individual entries and sum the differences. Below, recall that $[\gamma_0, \ldots, \gamma_{n-1}] = \pi^{(k)}$ from Proposition 4.1, and $\gamma_n = \rho^k$.

Lemma 4.4. *Let $\lambda = \lambda^{(k)}$. Then*

$$\frac{\lambda_{i,j-1}}{\gamma_{j-1}} - \frac{\lambda_{i,j}}{\gamma_j}\begin{cases}\geqslant 0 & 0 \leqslant i \leqslant j-2\\ \leqslant 0 & i \geqslant j+1\end{cases}$$

for all $1 \leqslant j \leqslant n-1$. In particular, for $j = n$ the inequality is valid for $\gamma_n = 1$.[8]

[8] For $j = n$, this immediately implies a corresponding inequality for $\gamma_n = \rho^k$. This result is only needed for the induction argument.

Proof. We use induction on k. For $k = 1$, no such entry exists; for $k = 2$, most of the comparisons are trivial from either $\lambda_{i,j-1} = 0$ or $\lambda_{i,j} = 0$, and the nontrivial ones are: $\frac{\lambda_{3,1}}{\gamma_1} - \frac{\lambda_{3,2}}{\gamma_2} = \frac{\rho}{2} - \frac{\rho^2+\rho+1}{\sqrt{2}} \leqslant 0$, $\frac{\lambda_{1,2}}{\gamma_2} - \frac{\lambda_{1,3}}{\gamma_3} = \frac{\rho+1}{\sqrt{2}} - \rho = 0$ (letting $\gamma_3 = 1$).

Now assume that the result holds for k, and consider $k+1$ with $\lambda = \lambda^{(k+1)}$ and corresponding γ. Then for $1 \leqslant j \leqslant n$, the inequalities hold from the recursive definition of λ and induction hypothesis, with $\frac{\lambda_{2n+1,n-1}}{\gamma_{n-1}} - \frac{\lambda_{2n+1,n}}{\gamma_n} = 0 - \frac{\lambda^{(k+1),\mathrm{sp}}_{2n+1,n}}{\gamma_n} \leqslant 0$. For $j = n+1$, note that $\lambda_{i,n} = 0$ for all $n+1 \leqslant i \leqslant 2n$ and $\lambda_{i,n+1} = 0$ for all $0 \leqslant i \leqslant n-1$; thus it only suffices to consider $i = 2n+1$, where $\frac{\lambda_{2n+1,n}}{\gamma_n} - \frac{\lambda_{2n+1,n+1}}{\gamma_{n+1}} = \frac{\rho^k}{\rho^{k-1}+1} - \frac{\sqrt{2}\rho}{\sqrt{2}} \leqslant 0$. For $n+2 \leqslant j \leqslant 2n+1$ the inequality readily follows from the induction hypothesis, with $\frac{\lambda_{n,2n}}{\gamma_{2n}} - \frac{\lambda_{n,2n+1}}{\gamma_{2n+1}} = \frac{\rho^2}{\rho} - \rho = 0$ (letting $\gamma_{2n+1} = 1$). □

Establishing these, now we return to item (i), i.e., the nonnegativity of μ. First we consider the first $n-1$ rows of $\widetilde{\mu}$. For $2 \leqslant i \leqslant n-1$, we have

$$\mu_{i,1} = \gamma_0 \left(-\frac{\lambda_{*,i}}{\gamma_i} + \frac{\lambda_{*,i-1}}{\gamma_{i-1}} \right) = 0\,,$$

since $\lambda_{*,j} = \gamma_j$ for all $0 \leqslant j \leqslant n-1$. Thus for $1 \leqslant j < i \leqslant n-1$, by Lemma 4.4,

$$\mu_{i,j} = \gamma_{j-1} \left(-\frac{\sum_{l=1}^{j-1} \lambda_{l,i}}{\gamma_i} + \frac{\sum_{l=1}^{j-1} \lambda_{l,i-1}}{\gamma_{i-1}} \right) \geqslant 0\,,$$

and similarly for $1 \leqslant i \leqslant n-1$ and $j > i$,

$$\mu_{i,j} = \gamma_{j-1} \left(\frac{\sum_{l=j}^{n} \lambda_{j,i}}{\gamma_i} - \frac{\sum_{l=j}^{n} \lambda_{j,i-1}}{\gamma_{i-1}} \right) \geqslant 0\,.$$

For the nth row of $\widetilde{\mu}$, for $1 \leqslant j \leqslant n-1$,

$$\begin{aligned}
\mu_{n,j} &= \frac{\gamma_{j-1}}{\gamma_{n-1}} (\lambda_{*,n-1} + \sum_{l=1}^{j-1} \lambda_{l,n-1}) - (\gamma_{j-1} + \sigma_j) \\
&= \frac{\gamma_{j-1}}{\gamma_{n-1}} \sum_{l=1}^{j-1} \lambda_{l,n-1} - \frac{\lambda_{j-1,n} + \lambda_{n,j-1}}{\gamma_n} \\
&\geqslant \frac{1}{\gamma_n} (\gamma_{j-1} \sum_{l=1}^{j-1} \lambda_{l,n} - \lambda_{j-1,n} - \lambda_{n,j-1}) \geqslant 0\,,
\end{aligned}$$

where the second equality is from $\lambda_{*,n-1} = \gamma_{n-1}$ and the definition of σ, and the final inequality is from the following lemma.

Lemma 4.5. *For $k \geqslant 2$ and $n = 2^k - 1$, $t_j^{(k)} := \pi_j^{(k)} \sum_{l=1}^{j-1} \lambda_{l,n}^{(k)} - \lambda_{j-1,n}^{(k)} - \lambda_{n,j-1}^{(k)} \geqslant 0$ for all $1 \leqslant j \leqslant n-1$.*

Proof. For $k = 2$, $t_1^{(2)} = 0$ and $t_2^{(2)} = 2\lambda_{1,3}^{(2)} - \lambda_{1,3}^{(2)} - \lambda_{3,1}^{(2)} = \rho - \rho = 0$.

Assume that the result holds for k, and consider $2n+1 = 2^{k+1} - 1$. Then clearly $t_j^{(k+1)} = 0$ for all $1 \leqslant j \leqslant n$, and $t_{n+1}^{(k+1)} = (1 + \rho^{k-1})\rho - \rho - \rho^k = 0$. For $n+2 \leqslant j \leqslant 2n$,

$$\begin{aligned}
t_j^{(k+1)} &= \pi_j^{(k+1)} \sum_{l=1}^{j-1} \lambda_{l,2n+1}^{(k+1)} - \lambda_{j-1,2n+1}^{(k+1)} - \lambda_{2n+1,j-1}^{(k+1)} \\
&= \pi_{j-n-1}^{(k)} \left(\rho + \sum_{l=1}^{(j-n-1)-1} \rho^2 \lambda_{l,n}^{(k)} \right) - \rho^2 \lambda_{(j-n-1)-1,2n+1}^{(k)} - (\rho^2 \lambda_{n,(j-n-1)-1}^{(k)} + \rho \pi_{j-n-1}^{(k)}) \\
&= \rho^2 t_{j-n-1}^{(k)} \geqslant 0\,.
\end{aligned}$$

□

Furthermore, from $\lambda_{1,0} = 1$ and $\lambda_{2,0} = \cdots = \lambda_{n,0} = 0$, $\begin{bmatrix}\mu_{*,1} & \dots & \mu_{*,n}\end{bmatrix} = -\mathbf{1}_n^T\widetilde{\mu} = e_1 + \widetilde{\gamma} + \widetilde{\sigma}$ which proves item (i) with $v_i = \begin{cases} -1 & i = 1 \\ 1 & i = * \\ 0 & \text{else} \end{cases}.$

For item (ii), it suffices to choose $\xi > 0$ such that $L - \frac{1}{\xi}vv^T$ is Laplacian. Since $v = -e_1 + e_{n+1}$, we only need to check that the $(1, n+1)$ entry of this matrix $-\sqrt{2} - \sigma_1\sigma_* + \frac{1}{\xi}$ is nonpositive. For this we can take $\xi = \frac{1}{\sqrt{2}}$, (note that for $k \geqslant 2$, $\sigma_1 = 0$ and thus the entry becomes 0), proving item (ii).

□

4.2 Proximal OGM

Optimized gradient method (OGM) [31] is a (n-step) first-order method with stepsize matrix H, whose entry is recursively defined as

$$\alpha_{i+1,j} = \begin{cases} \frac{\theta_i - 1}{\theta_{i+1}}\alpha_{i,j} & 0 \leqslant j \leqslant i-2 \\ \frac{\theta_i - 1}{\theta_{i+1}}(\alpha_{i,i-1} - 1) & j = i-1 \\ 1 + \frac{2\theta_i - 1}{\theta_{i+1}} & j = i\,. \end{cases} \tag{4.1}$$

Here, for each $n \in \mathbb{N}$ the sequence $\{\theta_i : 0 \leqslant i \leqslant n\}$ is defined as

$$\theta_i = \begin{cases} 1 & i = 0 \\ \frac{1+\sqrt{1+4\theta_{i-1}^2}}{2} & 1 \leqslant i \leqslant n-1 \\ \frac{1+\sqrt{1+8\theta_{i-1}^2}}{2} & i = n\,. \end{cases} \tag{4.2}$$

It is straightforward to check that $\theta_{i+1}^2 - \theta_{i+1} - \theta_i^2 = 0$ for $0 \leqslant i \leqslant n-2$, $\theta_n^2 - \theta_n - 2\theta_{n-1}^2 = 0$ and $\theta_n^2 \sim n^2/2$ (see e.g., [40]). The certificate for the convergence rate of OGM is rather simple, as presented in the following proposition.

Proposition 4.6 (Multipliers for OGM; [31, Theorem 2])**.** *OGM satisfies* (2.1) *with*

$$\begin{aligned} \lambda_{i,j} &= \begin{cases} 2\theta_i^2 & j = i+1, 0 \leqslant i \leqslant n-1 \\ 2\theta_j & i = *, 0 \leqslant j \leqslant n-1 \\ \theta_j & i = *, j = n \\ 0 & \textit{else}\,, \end{cases} \\ \gamma &= [2\theta_0, \ldots, 2\theta_{n-1}, \theta_n]\,, \\ R_n &= \theta_n^2\,. \end{aligned}$$

We establish a comparable rate for its composite extension as follows. Notably, this convergence rate is faster than that of FISTA and only differs by a small constant factor from the exactly optimal rate.

Theorem 4.7 (Convergence rate of proximal OGM)**.** *For $n \geqslant 2$, the composite extension of OGM satisfies*[9]

$$F(x_n) - F(x_*) \leqslant \frac{3+\sqrt{5}}{8\theta_n^2}\|x_0 - x_*\|^2 \sim \frac{3+\sqrt{5}}{4n^2}\|x_0 - x_*\|^2\,.$$

Proof. As before, we show items (i) and (ii) of Theorem 2.6. For item (i), first we use the following formulation of the stepsize matrix for OGM.

Lemma 4.8 (Factorization of H for OGM)**.** *Let H be the n-stepsize matrix for OGM. Then*

$$H = \operatorname{diag}(2\theta_0, \ldots, 2\theta_{n-1})U(\varphi_1, \ldots, \varphi_n)U(\theta_1, \ldots, \theta_n)^{-1}\,,$$

where $\varphi_i := \begin{cases} 1 + \frac{\theta_{i-1}}{2\theta_i} & i < n \\ 1 + \frac{\theta_{n-1}}{\theta_n} & i = n\,. \end{cases}$

[9] For $n = 1$, $F(x_n) - F(x_*) \leqslant \frac{2}{3\theta_n^2}\|x_0 - x_*\|^2 = \frac{1}{6}\|x_0 - x_*\|^2$ holds, which is tight. For $n \geqslant 3$, the constant factor in the convergence rate in Theorem 4.7 can be slightly improved; see the proof and Footnote 10 for details.

Proof. Consider an equivalent equality (note that each matrix is upper triangular)

$$HU(\theta_1,\dots,\theta_n)=\mathrm{diag}(2\theta_0,\dots,2\theta_{n-1})U(\varphi_1,\dots,\varphi_n)\,. \tag{4.3}$$

From [31, Lemma 4], we have $\sum_{j=k+1}^{i}\alpha_{j,k}+\theta_{i+1}\alpha_{i+1,k}=2\theta_k$ for all $0\leqslant k<i\leqslant n-1$, which is precisely the equality corresponding to the $(k+1,i+1)$ entry of (4.3). Similarly, from the same lemma we have $\theta_{i+1}\alpha_{i+1,i}=\theta_{i+1}+2\theta_i-1$ for all $0\leqslant i\leqslant n-1$, which is precisely the equality corresponding to the $(i+1,i+1)$ entry of (4.3), from $2\theta_i\varphi_{i+1}=\theta_{i+1}+2\theta_i-1$ which follows from the recursive definition of $\{\theta_i\}$. □

Now we calculate the term $(\widetilde{H}\widetilde{\lambda})^T$ in (3.2). From

$$[\widetilde{\lambda}]_{i,j}=\begin{cases}-2\theta_i^2 & j=i+1\\ 2\theta_i^2 & j=i+2\\ 0 & \text{else}\,,\end{cases}\qquad\text{thus}\qquad [U_n\widetilde{\lambda}]_{i,j}=\begin{cases}-2\theta_i^2 & j=i+1\\ -2\theta_{j-1} & j\geqslant i+2\\ 0 & \text{else}\,,\end{cases}$$

it can be observed that

$$U_n\widetilde{\lambda}=U(\theta_1,\dots,\theta_n)\begin{bmatrix}\mathbf{0}_{n-1} & \mathrm{diag}(-2\theta_1,\dots,-2\theta_{n-1})\\ 0 & \mathbf{0}_{n-1}^T\end{bmatrix}\,,$$

which implies

$$\begin{aligned}\widetilde{H}\widetilde{\lambda}&=\mathrm{diag}(2\theta_0,\dots,2\theta_{n-1})U(\varphi_1,\dots,\varphi_n)\begin{bmatrix}\mathbf{0}_{n-1} & \mathrm{diag}(-2\theta_1,\dots,-2\theta_{n-1})\\ 0 & \mathbf{0}_{n-1}^T\end{bmatrix}\\ &=\mathrm{diag}(2\theta_0,\dots,2\theta_{n-1})\left(\begin{bmatrix}\mathbf{0}_{n-1} & \mathrm{diag}(-\theta_0,\dots,-\theta_{n-2})\\ 0 & \mathbf{0}_{n-1}^T\end{bmatrix}+\begin{bmatrix} & -2\theta_1 & -2\theta_2 & \cdots & -2\theta_{n-1}\\ & & -2\theta_2 & \cdots & -2\theta_{n-1}\\ \mathbf{0}_{n-1} & & & \ddots & \vdots\\ & & & & -2\theta_{n-1}\\ 0 & & \mathbf{0}_{n-1}^T & & \end{bmatrix}\right)\,.\end{aligned}$$

Also, from (3.1) we obtain $\sigma=[\mathbf{0}_{n-1},\theta_n-1,1]$, which implies that

$$[(\widetilde{H}\widetilde{\lambda})^T+\widetilde{\gamma}(\widetilde{\gamma}+\widetilde{\sigma})^T]_{i,j}=\begin{cases}4\theta_{i-1}\theta_{j-1} & i\leqslant j\leqslant n-1\\ -2\theta_{j-1}^2 & i=j+1, j\leqslant n-1\\ 2(2\theta_{n-1}+\theta_n-1)\theta_{i-1} & j=n\\ 0 & \text{else}\,.\end{cases}$$

Thus

$$\begin{aligned}\widetilde{\mu}&=-\widetilde{H}^{-1}((\widetilde{H}\widetilde{\lambda})^T+\widetilde{\gamma}(\widetilde{\gamma}+\widetilde{\sigma})^T)\\ &=-U_n^{-1}U(\theta_1,\dots,\theta_n)U(\varphi_1,\dots,\varphi_n)^{-1}\mathrm{diag}(2\theta_0,\dots,2\theta_{n-1})^{-1}((\widetilde{H}\widetilde{\lambda})^T+\widetilde{\gamma}(\widetilde{\gamma}+\widetilde{\sigma})^T)\,,\end{aligned}$$

where

$$[\mathrm{diag}(2\theta_0,\dots,2\theta_{n-1})^{-1}((\widetilde{H}\widetilde{\lambda})^T+\widetilde{\gamma}(\widetilde{\gamma}+\widetilde{\sigma})^T)]_{i,j}=\begin{cases}2\theta_{j-1} & i\leqslant j\leqslant n-1\\ -(\theta_j-1) & i=j+1, j\leqslant n-1\\ 2\theta_{n-1}+\theta_n-1 & j=n\\ 0 & \text{else}\,.\end{cases}$$

Now we consider each column of $\widetilde{\mu}$, which boils down to solving the following equation. The proof is straightforward by solving the equation in the order of $x_n,x_{n-1},\dots,x_1$.

Lemma 4.9 (Solving linear system for jth column)**.** *For $1 \leqslant j \leqslant n$, define $x(j)$ to be the unique solution $x = [x_1, \dots, x_n]$ of*

$$U(\varphi_1, \dots, \varphi_n)x = 2\theta_{j-1} \begin{bmatrix} \mathbf{1}_j \\ \mathbf{0}_{n-j} \end{bmatrix} - (\theta_j - 1)e_{j+1}\,,$$

where for here, $e_{n+1} := \mathbf{0}_n$. Then

$$\begin{aligned} x_n &= \dots = x_{j+2} = 0, \\ x_{j+1} &= -\frac{\theta_j - 1}{\varphi_{j+1}} < 0, \\ x_j &= \frac{1}{\varphi_j}(2\theta_{j-1} - x_{j+1}) > 0, \\ x_i &= \frac{\varphi_{i+1} - 1}{\varphi_i} x_{i+1} = \frac{\theta_{i+1} - 1}{\theta_{i-1} + 2\theta_i} x_{i+1} > 0, 1 \leqslant i \leqslant j-1. \end{aligned}$$

Now fix $1 \leqslant j \leqslant n-1$. Then for $x = x(j)$, by Lemma 4.9 we have $x_{j+1} < 0$ and $x_i > 0$ for all $1 \leqslant i \leqslant j$. Thus with $[U_n^{-1}U(\theta_1, \dots, \theta_n)]_{i,j} = \begin{cases} \theta_i & i = j \\ 1 - \theta_j & j = i+1 \\ 0 & \text{else}\,, \end{cases}$ we have

$$\begin{bmatrix} \mu_{1,j} \\ \vdots \\ \mu_{j-1,j} \\ -\mu_{\bullet,j} \\ \mu_{j+1,j} \\ \vdots \\ \mu_{n,j} \end{bmatrix} = - \begin{bmatrix} \theta_1 & 1-\theta_2 & & & \\ & \theta_2 & 1-\theta_3 & & \\ & & \ddots & \ddots & \\ & & & \theta_{n-1} & 1-\theta_n \\ & & & & \theta_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

or equivalently,

$$\begin{aligned} \mu_{n,j} &= \dots = \mu_{j+2,j} = 0, \\ \mu_{j+1,j} &= -\theta_{j+1}x_{j+1} > 0, \\ \mu_{i,j} &= -(\theta_i x_i + (1-\theta_{i+1})x_{i+1}) = -\theta_i x_i + (\theta_{i-1} + 2\theta_i)x_i = (\theta_{i-1} + \theta_i)x_i > 0, 1 \leqslant i \leqslant j-1, \end{aligned}$$

and

$$\mu_{*,j} = \mu_{\bullet,j} - \sum_{i \notin \{j,*\}} \mu_{i,j} = \theta_1 x_1 + x_2 + \dots + x_n = (\theta_1 - \varphi_1)x_1 + 2\theta_{j-1} = (\theta_1 - \varphi_1)x_1 + \sigma_j + \lambda_{*,j-1}\,,$$

which proves that $\mu_{*,j} > 0$ and $v_j < 0$. It can be analogously shown that $\mu_{1,n}, \dots, \mu_{n-1,n} > 0$ and $v_n = \sigma_n + \lambda_{*,n-1} - \mu_{*,n} < 0$, where the only difference is that the factor $2\theta_{j-1}$ in the linear system in Lemma 4.9 is replaced by $2\theta_{n-1} + \theta_n - 1$, proving item (i).

For item (ii), first note that L is a symmetric matrix such that

$$[L]_{i,j} = \begin{cases} 4\theta_{i-1}^2 & i = j \leqslant n-1 \\ \theta_n^2 - 1 & i = j \in \{n, n+1\} \\ -2\theta_{i-1}^2 & j = i+1 \leqslant n \\ -2\theta_{n-1} - \theta_n + 1 & j = i+1 = n+1 \\ -2\theta_{i-1} & i \leqslant n-1, j = n+1 \\ 0 & \text{else}\,. \end{cases}$$

Also, for all $n \geqslant 2$, from

$$\begin{aligned} v_1 &= -2\theta_0(\theta_1 - \varphi_1)\frac{1}{\varphi_1}\left(1 + \frac{\theta_1 - 1}{2\theta_0\varphi_2}\right) \\ &= -2\left(\frac{1+\sqrt{5}}{2} - \frac{3+\sqrt{5}}{4}\right)(3-\sqrt{5})\left(1 + \frac{\sqrt{5}-1}{4\varphi_2}\right) \\ &= -2(\sqrt{5}-2)\left(1 + \frac{\sqrt{5}-1}{4\varphi_2}\right) \end{aligned} \tag{4.4}$$

and $1 \leqslant \varphi_2 < 2$ we have $-\frac{\sqrt{5}-1}{2} \leqslant v_1 < -\frac{1}{2}$. Thus choosing $\xi = \frac{\sqrt{5}-1}{4}$, $L - \frac{1}{\xi}vv^T$ is Laplacian: it only suffices to check whether the nondiagonal entries on the $(n+1)$th column are still nonpositive, which is true because the $(1, n+1)$ entry is $-2\theta_0 - \frac{v_1}{\xi} \leqslant -2 + \frac{\sqrt{5}-1}{2\xi} = 0$ and the $(i, n+1)$ entry for $2 \leqslant i \leqslant n$ is $-2\theta_{i-1} - \frac{v_i}{\xi} \leqslant -2 + \frac{1}{2\xi} < 0$ from $-\frac{1}{2} < v_i \leqslant 0$. Thus S is positive semidefinite, proving item (ii).[10] □

As noted in Section 1.3, the composite extension of OGM is equal to the proximal OGM (POGM) introduced in [47], by the following proposition. POGM updates as[11] (with $y_0 = z_0 = x_0$)

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

for $0 \leqslant k \leqslant n-1$ (recall that $\alpha_{k,k-1} = 1 + \frac{2\theta_{k-1}-1}{\theta_k}$; for $k = 0$, $\alpha_{k,k-1}$ is not used as $z_k - x_k = 0$), and outputs the final iterate x_n.

We note that the equivalence between POGM and the composite extension of OGM is natural, as they share common principles. As explained in [47, Section 4.3], POGM is constructed based on two principles: the algorithm being equivalent to OGM when $h \equiv 0$, and the algorithm staying at the optimum point (i.e., $x_{k-1} = x_k = x_*$ implies $x_{k+1} = x_*$). One can deduce that both of these are satisfied by composite extension.

Proposition 4.10 (Efficient form as POGM)**.** *Fix $n \in \mathbb{N}$, and let $\{(\widetilde{x}_k, \widetilde{y}_k, \widetilde{z}_k) : 0 \leqslant k \leqslant n\}$ be the iterates of POGM and $\{x_k : 0 \leqslant k \leqslant n\}$ be the iterates of the composite extension of OGM, with $\widetilde{x}_0 = x_0$. Then $\widetilde{x}_k = x_k$ for all $1 \leqslant k \leqslant n$.*

Proof. Define $x_0^- := x_0$ and $x_k^- := x_k + \alpha_{k,k-1}s_k$ for $1 \leqslant k \leqslant n$. We use (strong) induction on k to show that

$$\widetilde{x}_k = x_k, \widetilde{z}_k = x_k^-$$

for all k (the case $k = 0$ holds by definition).

For $k = 1$, we have

$$\widetilde{z}_1 = \widetilde{y}_1 + \frac{\theta_0}{\theta_1}(\widetilde{y}_1 - \widetilde{x}_0) = \widetilde{x}_0 - \left(1 + \frac{\theta_0}{\theta_1}\right)\nabla f(\widetilde{x}_0) = x_0 - \alpha_{1,0}g_0 = x_1^-,$$

[10] For $n = 1$, from $L = \begin{bmatrix} 3 & -3 \\ -3 & 3 \end{bmatrix}$ and $v_1 = -1$ we can choose $\xi = \frac{1}{3}$. For $n \geqslant 3$, the smallest value of ξ that our approach can take is $\xi = -\frac{v_1}{2} = \frac{15\sqrt{5}-17-\sqrt{1942-862\sqrt{5}}}{44} \approx 0.2894$ from (4.4), slightly smaller than $\frac{\sqrt{5}-1}{4} \approx 0.3090$.

[11] In [47], an additional notation of γ_k is introduced there instead of $\alpha_{k+1,k}$; by (4.1), they are identical.

and thus $\widetilde{x}_1 = \mathrm{prox}_{\alpha_{1,0}h}(x_1^-) = x_1$. Assume that the result hold for $1, \ldots, k$. Then

$$
\begin{aligned}
\widetilde{z}_{k+1} &= \widetilde{y}_{k+1} + \frac{\theta_k - 1}{\theta_{k+1}}\left(\widetilde{y}_{k+1} - \widetilde{y}_k + \frac{\widetilde{z}_k - \widetilde{x}_k}{\alpha_{k,k-1}}\right) + \frac{\theta_k}{\theta_{k+1}}(\widetilde{y}_{k+1} - \widetilde{x}_k) \\
&= x_k - g_k + \frac{\theta_k - 1}{\theta_{k+1}}(x_k - x_{k-1} - (g_k - g_{k-1}) + s_k) - \frac{\theta_k}{\theta_{k+1}} g_k \\
&= x_k + \frac{\theta_k - 1}{\theta_{k+1}}(g_{k-1} + s_k) - \left(1 + \frac{2\theta_k - 1}{\theta_{k+1}}\right) g_k - \frac{\theta_k - 1}{\theta_{k+1}} \sum_{j=0}^{k-1} \alpha_{k,j}(g_j + s_{j+1}) \\
&= x_k - \sum_{j=0}^{k-2} \frac{\theta_k - 1}{\theta_{k+1}} \alpha_{k,j}(g_j + s_{j+1}) - \frac{\theta_k - 1}{\theta_{k+1}}(\alpha_{k,k-1} - 1)(g_{k-1} + s_k) - \left(1 + \frac{2\theta_k - 1}{\theta_{k+1}}\right) g_k \\
&= x_{k+1}^-\,,
\end{aligned}
$$

where the first equality is from the induction hypothesis and the last equality is from the definition of $\alpha_{k+1,j}$. Thus, $\widetilde{x}_{k+1} = x_{k+1}$. □

5 Applications to gradient norm minimization

We now consider the performance metric of gradient norm rather than objective function. As in Section 4, we consider composite extensions of both stepsize-accelerated GD and OGM-G (i.e., a gradient norm version of OGM). For each result, it only suffices to prove item (i) of Theorem 2.8, as item (ii) is guaranteed by a certain choice of ξ' (see the details therein). As the arguments are fairly similar to that in Section 4, we only present the main conceptual steps and defer the details to Appendix C.

5.1 Stepsize-accelerated proximal GD

To the best of our knowledge, the silver stepsize schedule does not admit coefficients λ' satisfying (2.3). However, a slightly modified stepsize presented in [25] satisfies the condition. Compared to the silver stepsizes, this stepsize schedule has larger values in its first half.

Definition 5.1. *For $k \in \mathbb{N}$, define $\tau_1 := 4$ and (recall $\rho = 1 + \sqrt{2}$)*

$$
\tau_{k+1} := \frac{1}{2}\left(\tau_k + 4\rho^k + \sqrt{\tau_k^2 + 8\rho^k \tau_k}\right).
$$

Also, define $\eta_k := 1 + \frac{\sqrt{\tau_k^2 + 8\rho^k\tau_k} - \tau_k}{4}$. Finally, define the stepsize $w^{(k)}$ (of length $n = 2^k - 1$) as $w^{(1)} = [3/2]$,

$$
w^{(k+1)} := [w^{(k)}, \eta_k, \pi^{(k)}]\,.
$$

The multipliers $\lambda' = \lambda'^{(k)}$ for stepsize $w^{(k)}$ are similar to those for silver stepsizes in terms of their recursive definition.

Proposition 5.2 (Multipliers for unconstrained GD; [25, Proposition 1])**.** *For $k \in \mathbb{N}$ and $n = 2^k - 1$, GD with stepsize $w^{(k)}$ satisfies (2.3) with*

$$
\begin{aligned}
\lambda' &= \lambda'^{(k)}, \\
R'_n &= \tau_k - 1.
\end{aligned}
$$

Here, $\lambda'^{(k)}$ is recursively defined as

$$
\begin{aligned}
\lambda'^{(1)} &:= \begin{array}{c|cc} & 0 & 1 \\ \hline 0 & 0 & 2 \\ 1 & 1 & 0 \end{array}\,, \\
\lambda'^{(k+1)} &:= \underbrace{\lambda'^{(k+1),\mathrm{rec}}}_{\textit{recursion}} + \underbrace{\lambda'^{(k+1),\mathrm{sp}}}_{\textit{sparse correction}} + \underbrace{\lambda'^{(k+1),\mathrm{lr}}}_{\textit{low-rank correction}}\,,
\end{aligned}
$$

where $\lambda'^{(k+1),\mathrm{rec}}, \lambda'^{(k+1),\mathrm{sp}}, \lambda'^{(k+1),\mathrm{lr}}$ *are defined as (with* $\bar{\lambda}^{(k)}$ *as previously defined in Proposition 4.1)*

$$\begin{aligned}
\lambda'^{(k+1),\mathrm{rec}}_{i,j} &:= \lambda'^{(k)}_{i,j}\mathbf{1}_{\{0\leqslant i,j\leqslant n\}} + \frac{\tau_{k+1}}{\rho^{2k}}\bar{\lambda}^{(k)}_{i-n-1,j-n-1}\mathbf{1}_{\{n+1\leqslant i,j\leqslant 2n+1\}},\\
\lambda'^{(k+1),\mathrm{sp}}_{i,j} &:= \frac{\tau_{k+1}}{2\rho^{2k}}\mathbf{1}_{\{(i,j)=(n,2n+1)\}} + \left(\frac{\tau_{k+1}}{2\rho^{k}} - 1\right)\mathbf{1}_{\{(i,j)=(2n+1,n)\}},\\
\bar{\lambda}'^{(k+1),\mathrm{lr}}_{i,j} &:= \frac{\tau_{k+1}}{2\rho^{2k}}\pi^{(k)}_{j-n}\mathbf{1}_{\{i=n,n+1\leqslant j\leqslant 2n\}} + \frac{\tau_{k+1}}{2\rho^{2k}}\pi^{(k)}_{j-n}\mathbf{1}_{\{i=2n+1,n+1\leqslant j\leqslant 2n\}}.
\end{aligned}$$

Using these multipliers, we obtain an $O(1/n^{\log_2 \rho})$ rate for proximal GD with the same stepsizes. This establishes the phenomenon of stepsize acceleration for the composite setting, when measured in gradient norm—previously only known for the unconstrained setting [25].

Theorem 5.3 (Convergence rate of proximal GD)**.** *For* $k \in \mathbb{N}$ *and* $n = 2^k - 1$*, proximal GD with stepsize* $w^{(k)}$ *satisfies*

$$\|\nabla f(x_n) + s_n\|^2 \leqslant \frac{2\sqrt{2}}{\tau_k}(F_0 - F_n) \sim \frac{\sqrt{2}(\rho - \sqrt{\rho})}{n^{\log_2 \rho}}(F_0 - F_n)\,,$$

where $s_n \in \partial h(x_n)$*.*

Proof sketch. The proof is similar to that of Theorem 4.2. A key to that result was a structured inequality in $\lambda^{(k)}$ (Lemma 4.4), which also holds for $\lambda'^{(k)}$ from its recursive definition. See Appendix C.1 for details. □

5.2 Proximal OGM-G

OGM-G [33] is a (n-step) first-order method with stepsize matrix H, whose entry is recursively defined as (recall the definition of $\{\theta_i : 0 \leqslant i \leqslant n\}$ from (4.2))

$$\alpha_{i+1,j} = \begin{cases} \frac{\theta_{n-j-1}-1}{\theta_{n-j}}\alpha_{i+1,j+1} & 0 \leqslant j \leqslant i-2\\ \frac{\theta_{n-j-1}-1}{\theta_{n-j}}(\alpha_{i+1,i} - 1) & j = i-1\\ 1 + \frac{2\theta_{n-i-1}-1}{\theta_{n-i}} & j = i\,. \end{cases}$$

As in the case of OGM, the multipliers λ' for OGM-G are simple. Indeed, these can be considered as certain transformation of the multipliers λ for OGM in an appropriate sense [34].

Proposition 5.4 (Multipliers for OGM-G; [33, Theorem 6.1])**.** *OGM-G satisfies* (2.3) *with*

$$\begin{aligned}
(1/\theta_n^2)\lambda'_{i,j} &= \begin{cases} \frac{1}{2\theta^2_{n-j}} & j = i+1, 0 \leqslant i \leqslant n-1\\ \frac{1}{2\theta^2_{n-j-1}} - \frac{1}{2\theta^2_{n-j}} & i = n, 1 \leqslant j \leqslant n-1\\ \frac{1}{2\theta^2_{n-1}} - \frac{1}{\theta^2_n} & i = n, j = 0\\ 0 & \textit{else}, \end{cases}\\
R'_n &= \theta_n^2 - 1.
\end{aligned}$$

Using these multipliers, we establish the following convergence guarantee for the composite extension of OGM-G. This bound is almost 1/10 of the previously known best result in [34].

Theorem 5.5 (Convergence rate of proximal OGM-G)**.** *For* $n \geqslant 2$*, the composite extension of OGM-G satisfies*[12]

$$\|\nabla f(x_n) + s_n\|^2 \leqslant \frac{2(\sqrt{5}-1)}{\theta_n^2}(F_0 - F_n) \sim \frac{4(\sqrt{5}-1)}{n^2}(F_0 - F_n)\,,$$

where $s_n \in \partial h(x_n)$*.*

[12] For $n = 1$, $\|\nabla f(x_n) + s_n\|^2 \leqslant \frac{8}{3\theta_n^2}(F_0 - F_n) = \frac{2}{3}(F_0 - F_n)$ holds. See the proof for details.

Proof sketch. The overall proof structure is similar to that of proximal OGM (Theorem 4.7). The proof is more technical here (although straightforward) due to each component in the formula $\mu' = \widetilde{\lambda}' + \widetilde{H}^{-1}\widehat{\lambda}'$ (Definition 3.4) being complicated. In particular, for the nonnegativity of μ', some entries are nonnegative only through the sum of $\widetilde{\lambda}'$ and $\widetilde{H}^{-1}\widehat{\lambda}'$: for each j, the $(j-1, j)$ entry of $\widetilde{\lambda}'$ is negative and $(j-2, j)$ entry of $\widetilde{H}^{-1}\widehat{\lambda}'$ is negative. For details, see Appendix C.2. □

We also derive the following efficient representation of the composite extension of OGM-G, which we call *P-OGM-G* (proximal OGM-G). The derivation and analysis of this representation closely resemble those of POGM (Proposition 4.10). P-OGM-G updates as (with $y_0 = z_0 = x_0$)

$$\begin{aligned}
y_{k+1} &= x_k - \nabla f(x_k),\\
z_{k+1} &= y_{k+1} + \frac{(\theta_{n-k}-1)(2\theta_{n-k-1}-1)}{\theta_{n-k}(2\theta_{n-k}-1)}\left(y_{k+1} - y_k + \frac{z_k - x_k}{\alpha_{k,k-1}}\right) + \frac{2\theta_{n-k-1}-1}{2\theta_{n-k}-1}(y_{k+1} - x_k),\\
x_{k+1} &= \operatorname{prox}_{\alpha_{k+1,k}h}(z_{k+1}),
\end{aligned}$$

for $0 \leqslant k \leqslant n-1$ (recall that $\alpha_{k,k-1} = 1 + \frac{2\theta_{n-k}-1}{\theta_{n-k+1}}$; for $k = 0$, $\alpha_{k,k-1}$ is not used as $z_k - x_k = 0$), and outputs the final iterate x_n.

Proposition 5.6 (Efficient form as P-OGM-G)**.** *Fix $n \in \mathbb{N}$, and let $\{(\widetilde{x}_k, \widetilde{y}_k, \widetilde{z}_k) : 0 \leqslant k \leqslant n\}$ be the iterates of P-OGM-G and $\{x_k : 0 \leqslant k \leqslant n\}$ be the iterates of the composite extension of OGM-G, with $\widetilde{x}_0 = x_0$. Then $\widetilde{x}_k = x_k$ for all $1 \leqslant k \leqslant n$.*

Proof. As in the proof of Proposition 4.10, by defining $x_0^- := x_0$ and $x_k^- := x_k + \alpha_{k,k-1}s_k$ for $1 \leqslant k \leqslant n$, we inductively show that $\widetilde{x}_k = x_k$ and $\widetilde{z}_k = x_k^-$. For $k = 0$ this holds by definition. For $k = 1$, we have

$$\widetilde{z}_1 = \widetilde{y}_1 + \frac{(\theta_n - 1)(2\theta_{n-1}-1)}{\theta_n(2\theta_n - 1)}(\widetilde{y}_1 - \widetilde{y}_0) + \frac{2\theta_{n-1}-1}{2\theta_n - 1}(\widetilde{y}_1 - \widetilde{x}_0) = \widetilde{x}_0 - \left(1 + \frac{2\theta_{n-1}-1}{\theta_n}\right)\nabla f(\widetilde{x}_0) = x_0 - \alpha_{1,0}g_0 = x_1^-,$$

and thus $\widetilde{x}_1 = x_1$. Assume that the result holds for $1, \ldots, k$. Then

$$\begin{aligned}
\widetilde{z}_{k+1} &= \widetilde{y}_{k+1} + \frac{(\theta_{n-k}-1)(2\theta_{n-k-1}-1)}{\theta_{n-k}(2\theta_{n-k}-1)}\left(\widetilde{y}_{k+1} - \widetilde{y}_k + \frac{\widetilde{z}_k - \widetilde{x}_k}{\alpha_{k,k-1}}\right) + \frac{2\theta_{n-k-1}-1}{2\theta_{n-k}-1}(\widetilde{y}_{k+1} - \widetilde{x}_k)\\
&= x_k - g_k + \frac{(\theta_{n-k}-1)(2\theta_{n-k-1}-1)}{\theta_{n-k}(2\theta_{n-k}-1)}(x_k - x_{k-1} - (g_k - g_{k-1}) + s_k) - \frac{2\theta_{n-k-1}-1}{2\theta_{n-k}-1}g_k\\
&= x_k + \frac{(\theta_{n-k}-1)(2\theta_{n-k-1}-1)}{\theta_{n-k}(2\theta_{n-k}-1)}(g_{k-1} + s_k) - \left(1 + \frac{2\theta_{n-k-1}-1}{\theta_{n-k}}\right)g_k\\
&\quad - \frac{(\theta_{n-k}-1)(2\theta_{n-k-1}-1)}{\theta_{n-k}(2\theta_{n-k}-1)}\sum_{j=0}^{k-1}\alpha_{k,j}(g_j + s_{j+1})\\
&= x_k - \sum_{j=0}^{k-2}\frac{(\theta_{n-k}-1)(2\theta_{n-k-1}-1)}{\theta_{n-k}(2\theta_{n-k}-1)}\alpha_{k,j}(g_j + s_{j+1})\\
&\quad - \frac{(\theta_{n-k}-1)(2\theta_{n-k-1}-1)}{\theta_{n-k}(2\theta_{n-k}-1)}(\alpha_{k,k-1} - 1)(g_{k-1} + s_k) - \left(1 + \frac{2\theta_{n-k-1}-1}{\theta_{n-k}}\right)g_k\\
&= x_{k+1}^-,
\end{aligned}$$

where the first equality is from the induction hypothesis and the last equality is from [33, equation 32]. Thus, $\widetilde{x}_{k+1} = x_{k+1}$. □

6 Discussion

In this paper, we developed a general-purpose result for extending optimized first-order methods from the unconstrained setting to the composite setting, and we applied this to different combinations of algorithms and performance metrics. Our work suggests several directions for further inquiry, such as the following.

Different classes of methods. Our result connects methods in unconstrained and composite settings; can more general connections be made? These can be, for example, between different problem settings (as in this paper), between different performance metrics (as in H-duality), or between more general algorithms (e.g., that use line search or adaptive updates). Developing these connections is a fundamental question in its own right and could help unify the design and analysis of algorithms.

Further understanding of optimized methods. The starting point for this work is the observation that optimized methods admit simple proof structures, namely the sum-of-squares term is of rank 0 or 1 (see the technical overview in Section 1.2). This phenomenon holds beyond the optimized methods we covered here [45, Chapter 5], and for specific choices of stepsizes for GD it admits an equivalent characterization in terms of worst-case functions [24, Proposition 4]. On the other hand, it is unclear whether proofs for "non-optimized" methods such as constant-stepsize GD or AGD admit such structures (see e.g., [16, 31]). It is an interesting open question to understand the generality of this phenomenon and whether it hints at an underlying unified theory for optimized methods.

A full reduction. While our main results (Theorems 2.6 and 2.8) provide reduction-style approaches that are much simpler to invoke than proving convergence rates from scratch, these reductions are not fully black-box as one needs to verify items (i) and (ii). It would be helpful for the design and analysis of algorithms if these reductions could be made fully black-box, or more generally if there are other unified approaches for solving the semidefinite programs arising from PEP-type analyses.

Acknowledgements

This work was partially supported by a Sloan Research Fellowship and a Seed Grant Award from Apple.

References

[1] Jason M. Altschuler. Greed, hedging, and acceleration in convex optimization. Master's thesis, Massachusetts Institute of Technology, 2018.

[2] Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization. *Math. Program.*, 2024.

[3] Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *J. ACM*, 72(2):1–38, 2025.

[4] Rina Foygel Barber and Wooseok Ha. Gradient descent with non-convex constraints: local concavity determines convergence. *Inf. Inference*, 7(4):755–806, 2018.

[5] Mathieu Barré, Adrien Taylor, and Alexandre d'Aspremont. Complexity guarantees for Polyak steps with momentum. In *Conference on Learning Theory*, volume 125, pages 452–478. PMLR, 2020.

[6] Amir Beck. *First-order methods in optimization*, volume 25 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, 2017.

[7] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Process.*, 18(11):2419–2434, 2009.

[8] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[9] Jinho Bok and Jason M. Altschuler. Accelerating proximal gradient descent via silver stepsizes. *arXiv preprint arXiv:2412.05497, to appear in Conference on Learning Theory*, 2025.

[10] Nizar Bousselmi, Julien M. Hendrickx, and François Glineur. Interpolation conditions for linear operators and applications to performance estimation problems. *SIAM J. Optim.*, 34(3):3033–3063, 2024.

[11] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

[12] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. "Convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International Conference on Machine Learning*, volume 70, pages 654–663. PMLR, 2017.

[13] Antoine Daccache. Performance estimation of the gradient method with fixed arbitrary step sizes. Master's thesis, UCL - Ecole polytechnique de Louvain, 2019.

[14] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

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

[16] Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Math. Program.*, 145(1–2):451–482, 2014.

[17] Yoel Drori and Marc Teboulle. An optimal variant of Kelley's cutting-plane method. *Math. Program.*, 160(1-2):321–351, 2016.

[18] Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021.

[19] Diego Eloi. Worst-case functions for the gradient method with fixed variable step sizes. Master's thesis, UCL - Ecole polytechnique de Louvain, 2022.

[20] Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

[21] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2):59–99, 2016.

[22] Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Optimal first-order methods for convex functions with a quadratic upper bound. *arXiv preprint arXiv:2205.15033*, 2022.

[23] Benjamin Grimmer. Provably faster gradient descent via long steps. *SIAM J. Optim.*, 34(3):2588–2608, 2024.

[24] Benjamin Grimmer, Kevin Shu, and Alex L. Wang. Composing optimized stepsize schedules for gradient descent. *arXiv preprint arXiv:2410.16249*, 2024.

[25] Benjamin Grimmer, Kevin Shu, and Alex L. Wang. Accelerated objective gap and gradient norm convergence for gradient descent via long steps. *INFORMS J. Optim.*, 7(2):156–169, 2025.

[26] Guoyong Gu and Junfeng Yang. Tight convergence rate in subgradient norm of the proximal point algorithm. *arXiv preprint arXiv:2301.03175*, 2023.

[27] Uijeong Jang, Shuvomoy Das Gupta, and Ernest K. Ryu. Computer-assisted design of accelerated composite optimization methods: OptISTA. *arXiv preprint arXiv:2305.15704*, 2024.

[28] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[29] Koulik Khamaru and Martin J. Wainwright. Convergence guarantees for a class of non-convex and non-smooth optimization problems. *J. Mach. Learn. Res.*, 20(154):1–52, 2019.

[30] Donghwan Kim. Accelerated proximal point method for maximally monotone operators. *Math. Program.*, 190(1-2):57–87, 2021.

[31] Donghwan Kim and Jeffrey A. Fessler. Optimized first-order methods for smooth convex minimization. *Math. Program.*, 159(1-2):81–107, 2016.

[32] Donghwan Kim and Jeffrey A. Fessler. Another look at the fast iterative shrinkage/thresholding algorithm (FISTA). *SIAM J. Optim.*, 28(1):223–250, 2018.

[33] Donghwan Kim and Jeffrey A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *J. Optim. Theory Appl.*, 188(1):192–219, 2021.

[34] Jaeyeon Kim, Asuman Ozdaglar, Chanwoo Park, and Ernest K. Ryu. Time-reversed dissipation induces duality between minimizing gradient norm and function value. In *Advances in Neural Information Processing Systems*, volume 36, pages 23389–23440, 2023.

[35] Jaeyeon Kim, Chanwoo Park, Asuman Ozdaglar, Jelena Diakonikolas, and Ernest K. Ryu. Mirror duality in convex optimization. *arXiv preprint arXiv:2311.17296*, 2023.

[36] Jungbin Kim. A proof of the exact convergence rate of gradient descent. *arXiv preprint arXiv:2412.04427*, 2025.

[37] Jongmin Lee, Chanwoo Park, and Ernest Ryu. A geometric structure of acceleration and its role in making gradients small fast. In *Advances in Neural Information Processing Systems*, volume 34, pages 11999–12012, 2021.

[38] Yu. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[39] Yurii Nesterov. *Lectures on convex optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, 2018.

[40] Chanwoo Park, Jisun Park, and Ernest K. Ryu. Factor-$\sqrt{2}$ acceleration of accelerated gradient methods. *Appl. Math. Optim.*, 88(77):1–38, 2023.

[41] Jisun Park and Ernest K. Ryu. Exact optimal accelerated complexity for fixed-point iterations. In *International Conference on Machine Learning*, volume 162, pages 17420–17457. PMLR, 2022.

[42] Teodor Rotaru, François Glineur, and Panagiotis Patrinos. Exact worst-case convergence rates of gradient descent: a complete analysis for all constant stepsizes over nonconvex and convex functions. *arXiv preprint arXiv:2406.17506*, 2024.

[43] Ernest K. Ryu, Adrien B. Taylor, Carolina Bergeling, and Pontus Giselsson. Operator splitting performance estimation: tight contraction factors and optimal parameter selection. *SIAM J. Optim.*, 30(3):2251–2271, 2020.

[44] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *J. Mach. Learn. Res.*, 17(153):1–43, 2016.

[45] Adrien Taylor. *Towards principled and systematic approaches to the analysis and design of optimization algorithms*. Habilitation à diriger des recherches, Université Paris Sciences & Lettres, 2024.

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

[47] Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM J. Optim.*, 27(3):1283–1313, 2017.

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

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

[50] Marc Teboulle and Yakov Vaisbourd. An elementary approach to tight worst case complexity analysis of gradient based methods. *Math. Program.*, 201(1-2):63–96, 2023.

[51] Bofan Wang, Shiqian Ma, Junfeng Yang, and Danqing Zhou. Relaxed proximal point algorithm: Tight complexity bounds and acceleration without momentum. *arXiv preprint arXiv:2410.08890*, 2024.

[52] Ashia C. Wilson, Ben Recht, and Michael I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *J. Mach. Learn. Res.*, 22(113):1–34, 2021.

[53] Taeho Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning*, volume 139, pages 12098–12109. PMLR, 2021.

[54] TaeHo Yoon and Ernest K. Ryu. Accelerated minimax algorithms flock together. *SIAM J. Optim.*, 35(1):180–209, 2025.

[55] David Young. On Richardson's method for solving linear systems with positive definite matrices. *J. Math. Physics*, 32(1-4):243–255, 1953.

[56] Moslem Zamani, Hadi Abbaszadehpeivasti, and Etienne de Klerk. The exact worst-case convergence rate of the alternating direction method of multipliers. *Math. Program.*, 208(1-2):243–276, 2024.

[57] Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *arXiv preprint arXiv:2307.11134*, 2023.

[58] Zehao Zhang and Rujun Jiang. Accelerated gradient descent by concatenation of stepsize schedules. *arXiv preprint arXiv:2410.12395*, 2024.

[59] Zihan Zhang, Jason D. Lee, Simon S. Du, and Yuxin Chen. Anytime acceleration of gradient descent. *arXiv preprint arXiv:2411.17668*, 2024.

A Proof of Theorem 2.6

A.1 Implications of (2.1)

In order to establish (2.2), we first parse out (2.1) by comparing the coefficients on its each side.

Linear form. By comparing the coefficients of the linear form in $\{f_i\}$, we obtain

$$\begin{aligned} \lambda_{i,\bullet} - \lambda_{\bullet,i} &= 0, \quad 0 \leqslant i \leqslant n-1, \\ \lambda_{n,\bullet} - \lambda_{\bullet,n} &= -R_n, \\ \lambda_{*,\bullet} &= R_n. \end{aligned} \tag{A.1}$$

Quadratic form. Here, we compare the coefficients of the quadratic form in $x_0 - x_*, \{g_i\}$. The coefficient of $\langle g_i, g_j\rangle$ for $0 \leqslant i < j \leqslant n$ on the left hand side of (2.1) is

$$\sum_{k=i+1}^{n} \widetilde{\alpha}_{k,i}\lambda_{k,j} - \widetilde{\alpha}_{j,i}\lambda_{\bullet,j} + \sum_{k=j+1}^{n} \widetilde{\alpha}_{k,j}\lambda_{k,i} + \lambda_{i,j} + \lambda_{j,i} + \gamma_i\gamma_j = 0\,,$$

and for $0 \leqslant i = j \leqslant n$ the coefficient is $\sum_{k=i+1}^{n} \widetilde{\alpha}_{k,i}\lambda_{k,i} - \frac{1}{2}(\lambda_{\bullet,i} + \lambda_{i,\bullet}) + \frac{1}{2}\gamma_i^2 = 0$. This is equivalent to

$$\widehat{\lambda} + \widetilde{H}\widetilde{\lambda} + (\widetilde{H}\widetilde{\lambda})^T = -\widetilde{\gamma}\widetilde{\gamma}^T, \tag{A.2}$$

$$\begin{bmatrix} \lambda_{0,n} + \lambda_{n,0} \\ \vdots \\ \lambda_{n-1,n} + \lambda_{n,n-1} \end{bmatrix} + \widetilde{H} \begin{bmatrix} \lambda_{1,n} \\ \vdots \\ \lambda_{n-1,n} \\ -\lambda_{\bullet,n} \end{bmatrix} = -\gamma_n\widetilde{\gamma}, \tag{A.3}$$

$$\lambda_{\bullet,n} + \lambda_{n,\bullet} = \gamma_n^2. \tag{A.4}$$

Also, by comparing the coefficient of $\langle x_0 - x_*, g_i\rangle$ for $0 \leqslant i \leqslant n$, we obtain

$$\lambda_{*,i} = \gamma_i, \quad 0 \leqslant i \leqslant n\,. \tag{A.5}$$

A.2 Verification of (2.2)

Assuming (2.1) is true, we show that (2.2) is true by matching coefficients. It is clear that both sides of (2.2) are linear in $\{f_i\}, \{h_i\}$ and quadratic in $x_0 - x_*, \{g_i\}, \{s_i\}$.

Linear form. The coefficients of $\{f_i\}$ are from (2.1). For the coefficients of $\{h_i\}$, it suffices to show that

$$\widetilde{\mu}\mathbf{1}_n = -R_n e_n\,.$$

From $\widetilde{\mu} = \widetilde{\lambda} + \widetilde{H}^{-1}(\widehat{\lambda} - \widetilde{\gamma}\widetilde{\sigma}^T)$ and $\widetilde{\lambda}\mathbf{1}_n = \begin{bmatrix} -\lambda_{1,n} \\ \vdots \\ -\lambda_{n-1,n} \\ \lambda_{\bullet,n} - R_n \end{bmatrix}$ from (A.1), it suffices to show that

$$\widetilde{H}^{-1}(\widehat{\lambda} - \widetilde{\gamma}\widetilde{\sigma}^T)\mathbf{1}_n = \begin{bmatrix} \lambda_{1,n} \\ \vdots \\ \lambda_{n-1,n} \\ -\lambda_{\bullet,n} \end{bmatrix} \Leftrightarrow (\widehat{\lambda} - \widetilde{\gamma}\widetilde{\sigma}^T)\mathbf{1}_n = \widetilde{H} \begin{bmatrix} \lambda_{1,n} \\ \vdots \\ \lambda_{n-1,n} \\ -\lambda_{\bullet,n} \end{bmatrix}.$$

The right hand side is equal to the left hand side because

$$\begin{aligned}
-\begin{bmatrix}\lambda_{0,n}+\lambda_{n,0}\\ \vdots \\ \lambda_{n-1,n}+\lambda_{n,n-1}\end{bmatrix}-\gamma_n\widetilde{\gamma} &= -\begin{bmatrix}\lambda_{0,n}+\lambda_{n,0}\\ \vdots \\ \lambda_{n-1,n}+\lambda_{n,n-1}\end{bmatrix}-\widetilde{\gamma}-(\sum\widetilde{\sigma})\widetilde{\gamma}\\
&= -\begin{bmatrix}\lambda_{0,n}+\lambda_{n,0}+\lambda_{*,0}\\ \vdots \\ \lambda_{n-1,n}+\lambda_{n,n-1}+\lambda_{*,n-1}\end{bmatrix}-(\sum\widetilde{\sigma})\widetilde{\gamma}\\
&= (\widehat{\lambda}-\widetilde{\gamma}\widetilde{\sigma}^T)\mathbf{1}_n\,,
\end{aligned}$$

where the first equality is from

$$\sum\sigma=\sum\widetilde{\sigma}+\sigma_*=\frac{\lambda_{\bullet,n}+\lambda_{n,\bullet}}{\gamma_n}=\frac{\gamma_n^2}{\gamma_n}=\gamma_n\,, \tag{A.6}$$

by (A.4) and the second equality is from (A.5).

Quadratic form. Now, we compare the coefficients of the quadratic forms (in $x_0-x_*, \{g_i\}, \{s_i\}$). The analysis is more technically involved than that for the linear form, because Q_{ij} (in unconstrained setting) is not equal to Q^f_{ij} (in composite setting). Thus, we define and analyze an intermediate quantity that connects these quantities. After this, the comparison is conceptually straightforward yet algebraically tedious.

Setup. It is convenient to first characterize $\sum_{i,j}\lambda_{i,j}Q^f_{ij}$, from (2.1) (which is written in Q_{ij}, for unconstrained setting). For this, define the *pseudo-co-coercivities* $\widetilde{Q}^f_{ij}$ obtained by replacing $(x_0-x_*, g_0, g_1, \dots, g_n)$ from Q_{ij} with $(x_0-x_*, g_0+s_1, g_1+s_2, \dots, g_{n-1}+s_n, g_n+s_{n+1})$, with $s_{n+1}:=0$. Then (2.1) implies

$$\sum_{i,j}\lambda_{i,j}\widetilde{Q}^f_{ij}+\frac{1}{2}\|x_0-x_*-\left[g_0+s_1|g_1+s_2|\dots|g_n+s_{n+1}\right]\gamma\|^2=R_n(f_*-f_n)+\frac{1}{2}\|x_0-x_*\|^2\,. \tag{A.7}$$

The difference between the actual co-coercivities Q^f_{ij} and pseudo-co-coercivities $\widetilde{Q}^f_{ij}$ satisfies

$$\begin{aligned}
Q^f_{ij}-\widetilde{Q}^f_{ij} &= \langle s_{j+1}, x_i-x_j\rangle-\frac{1}{2}\|g_i-g_j\|^2+\frac{1}{2}\|g_i+s_{i+1}-g_j-s_{j+1}\|^2\\
&= \langle s_{j+1}, x_i-x_j\rangle+\frac{1}{2}\|s_{i+1}-s_{j+1}\|^2+\langle g_i-g_j, s_{i+1}-s_{j+1}\rangle\,,
\end{aligned} \tag{A.8}$$

with $s_{*+1}:=s_*$. Therefore, by adding $\sum_{i,j}\lambda_{i,j}(Q^f_{ij}-\widetilde{Q}^f_{ij})$ on both sides of (A.7) using (A.8),

$$\begin{aligned}
&\sum_{i,j}\lambda_{i,j}Q^f_{ij}+\frac{1}{2}\|x_0-x_*-\left[g_0+s_1|g_1+s_2|\dots|g_n+s_{n+1}\right]\gamma\|^2\\
&=R_n(f_*-f_n)+\frac{1}{2}\|x_0-x_*\|^2+\sum_{i\neq *,j}\lambda_{i,j}\left(\langle s_{j+1}, x_i-x_j\rangle+\frac{1}{2}\|s_{i+1}-s_{j+1}\|^2+\langle g_i-g_j, s_{i+1}-s_{j+1}\rangle\right)\\
&\quad+\sum_j\lambda_{*j}\left(\langle s_{j+1}, x_*-x_j\rangle+\frac{1}{2}\|s_{j+1}\|^2-\frac{1}{2}\|s_*\|^2+\langle g_j, s_{j+1}-s_*\rangle\right)\,,
\end{aligned}$$

where we split the sum $\sum_{i,j} \lambda_{i,j}(Q_{ij} - \widetilde{Q}^f_{ij})$ based on $i \neq *$ and $i = *$. Thus, (2.2) is equivalent to

$$\begin{aligned}
\sum_{i,j} \mu_{i,j} Q^h_{ij} = R_n(h_* - h_n) + \frac{1}{2}\xi\|x_0 - x_*\|^2 \\
&+ \left\langle x_0 - x_* - \left[g_0 + s_1|g_1 + s_2|\dots|g_n + s_{n+1}\right]\gamma, \left[s_1|s_2|\dots|s_n|s_*\right]\sigma\right\rangle - \frac{1}{2}\left\|\left[s_1|s_2|\dots|s_n|s_*\right]\sigma\right\|^2 \\
&- \sum_{i\neq *,j} \lambda_{i,j}\left(\langle s_{j+1}, x_i - x_j\rangle + \frac{1}{2}\|s_{i+1} - s_{j+1}\|^2 + \langle g_i - g_j, s_{i+1} - s_{j+1}\rangle\right) \\
&- \sum_{j} \lambda_{*j}\left(\langle s_{j+1}, x_* - x_j\rangle + \frac{1}{2}\|s_{j+1}\|^2 - \frac{1}{2}\|s_*\|^2 + \langle g_j, s_{j+1} - s_*\rangle\right) - \frac{1}{2}\operatorname{Tr}(VSV^T)\,.
\end{aligned} \tag{A.9}$$

Category of vectors. For convenience, we categorize the vectors (that constitute the quadratic form) into different types as follows:

$$\begin{aligned}
g &: g_0, \dots, g_n \ (\underline{\text{g}}\text{radients}), \\
s &: s_1, \dots, s_n \ (\underline{\text{s}}\text{ubgradients}), \\
i &: x_0 - x_* \ (\underline{\text{i}}\text{nitial distance}), \\
o &: s_* = -g_* \ (\underline{\text{o}}\text{ptimum}).
\end{aligned} \tag{A.10}$$

In our subsequent analysis, we divide into different cases based on the combination of these types.

Coefficients of (g, g). These are directly from (2.1).

Coefficients of (g, s). The coefficient of $\langle g_i, s_j\rangle$ on the left hand side of (A.9) for $0 \leqslant i \leqslant n-1, 1 \leqslant j \leqslant n$ is

$$\sum_{k=i+1}^{n} \widetilde{\alpha}_{k,i}\mu_{k,j} - \widetilde{\alpha}_{j,i}\mu_{\bullet,j}\mathbf{1}\{i \leqslant j-1\}\,,$$

whereas the corresponding coefficient on the right hand side of (A.9) is

$$-\gamma_i\sigma_j + \sum_{k=i+1}^{n} \widetilde{\alpha}_{k,i}\lambda_{k,j-1} - \widetilde{\alpha}_{j-1,i}\lambda_{\bullet,j-1}\mathbf{1}\{i \leqslant j-2\} + \lambda_{i,j-1} + \lambda_{j-1,i} - (\lambda_{\bullet,j-1} + \lambda_{j-1,\bullet})\mathbf{1}\{i = j-1\}\,.$$

Thus, the coefficients being matched is equivalent to (with a correspondence of entry $(i,j) \Leftrightarrow \langle g_{i-1}, s_j\rangle$)

$$\widetilde{H}\widetilde{\mu} = -\widetilde{\gamma}\widetilde{\sigma}^T + \widetilde{H}\widetilde{\lambda} + \widehat{\lambda}\,,$$

which is true from (3.2). Note that the alternative form in (3.2) is obtained from (A.2).

The coefficient of $\langle g_n, s_j\rangle$ for $1 \leqslant j \leqslant n$ on the left hand side of (A.9) is 0. For the right hand side it is $-\gamma_n\sigma_j + \lambda_{j-1,n} + \lambda_{n,j-1} = 0$ from (3.1).

Coefficients of (s, s). The coefficient of $\langle s_i, s_j\rangle$ for $1 \leqslant i \leqslant j \leqslant n$ on the left hand side of (A.9) is

$$\begin{cases} \sum_{k=i}^{n} \widetilde{\alpha}_{k,i-1}\mu_{k,j} - \widetilde{\alpha}_{j,i-1}\mu_{\bullet,j} + \sum_{k=j}^{n} \widetilde{\alpha}_{k,j-1}\mu_{k,i} & j > i \\ \sum_{k=i}^{n} \widetilde{\alpha}_{k,i-1}\mu_{k,i} - \widetilde{\alpha}_{i,i-1}\mu_{\bullet,i} & j = i\,, \end{cases}$$

and the corresponding coefficient on the right hand side is

$$\begin{aligned}
&-\gamma_{i-1}\sigma_j - \gamma_{j-1}\sigma_i - \sigma_i\sigma_j + \sum_{k=i}^{n} \widetilde{\alpha}_{k,i-1}\lambda_{k,j-1} - \widetilde{\alpha}_{j-1,i-1}\lambda_{\bullet,j-1} + \sum_{k=j}^{n} \widetilde{\alpha}_{k,j-1}\lambda_{k,i-1} + \lambda_{i-1,j-1} + \lambda_{j-1,i-1} \\
&\quad + \lambda_{i-1,j-1} + \lambda_{j-1,i-1} + \sigma_i\sigma_j \\
&= -\gamma_{i-1}\sigma_j - \gamma_{j-1}\sigma_i + \sum_{k=i}^{n} \widetilde{\alpha}_{k,i-1}\lambda_{k,j-1} - \widetilde{\alpha}_{j-1,i-1}\lambda_{\bullet,j-1} + \sum_{k=j}^{n} \widetilde{\alpha}_{k,j-1}\lambda_{k,i-1} + 2(\lambda_{i-1,j-1} + \lambda_{j-1,i-1})
\end{aligned}$$

if $i < j$ and

$$\begin{aligned}
&- \gamma_{i-1}\sigma_i - \frac{1}{2}\sigma_i^2 + \sum_{k=i}^{n} \widetilde{\alpha}_{k,i-1}\lambda_{k,i-1} - \mu_{\bullet,i}\widetilde{\alpha}_{i,i-1} - \frac{1}{2}(\lambda_{i-1,\bullet} + \lambda_{\bullet,i-1}) - \frac{1}{2}(\lambda_{i-1,\bullet} + \lambda_{\bullet,i-1}) + \frac{1}{2}\sigma_i^2 \\
&= -\gamma_{i-1}\sigma_i + \sum_{k=i}^{n} \widetilde{\alpha}_{k,i-1}\lambda_{k,i-1} - \mu_{\bullet,i}\widetilde{\alpha}_{i,i-1} - (\lambda_{i-1,\bullet} + \lambda_{\bullet,i-1})
\end{aligned}$$

if $i = j$. Thus the coefficients matching is equivalent to

$$\widetilde{H}\widetilde{\mu} + (\widetilde{H}\widetilde{\mu})^T = -(\widetilde{\gamma}\widetilde{\sigma}^T + \widetilde{\sigma}\widetilde{\gamma}^T) + \widetilde{H}\widetilde{\lambda} + (\widetilde{H}\widetilde{\lambda})^T + 2\widehat{\lambda}\,,$$

which holds from (3.2) as $\widetilde{H}\widetilde{\mu} = \widetilde{H}\widetilde{\lambda} + \widehat{\lambda} - \widetilde{\gamma}\widetilde{\sigma}^T$.

Coefficients of remaining combinations. The rest of the combinations for inner products involves $x_0 - x_*$ or $s_* = -g_*$, which only appears in a few entries of (A.9).

- $(g, o), (s, o), (o, o)$: In (A.9), the left hand side has coefficients 0, and the right hand side has coefficients (respectively for $\langle g_i, s_* \rangle, \langle s_j, s_* \rangle, \|s_*\|^2$) $-\gamma_i\sigma_* + \lambda_{*,i} = 0, -\gamma_{j-1}\sigma_* - \sigma_j\sigma_* + \gamma_{j-1}\sigma_* + \sigma_j\sigma_* = 0, -\frac{1}{2}\sigma_*^2 + \frac{1}{2}\sigma_*^2 = 0$.
- $(i, i), (g, i)$: Straightforward from (2.1).
- $(s, i), (o, i)$: In (A.9), for $\langle s_j, x_0 - x_* \rangle$ the left hand side has coefficients $\mu_{*,j}$ and the right hand side has coefficients $\sigma_j + \lambda_{*,j-1} - v_j = \mu_{*,j}$.

A.3 Other properties

In this section, we conclude the proof of Theorem 2.6 by providing the remaining details.

$\sum v = 0$. From the definition of v in (3.3),

$$\begin{aligned}
\sum v &= \sum_{k=1}^{n}(\sigma_k + \lambda_{*,k-1} - \mu_{*,k}) + \sigma_* \\
&= \sum \sigma + (\lambda_{*,\bullet} - \lambda_{*,n}) - \mu_{*,\bullet}\,,
\end{aligned}$$

where $\sum \sigma = \lambda_{*,n} = \gamma_n$ from (A.5) and (A.6). Thus $\sum v = R_n - \mu_{*,\bullet}$. The term $\mu_{*,\bullet}$ can be calculated as follows:

$$\begin{aligned}
\mu_{*,\bullet} &= -\mathbf{1}_n^T \widetilde{\mu} \mathbf{1}_n \\
&= -\mathbf{1}_n^T \widetilde{H}^{-1} \left(- \begin{bmatrix} \lambda_{n,0} + \lambda_{*,0} + \lambda_{0,n} \\ \vdots \\ \lambda_{n,n-1} + \lambda_{*,n-1} + \lambda_{n-1,n} \end{bmatrix} - \left(\sum \widetilde{\sigma}\right) \widetilde{\gamma} \right) - \mathbf{1}_n^T \begin{bmatrix} \lambda_{1,\bullet} - \lambda_{1,n} - \lambda_{\bullet,1} \\ \vdots \\ \lambda_{n-1,\bullet} - \lambda_{n-1,n} - \lambda_{\bullet,n-1} \\ \lambda_{n,\bullet} \end{bmatrix} \\
&= \mathbf{1}_n^T \widetilde{H}^{-1} (\widetilde{\gamma} + \widetilde{\sigma}) \gamma_n + \lambda_{\bullet,n} - \lambda_{*,n} - \lambda_{n,\bullet} \\
&= -\mathbf{1}_n^T \begin{bmatrix} \lambda_{1,n} \\ \vdots \\ -\lambda_{\bullet,n} \end{bmatrix} - \lambda_{*,n} + R_n \\
&= \lambda_{*,n} - \lambda_{*,n} + R_n = R_n\,,
\end{aligned}$$

where the second equality is from the definition of $\widetilde{\mu}$ (3.2), the third equality is from (A.1) and (A.6), and the fourth equality is from (A.3). Thus, $\sum v = 0$.

L is Laplacian. Recall from (3.3) that $L = -\begin{bmatrix} \widehat{\lambda} & \widetilde{\gamma} \\ \widetilde{\gamma}^T & -\lambda_{*,\bullet} \end{bmatrix} - \sigma\sigma^T$. Here, the nondiagonal entries are clearly nonpositive. The jth row sum for $1 \leqslant j \leqslant n$ is

$$\lambda_{j-1,n} + \lambda_{n,j-1} + \lambda_{*,j-1} - \gamma_{j-1} - \left(\sum \sigma\right)\sigma_j = \sigma_j \left(\gamma_n - \sum \sigma\right) = 0\,,$$

where the first equality is from the definition of σ (3.1), $\lambda_{*,j-1} = \gamma_{j-1}$ from (A.5) and the last equality is from (A.6). The $(n+1)$th row sum is $-\sum \widetilde{\gamma} + \lambda_{*,\bullet} - (\sum \sigma)\,\sigma_* = \lambda_{*,\bullet} - \sum \gamma = 0$ from (A.5).

B Proof of Theorem 2.8

As in Appendix A, we show that (2.4) is true by matching coefficients (whose both sides are linear in $\{f_i\}, \{h_i\}$ and quadratic in $x_0 - x_*, \{g_i\}, \{s_i\}$), assuming (2.3) is true.

B.1 Implications of (2.3)

We record properties of λ' by comparing the coefficients in (2.3).

Linear form. The coefficients of the linear form in $\{f_i\}$ yield

$$\begin{aligned} \lambda'_{0,\bullet} - \lambda'_{\bullet,0} &= 1, \\ \lambda'_{i,\bullet} - \lambda'_{\bullet,i} &= 0\,, \quad i \in \{1, \dots, n-1, *\}, \\ \lambda'_{n,\bullet} - \lambda'_{\bullet,n} &= -1. \end{aligned} \tag{B.1}$$

Quadratic form. The coefficients of the quadratic form in $x_0 - x_*, \{g_i\}$ yield

$$\widehat{\lambda}' + \widetilde{H}\widetilde{\lambda}' + (\widetilde{H}\widetilde{\lambda}')^T = \mathbf{0}_{n\times n}, \tag{B.2}$$

$$\begin{bmatrix} \lambda'_{0,n} + \lambda'_{n,0} \\ \vdots \\ \lambda'_{n-1,n} + \lambda'_{n,n-1} \end{bmatrix} + \widetilde{H} \begin{bmatrix} \lambda'_{1,n} \\ \vdots \\ -\lambda'_{\bullet,n} \end{bmatrix} = \mathbf{0}_n, \tag{B.3}$$

$$\lambda'_{\bullet,n} + \lambda'_{n,\bullet} = R'_n. \tag{B.4}$$

B.2 Verification of (2.4) and positive semidefiniteness

Now, we compare the coefficients in (2.4). The derivation for each case is very similar to that in Appendix A.2 and thus we focus on the equivalent forms presented in matrix.

Linear form. The coefficients for $\{f_i\}$ are from (2.3). For the coefficients for $\{h_i\}$, it suffices to show that

$$\widetilde{\mu}'\mathbf{1}_n = -e_n\,.$$

From $\widetilde{\mu}' = \widetilde{\lambda}' + \widetilde{H}^{-1}\widehat{\lambda}'$ and $\widetilde{\lambda}'\mathbf{1} = \begin{bmatrix} -\lambda'_{1,n} \\ \vdots \\ -\lambda'_{n-1,n} \\ \lambda'_{\bullet,n} - 1 \end{bmatrix}$ (from (B.1)), it suffices to show that

$$\widetilde{H}^{-1}\widehat{\lambda}'\mathbf{1}_n = \begin{bmatrix} \lambda'_{1,n} \\ \vdots \\ \lambda'_{n-1,n} \\ -\lambda'_{\bullet,n} \end{bmatrix} \Leftrightarrow \widehat{\lambda}'\mathbf{1}_n = \widetilde{H} \begin{bmatrix} \lambda'_{1,n} \\ \vdots \\ \lambda'_{n-1,n} \\ -\lambda'_{\bullet,n} \end{bmatrix},$$

which holds from (B.3).

Quadratic form. By using the pseudo-co-coercivities defined in Appendix A.2, (2.4) given (2.3) is equivalent to

$$\begin{aligned}\sum_{i,j} \mu'_{i,j} Q^h_{ij} = \frac{R'_n}{2}\|g_n\|^2 - \frac{R'_n}{2}(1-\xi')\|g_n + s_n\|^2 + h_0 - h_* \\ - \sum_{i,j} \lambda'_{i,j}\left(\langle s_{j+1}, x_i - x_j\rangle + \frac{1}{2}\|s_{i+1} - s_{j+1}\|^2 + \langle g_i - g_j, s_{i+1} - s_{j+1}\rangle\right) \\ - \frac{1}{2}\operatorname{Tr}(V'S'(V')^T)\,.\end{aligned} \tag{B.5}$$

Note that the term $-\frac{R'_n}{2}(1-\xi')\|g_n + s_n\|^2$ in (B.5) is cancelled by the summand $-R'_n(1-\xi')(e_1 + e_{n+1})(e_1 + e_{n+1})^T$ in S' from (3.5). Also, for the types of vectors defined in (A.10), we only need to consider type g and s as we do not use the index $*$.

Coefficients of (g, g). These are directly from (2.3).

Coefficients of (g, s). The coefficients of $\langle g_i, s_j\rangle$ for $0 \leqslant i \leqslant n-1, 1 \leqslant j \leqslant n$ being matched is equivalent to

$$\widetilde{H}\widetilde{\mu}' = \widetilde{H}\widetilde{\lambda}' + \widehat{\lambda}'\,,$$

which holds from (B.3) and $\widetilde{H}\widetilde{\mu}' = -(\widetilde{H}\widetilde{\lambda})^T$ from (3.4).

The coefficient of $\langle g_n, s_j\rangle$ for $1 \leqslant j \leqslant n$ on the left hand side of (B.5) is 0; for the right hand side it is $\lambda_{j-1,n} + \lambda_{n,j-1} - v'_j = 0$ from (3.5).

Coefficients of (s, s). By considering the coefficients of $\langle s_i, s_j\rangle$ for $1 \leqslant i \leqslant j \leqslant n$, it suffices to show

$$\widetilde{H}\widetilde{\mu}' + (\widetilde{H}\widetilde{\mu}')^T = \widetilde{H}\widetilde{\lambda}' + (\widetilde{H}\widetilde{\lambda}')^T + 2\widehat{\lambda}'\,,$$

which follows from the corresponding result for the coefficients of (g, s).

Structure of S' and diagonal dominance. Recall the definition of S' from (3.5):

$$S' = \begin{bmatrix} R'_n & (v')^T \\ v' & -\widehat{\lambda}' \end{bmatrix} - R'_n(1-\xi')(e_1 + e_{n+1})(e_1 + e_{n+1})^T\,,$$

where $v' = [v'_1, \ldots, v'_n]$ with $v'_i = \lambda'_{i-1,n} + \lambda'_{n,i-1}$.

The first summand $\begin{bmatrix} R'_n & (v')^T \\ v' & -\widehat{\lambda}' \end{bmatrix}$ is diagonally dominant: in the first row, $v'_i \geqslant 0$ with $R'_n = \lambda'_{\bullet,n} + \lambda'_{n,\bullet} = \sum v'$ from (B.4); in the ith row for $2 \leqslant i \leqslant n+1$, the diagonal entry is $\lambda'_{\bullet,i-1} + \lambda'_{i-1,\bullet}$, and the absolute sum of the remaining entries is $v'_{i-1} + \sum_{k\notin\{i-1,n\}}(\lambda'_{k,i-1} + \lambda'_{i-1,k}) = \lambda'_{\bullet,i-1} + \lambda'_{i-1,\bullet}$.

Thus, if we choose $\xi' \in [0, 1)$ such that the $(1, n+1)$ entry of S' is nonnegative, i.e., $\lambda'_{n-1,n} + \lambda'_{n,n-1} - R'_n(1-\xi') \geqslant 0$, S' is diagonally dominant. In particular, we can choose $\xi' = 1 - \frac{\lambda'_{n-1,n} + \lambda'_{n,n-1}}{R'_n}$ (which is always in $[0, 1]$ from (B.4)).

C Deferred details for Section 5

C.1 Stepsize-accelerated proximal GD

Proof of Theorem 5.3. As in the case of objective function minimization, we first show that the following key lemma holds.

Lemma C.1. *For $k \in \mathbb{N}$ and $n = 2^k - 1$, let $\lambda' = \lambda'^{(k)}$. Then*

$$\frac{\lambda'_{i,j-1}}{\gamma'_{j-1}} - \frac{\lambda'_{i,j}}{\gamma'_j} \begin{cases} \geqslant 0 & 0 \leqslant i \leqslant j-2 \\ \leqslant 0 & i \geqslant j+1 \end{cases} \tag{C.1}$$

for all $1 \leqslant j \leqslant n$, where $[\gamma'_0, \ldots, \gamma'_{n-1}] := w^{(k)}$ and $\gamma'_n := 1$.

Proof. For $k = 2$ (note that no such entries are included for $k = 1$), all inequalities are straightforward (as either $\lambda'_{i,j-1}$ or $\lambda'_{i,j}$ is 0) except $\frac{\lambda'_{3,1}}{\gamma'_1} - \frac{\lambda'_{3,2}}{\gamma'_2} = \frac{\sqrt{2}}{1+\sqrt{2}} - \frac{2+\sqrt{2}}{\sqrt{2}} \leqslant 0$ and $\frac{\lambda'_{1,2}}{\gamma'_2} - \frac{\lambda'_{1,3}}{\gamma'_3} = \frac{\sqrt{2}}{\sqrt{2}} - 1 = 0$.

Assume that the result hold for k, and consider $k+1$. For $1 \leqslant j \leqslant n$, (C.1) holds by the induction hypothesis with $w^{(k+1)} = [w^{(k)}, \eta_k, \pi^{(k)}]$.

For $j = n+1$, we have $\lambda'_{i,j-1} = 0$ for all $n+1 \leqslant i \leqslant 2n$ and $\lambda'_{i,j} = 0$ for all $0 \leqslant i \leqslant n-1$. Thus, it only suffices to consider $i = 2n+1$ where

$$\frac{\lambda'_{2n+1,n}}{\gamma'_n} - \frac{\lambda'_{2n+1,n+1}}{\gamma'_{n+1}} = \frac{1}{\eta_k}\left(\frac{\tau_{k+1}}{2\rho^k} - 1\right) - \frac{1}{\sqrt{2}}\frac{\tau_{k+1}}{2\rho^{2k}}\sqrt{2} \leqslant 0 \;\Leftrightarrow\; \eta_k \geqslant \rho^k\left(1 - \frac{2\rho^k}{\tau_{k+1}}\right),$$

which holds from $\eta_k = 1 + \frac{\sqrt{\tau_k(\tau_k+8\rho^k)}-\tau_k}{4} = 1 + \rho^k - \frac{4\rho^k}{\sqrt{\tau_k(\tau_k+8\rho^k)}+\tau_k+4\rho^k} = 1 + \rho^k\left(1 - \frac{2\rho^k}{\tau_{k+1}}\right)$.

For $n+2 \leqslant j \leqslant 2n+1$, (C.1) holds from Lemma 4.4. □

By Lemma C.1, for $1 \leqslant j < i \leqslant n-1$,

$$\mu'_{i,j} = \gamma'_{j-1}\left(\frac{\sum_{l=0}^{j-1}\lambda'_{l,i-1}}{\gamma'_{i-1}} - \frac{\sum_{l=0}^{j-1}\lambda'_{l,i}}{\gamma'_i}\right) \geqslant 0\,,$$

and for $1 \leqslant i < j \leqslant n$,

$$\mu'_{i,j} = \gamma'_{j-1}\left(\frac{\sum_{l=j}^{n}\lambda'_{l,i}}{\gamma'_i} - \frac{\sum_{l=j}^{n}\lambda'_{l,i-1}}{\gamma'_{i-1}}\right) \geqslant 0\,.$$

For $i = n$ and $1 \leqslant j \leqslant n-1$,

$$\mu'_{n,j} = \frac{\gamma'_{j-1}}{\gamma'_{n-1}}\sum_{l=0}^{j-1}\lambda'_{l,n-1} \geqslant 0\,,$$

and for $i = 0$ and $1 \leqslant j \leqslant n$, $\begin{bmatrix}\mu'_{0,1} & \ldots & \mu'_{0,n}\end{bmatrix} = -\mathbf{1}_n^T\widetilde{\mu} = e_1$.

For item (ii), as in Theorem 2.8 we take $\xi' = 1 - \frac{\lambda'_{n,n-1}+\lambda'_{n-1,n}}{R'_n}$. For $k = 1$ (recall $n = 2^k - 1$) we have $\lambda'_{n,n-1} + \lambda'_{n-1,n} = 3$; for $k \geqslant 2$,

$$\begin{aligned}
\lambda'_{n,n-1} + \lambda'_{n-1,n} &= \frac{\tau_k}{\rho^{2(k-1)}}(\bar{\lambda}'^{(k-1)}_{(n-1)/2,(n-3)/2} + \bar{\lambda}'^{(k-1)}_{(n-3)/2,(n-1)/2}) + \frac{\tau_k}{2\rho^{2(k-1)}}\sqrt{2} \\
&= \frac{\tau_k}{\rho^{2(k-1)}}\left(\frac{1}{\sqrt{2}}(\rho^{2k-3} - 1) + \rho^{2k-3}\right) + \frac{\tau_k}{2\rho^{2(k-1)}}\sqrt{2} \\
&= \frac{\tau_k}{\sqrt{2}}\,,
\end{aligned}$$

where the formula for $a_k := \bar{\lambda}'^{(k-1)}_{(n-1)/2,(n-3)/2}, b_k := \bar{\lambda}'^{(k-1)}_{(n-3)/2,(n-1)/2}(k \geqslant 2)$ are obtained from respectively solving $a_2 = 1, a_{l+1} = \rho^2 a_l + \rho\sqrt{2}$ and $b_2 = \rho, b_{l+1} = \rho^2 b_l$ from the recursive definition in Proposition 5.2. Finally, from [25, Lemma 4], $\frac{1}{\tau_k} \sim \frac{1}{(\rho-1)(1+1/\sqrt{\rho})\rho^k} \sim \frac{\rho-\sqrt{\rho}}{2n^{\log_2\rho}}$. □

C.2 Proximal OGM-G

Before the proof of Theorem 5.5, we present the following lemma that will be used throughout our analysis. This quantitatively characterizes the sequences $\{\theta_i\}$ and $\{\varphi_i\}$.

Lemma C.2. *(a) (Difference of $\{\theta_i\}$). $\theta_{i+1} - \theta_i > \frac{1}{2}$ for all $0 \leqslant i \leqslant n-1$ and $\theta_{i+1} - \theta_i < \frac{3}{4}$ for all $0 \leqslant i \leqslant n-2$.*

(b) (Concentration of $\{\varphi_i\}$). $\varphi_{i+1} \geqslant \frac{4}{3}, \varphi_{i+2} \geqslant \frac{15}{11}, \varphi_{i+3} \geqslant \frac{7}{5}$ for all $i \geqslant 1$ and $\varphi_i \leqslant \frac{3}{2}, \varphi_{i+1} \leqslant 1 + \frac{1}{\sqrt{2}}$ for all $1 \leqslant i \leqslant n-1$.

Proof. (a) For $0 \leqslant i \leqslant n-2$, $\theta_i + \frac{1}{2} < \theta_{i+1} = \frac{1+\sqrt{1+4\theta_i^2}}{2} < \frac{1+\sqrt{4\theta_i^2+2\theta_i+\frac{1}{4}}}{2} = \theta_i + \frac{3}{4}$. $\theta_n - \theta_{n-1} > \frac{1}{2}$ directly follows.

(b) The upper bounds hold from $\theta_{i-1} \leqslant \theta_i \Rightarrow \varphi_i = 1 + \frac{\theta_{i-1}}{2\theta_i} \leqslant \frac{3}{2}$ for all $1 \leqslant i \leqslant n-1$ and $\sqrt{2}\theta_{n-1} \leqslant \theta_n \Rightarrow \varphi_n = 1 + \frac{\theta_{n-1}}{\theta_n} \leqslant 1 + \frac{1}{\sqrt{2}}$. For the lower bounds, it is straightforward to see that φ_i is increasing in i and $\varphi_n \geqslant \frac{3}{2}$. Thus, it suffices to show that $\varphi_2 \geqslant \frac{4}{3}, \varphi_3 \geqslant \frac{15}{11}, \varphi_4 \geqslant \frac{7}{5}$ (for $n \geqslant 5$).

These are respectively equivalent to

$$\theta_2 \leqslant \frac{3}{2}\theta_1, \quad \theta_3 \leqslant \frac{11}{8}\theta_2, \quad \theta_4 \leqslant \frac{5}{4}\theta_3 .$$

From (a), we know that $\theta_i \geqslant \frac{i}{2} + 1$ and in particular $\theta_1 \geqslant \frac{3}{2}, \theta_2 \geqslant 2$. Thus from (a), $\theta_2 \leqslant \theta_1 + \frac{3}{4} \leqslant \frac{3}{2}\theta_1$ and $\theta_3 \leqslant \theta_2 + \frac{3}{4} \leqslant \frac{11}{8}\theta_2$, proving the first two inequalities. The last inequality holds from $\theta_4 = \frac{1+\sqrt{1+4\theta_3^2}}{2} \leqslant \frac{1+\frac{5}{2}\theta_3-1}{2} = \frac{5}{4}\theta_3$, from $\theta_3 \geqslant \frac{5}{2}$.

□

Proof of Theorem 5.5. For item (i), we calculate $\widetilde{\mu}' = \widetilde{\lambda}' + \widetilde{H}^{-1}\widehat{\lambda}'$ and show the corresponding nonnegativity. For notational convenience, we will mainly consider these matrices scaled by $1/\theta_n^2$.

First, note that from

$$[(1/\theta_n^2)\widetilde{\lambda}']_{i,j} = \begin{cases} -\frac{1}{2\theta_{n-i-1}^2} & j = i+1 \\ \frac{1}{2\theta_{n-i-1}^2} & j = i+2 \\ \frac{1}{2\theta_{n-1}^2} - \frac{1}{\theta_n^2} & i = n, j = 1 \\ \frac{1}{2\theta_{n-j}^2} - \frac{1}{2\theta_{n-j+1}^2} & i = n, 2 \leqslant j \leqslant n \\ 0 & \text{else}\,, \end{cases}$$

its $(j-1, j)$ entries are negative, whereas the rest are nonnegative.

The main technical part is to calculate $\widetilde{H}^{-1}\widehat{\lambda}'$. First, note that the stepsize matrix H of OGM-G satisfies the following (the proof is analogous to that of Lemma 4.8 and hence is omitted):

$$\begin{aligned} H &= U(\theta_n, \ldots, \theta_1)^{-1} U(\varphi_n, \ldots, \varphi_1) \mathrm{diag}(2\theta_{n-1}, \ldots, 2\theta_0) \\ \Rightarrow \widetilde{H}^{-1} &= U_n^{-1} \mathrm{diag}(1/(2\theta_{n-1}), \ldots, 1/(2\theta_0)) U(\varphi_n, \ldots, \varphi_1)^{-1} U(\theta_n, \ldots, \theta_1) . \end{aligned}$$

Thus with $[(1/\theta_n^2)\widehat{\lambda}']_{i,j} = \begin{cases} \frac{1}{2\theta_{n-i}^2} & i = j-1 \\ \frac{1}{2\theta_{n-j}^2} & j = i-1 \\ \frac{1}{\theta_n^2} - \frac{1}{\theta_{n-1}^2} & i = j = 1 \\ -\frac{1}{\theta_{n-i}^2} & i = j \geqslant 2\,, \end{cases}$ we have

$$[(1/\theta_n^2)U(\theta_n,\ldots,\theta_1)\widehat{\lambda}']_{i,j} = \begin{cases} \frac{1}{2\theta_{n-j+1}^2} - \frac{1}{2\theta_{n-j}^2} & i+2 \leqslant j \leqslant n-1 \\ \frac{1}{2\theta_1^2} - \frac{1}{\theta_0^2} & i+2 \leqslant j = n \\ \frac{\theta_{n-i+1}}{2\theta_{n-i}^2} - \frac{1}{2\theta_{n-i-1}^2} & i+1 = j \leqslant n-1 \\ \frac{\theta_2}{2\theta_1^2} - \frac{1}{\theta_0^2} & i+1 = j = n \\ -\frac{\theta_n}{2\theta_{n-1}^2} & i = j = 1 \\ -\frac{2\theta_{n-i+1}-1}{2\theta_{n-i}^2} & 2 \leqslant i = j \leqslant n-1 \\ -\frac{\theta_1}{\theta_0^2} & i = j = n \\ \frac{1}{2\theta_{n-i+1}} & i = j+1 \\ 0 & \text{else}\,. \end{cases}$$

Similarly as before, we fix j and solve an equation with respect to the jth column. We divide this step into two cases.

Typical columns ($2 \leqslant j \leqslant n-1$). Here we consider a jth column of $(1/\theta_n^2)U(\theta_n,\ldots,\theta_1)\widehat{\lambda}'$, which is given as

$$\begin{bmatrix} \left(\frac{1}{2\theta_{n-j+1}^2} - \frac{1}{2\theta_{n-j}^2}\right)\mathbf{1}_{j-2} \\ \frac{\theta_{n-j+2}}{2\theta_{n-j+1}^2} - \frac{1}{2\theta_{n-j}^2} \\ -\frac{2\theta_{n-j+1}-1}{2\theta_{n-j}^2} \\ \frac{1}{2\theta_{n-j}} \\ \mathbf{0}_{n-j-1} \end{bmatrix}.$$

This column, after premultiplying $U(\varphi_n,\ldots,\varphi_1)^{-1}$, satisfies the following properties. The proof for the following lemma is similar to that of Lemma 4.9 and is deferred until later.

Lemma C.3 (Linear system for typical columns)**.** *For $2 \leqslant j \leqslant n-1$, let $x'(j)$ be the unique solution $x' = [x'_1,\ldots,x'_n]$ of*

$$U(\varphi_n,\ldots,\varphi_1)x' = 2\theta_{n-j+1}\theta_{n-j}^2 \begin{bmatrix} \left(\frac{1}{2\theta_{n-j+1}^2} - \frac{1}{2\theta_{n-j}^2}\right)\mathbf{1}_{j-2} \\ \frac{\theta_{n-j+2}}{2\theta_{n-j+1}^2} - \frac{1}{2\theta_{n-j}^2} \\ -\frac{2\theta_{n-j+1}-1}{2\theta_{n-j}^2} \\ \frac{1}{2\theta_{n-j}} \\ \mathbf{0}_{n-j-1} \end{bmatrix} = \begin{bmatrix} -\mathbf{1}_{j-2} \\ \theta_{n-j+2}\theta_{n-j+1} - \theta_{n-j+2} - \theta_{n-j+1} \\ -(2\theta_{n-j+1}-1)\theta_{n-j+1} \\ \theta_{n-j+1}\theta_{n-j} \\ \mathbf{0}_{n-j-1} \end{bmatrix}.$$

Then the following hold.

(a) $x'_n = \cdots = x'_{j+2} = 0, x'_{j+1} > 0, x'_j < 0$ and $x'_{j-1} > 0$.

(b) For $j \geqslant 3$, $x'_{j-2} < 0$ and $x'_k = \frac{\varphi_{n-k}-1}{\varphi_{n-k+1}}x'_{k+1}$ for all $1 \leqslant k \leqslant j-3$.

(c) $-\theta_{n-j+1}^2 + \frac{1}{2}x'_{j-1} - \frac{\theta_{n-j+1}}{2\theta_{n-j}}x'_j \geqslant 0$.

(d) For $j \geqslant 3$, $\theta_{n-j}^2 + \frac{\theta_{n-j+1}}{2\theta_{n-j+2}}x'_{j-2} - \frac{1}{2}x'_{j-1} \geqslant 0$.

Note that $[U_n^{-1}\mathrm{diag}(1/(2\theta_{n-1}),\ldots,1/(2\theta_0))]_{i,j} = \begin{cases} \frac{1}{2\theta_{n-i}} & i=j \\ -\frac{1}{2\theta_{n-j}} & j=i+1 \\ 0 & \text{else} \end{cases}$ holds. Thus with Lemma C.3 and $x' = x'(j)$, the (i,j) entry of $(1/\theta_n^2)\widetilde{H}^{-1}\widehat{\lambda}' = (1/\theta_n^2)U_n^{-1}\mathrm{diag}(1/(2\theta_{n-1}),\ldots,1/(2\theta_0))U_n(\varphi_n,\ldots,\varphi_1)^{-1}U(\theta_n,\ldots,\theta_1)\widehat{\lambda}'$ is given as

$$\begin{cases} \frac{1}{2\theta_{n-j+1}\theta_{n-j}^2}\left(\frac{1}{2\theta_{n-i}}x_i' - \frac{1}{2\theta_{n-i-1}}x_{i+1}'\right) & 1 \leqslant i \leqslant n-1 \\ \frac{1}{2\theta_{n-j+1}\theta_{n-j}^2}\left(\frac{1}{2\theta_0}x_n'\right) \geqslant 0 & i = n\,. \end{cases}$$

In particular, except for $i \in \{j-2, j\}$, the (i,j) entry of $(1/\theta_n^2)\widetilde{H}^{-1}\widehat{\lambda}'$ is nonnegative as

$$\begin{aligned} \frac{1}{2\theta_{n-i}}x_i' - \frac{1}{2\theta_{n-i-1}}x_{i+1}' \geqslant \left(\frac{1}{2\theta_{n-i}} - \frac{1}{2\theta_{n-i-1}}\right)x_{i+1}' \geqslant 0,\ 1 \leqslant i \leqslant j-3, \\ \frac{1}{2\theta_{n-j+1}}x_{j-1}' - \frac{1}{2\theta_{n-j}}x_j' \geqslant 0, \\ \frac{1}{2\theta_{n-j-1}}x_{j+1}' \geqslant 0. \end{aligned}$$

by Lemma C.3 (a) and (b).

Summing up what we have shown so far, for $(1/\theta_n^2)(\widetilde{\lambda} + \widetilde{H}^{-1}\widehat{\lambda})$ it suffices to show that the (i,j) entry is nonnegative for $i \in \{j-2, j-1\}$. For $i = j-2$, this entry is given as (from Lemma C.3 (c))

$$\frac{1}{2\theta_{n-j+1}^2} + \frac{1}{2\theta_{n-j+1}\theta_{n-j}^2}\left(\frac{1}{2\theta_{n-j+2}}x_{j-2}' - \frac{1}{2\theta_{n-j+1}}x_{j-1}'\right) = \frac{1}{2\theta_{n-j+1}^2\theta_{n-j}^2}\left(\theta_{n-j}^2 + \frac{\theta_{n-j+1}}{2\theta_{n-j+2}}x_{j-2}' - \frac{1}{2}x_{j-1}'\right) \geqslant 0\,,$$

and for $i = j-1$ this entry is given as (from Lemma C.3 (d))

$$-\frac{1}{2\theta_{n-j}^2} + \frac{1}{2\theta_{n-j+1}\theta_{n-j}^2}\left(\frac{1}{2\theta_{n-j+1}}x_{j-1}' - \frac{1}{2\theta_{n-j}}x_j'\right) = \frac{1}{2\theta_{n-j+1}^2\theta_{n-j}^2}\left(-\theta_{n-j+1}^2 + \frac{1}{2}x_{j-1}' - \frac{\theta_{n-j+1}}{2\theta_{n-j}}x_j'\right) \geqslant 0\,.$$

Special columns ($j \in \{1, n\}$). The first column of $(1/\theta_n^2)U(\theta_n,\ldots,\theta_1)\widehat{\lambda}'$ is given as

$$\begin{bmatrix} -\frac{\theta_n}{2\theta_{n-1}^2} \\ \frac{1}{2\theta_{n-1}} \\ \mathbf{0}_{n-2} \end{bmatrix},$$

from which it is trivial to check that the $(i,1)$ entry of $(1/\theta_n^2)\widetilde{H}^{-1}\widehat{\lambda}$ is nonnegative for all $2 \leqslant i \leqslant n$ (in particular, is equal to 0 for all $3 \leqslant i \leqslant n$).

The nth column of $(1/\theta_n^2)U(\theta_n,\ldots,\theta_1)\widehat{\lambda}'$ is given as

$$\begin{bmatrix} (\frac{1}{2\theta_1^2} - \frac{1}{\theta_0^2})\mathbf{1}_{n-2} \\ \frac{\theta_2}{2\theta_1^2} - \frac{1}{\theta_0^2} \\ -\frac{\theta_1}{\theta_0} \end{bmatrix}.$$

As in the case of typical columns, we solve the corresponding linear system as follows.

Lemma C.4 (Linear system for nth column)**.** *Let $y' = [y_1', \ldots, y_n']$ be the unique solution of*

$$U(\varphi_n,\ldots,\varphi_1)y' = 2\theta_1^2\begin{bmatrix} (\frac{1}{2\theta_1^2} - \frac{1}{\theta_0^2})\mathbf{1}_{n-2} \\ \frac{\theta_2}{2\theta_1^2} - \frac{1}{\theta_0^2} \\ -\frac{\theta_1}{\theta_0} \end{bmatrix} = \begin{bmatrix} -(2\theta_1^2 - 1)\mathbf{1}_{n-2} \\ \theta_2 - 2\theta_1^2 \\ -2\theta_1^3 \end{bmatrix}.$$

Then the following hold.

(a) $y'_n = -4\theta_1 < 0, y'_{n-1} = \frac{1}{\varphi_2}(\theta_2 + 2\theta_1 - 2) > 0, y'_{n-2} = -\frac{1}{\varphi_3}(\theta_2 - 1 - (\varphi_2 - 1)y'_{n-1}) < 0.$

(b) $y'_k = \frac{\varphi_{n-k}-1}{\varphi_{n-k+1}} y'_{k+1}$ *for all* $1 \leqslant k \leqslant n-3.$

For $1 \leqslant i \leqslant n-3$, the (i,n) entry of $(1/\theta_n^2)\widetilde{H}^{-1}\widehat{\lambda}'$ is nonnegative from $\frac{1}{2\theta_{n-i}}y'_i - \frac{1}{2\theta_{n-i-1}}y'_{i+1} \geqslant \left(\frac{1}{2\theta_{n-i}} - \frac{1}{2\theta_{n-i-1}}\right) y'_{i+1} \geqslant 0.$

The $(n-2,n)$ entry of $(1/\theta_n^2)(\widetilde{\lambda} + \widetilde{H}^{-1}\widehat{\lambda})$ is $\frac{1}{2\theta_1^2}\left(\frac{1}{2\theta_2}y'_{n-2} - \frac{1}{2\theta_1}y'_{n-1}\right) + \frac{1}{2\theta_1^2} \geqslant \frac{1}{2\theta_1^3}\left(\theta_1 - \frac{1}{2}y'_{n-1} + \frac{1}{2}y'_{n-2}\right)$ where

$$\begin{aligned}
\theta_1 - \frac{1}{2}y'_{n-1} + \frac{1}{2}y'_{n-2} &= \theta_1 - \frac{1}{2}y'_{n-1} - \frac{1}{2\varphi_3}(\theta_2 - 1 - (\varphi_2 - 1)y'_{n-1}) \\
&\geqslant \theta_1 - \frac{1}{2}y'_{n-1} - \frac{3}{8}(\theta_2 - 1 - (\varphi_2 - 1)y'_{n-1}) \\
&= \theta_1 - \frac{3}{8}(\theta_2 - 1) + \frac{3\varphi_2 - 7}{8}y'_{n-1} \\
&= \theta_1 - \frac{3}{8}(\theta_2 - 1) + \left(\frac{3}{8} - \frac{7}{8\varphi_2}\right)(\theta_2 + 2\theta_1 - 2) \\
&\geqslant \theta_1 - \frac{3}{8}(\theta_2 - 1) - \frac{9}{32}(\theta_2 + 2\theta_1 - 2) \\
&= \frac{7}{16}\theta_1 - \frac{21}{32}\theta_2 + \frac{15}{16} \\
&\geqslant \frac{7}{16}\theta_1 - \frac{21}{32}\left(\theta_1 + \frac{3}{4}\right) + \frac{15}{16} \\
&\geqslant \frac{7(2 - \theta_1)}{32} \geqslant 0\,,
\end{aligned}$$

and the $(n-1,n)$ entry of $(1/\theta_n^2)(\widetilde{\lambda} + \widetilde{H}^{-1}\widehat{\lambda})$ is $\frac{1}{2\theta_1^2}\left(\frac{1}{2\theta_1}y'_{n-1} - \frac{1}{2\theta_0}y'_n\right) - \frac{1}{2\theta_0^2} \geqslant \frac{1}{2\theta_1^3}\left(\frac{1}{2}y'_{n-1} - \frac{3}{4}y'_n - \theta_1^3\right)$ where

$$\begin{aligned}
-\theta_1^3 + \frac{1}{2}y'_{n-1} - \frac{3}{4}y'_n &= -\theta_1^3 + \frac{1}{2\varphi_2}(\theta_2 + 2\theta_1 - 2) + 3\theta_1 \\
&= \theta_1 - 1 + \frac{1}{2\varphi_2}(\theta_2 + 2\theta_1 - 2) \geqslant 0\,.
\end{aligned}$$

Nonnegativity of $\mu_{0,j}$. Finally, for $\begin{bmatrix}\mu_{0,1} & \dots & \mu_{0,n}\end{bmatrix} = -\mathbf{1}_n^T(\widetilde{\lambda}' + \widetilde{H}^{-1}\widehat{\lambda}')$, we have

$$-(1/\theta_n^2)\mathbf{1}_n^T\widetilde{\lambda}' = \begin{bmatrix}\frac{1}{\theta_n^2} - \frac{1}{2\theta_{n-1}^2} & \frac{1}{2\theta_{n-1}^2} & \mathbf{0}_{n-2}^T\end{bmatrix}$$

and

$$\begin{aligned}
-(1/\theta_n^2)\mathbf{1}_n^T(\widetilde{H}^{-1}\widehat{\lambda}') &= -\mathbf{1}_n^T U_n^{-1}\mathrm{diag}(1/(2\theta_{n-1}), \dots, 1/(2\theta_0))(1/\theta_n^2)U(\varphi_n, \dots, \varphi_1)^{-1}U(\theta_n, \dots, \theta_1)\widehat{\lambda}' \\
&= -\frac{1}{2\theta_{n-1}}e_1^T(1/\theta_n^2)U(\varphi_n, \dots, \varphi_1)^{-1}U(\theta_n, \dots, \theta_1)\widehat{\lambda}'\,.
\end{aligned}$$

We know that $(1,j)$ entry of $-(1/\theta_n^2)U(\varphi_n, \dots, \varphi_1)^{-1}U(\theta_n, \dots, \theta_1)\widehat{\lambda}'$ is nonnegative for all $j \neq 2$. Thus it suffices to verify the nonnegativity of the first two entries of $-(1/\theta_n^2)\mathbf{1}_n^T(\widetilde{\lambda}' + \widetilde{H}^{-1}\widehat{\lambda}')$. The first entry is $\frac{1}{\theta_n^2} - \frac{1}{2\theta_{n-1}^2} + \frac{4}{\theta_{n-1}^3\varphi_n}\left(\theta_n + \frac{\theta_{n-1}}{\varphi_{n-1}}\right) = \frac{1}{2\theta_{n-1}^3}\left(\frac{1}{2\varphi_n}\left(\theta_n + \frac{\theta_{n-1}}{\varphi_{n-1}}\right) - \theta_{n-1} + \frac{2\theta_{n-1}^3}{\theta_n^2}\right)$ where from Lemma C.2,

$$\begin{aligned}
\frac{1}{2\varphi_n}\left(\theta_n + \frac{\theta_{n-1}}{\varphi_{n-1}}\right) - \theta_{n-1} + \frac{2\theta_{n-1}^3}{\theta_n^2} &\geqslant \frac{1}{2+\sqrt{2}}\left(\theta_n + \frac{\theta_{n-1}}{\varphi_{n-1}}\right) - \theta_{n-1} + \frac{(\theta_n - 1)\theta_{n-1}}{\theta_n} \\
&\geqslant \frac{1}{2+\sqrt{2}}\left(\sqrt{2}\theta_{n-1} + \frac{2}{3}\theta_{n-1}\right) - \theta_{n-1} + \frac{1}{2}\theta_{n-1} \geqslant 0\,.
\end{aligned}$$

The second entry, for $n \geqslant 3$, is (letting $x' = x'(2)$) $-\frac{1}{4\theta_{n-1}^2\theta_{n-2}^2}x'_1 + \frac{1}{2\theta_{n-1}^2} = \frac{1}{2\theta_{n-1}^2\theta_{n-2}^2}(-\frac{1}{2}x'_1 + \theta_{n-2}^2)$ where

$$\begin{aligned}
-\frac{1}{2}x'_1 + \theta_{n-2}^2 &= -\frac{1}{2\varphi_n}(\theta_n(\theta_{n-1}-1) + 2\theta_{n-2}^2 + (\varphi_{n-1}-1)x'_2) + \theta_{n-2}^2 \\
&\geqslant -\frac{5}{16}\left(\theta_n(\theta_{n-1}-1) + 2(\theta_{n-1}^2 - \theta_{n-1}) - \frac{\varphi_{n-1}-1}{\varphi_{n-1}}(2\theta_{n-1}^2 - \theta_{n-1} + x'_3)\right) + (\theta_{n-1}^2 - \theta_{n-1}) \\
&\geqslant -\frac{5}{16}\left(\frac{10}{7}(\theta_{n-1}^2 - 1) + 2(\theta_{n-1}^2 - \theta_{n-1}) - \frac{1}{4}(2\theta_{n-1}^2 - \theta_{n-1}) - \frac{1}{4}x'_3\right) + (\theta_{n-1}^2 - \theta_{n-1}) \\
&\geqslant -\frac{5}{16}\left(\frac{10}{7}(\theta_{n-1}^2 - 1) + 2(\theta_{n-1}^2 - \theta_{n-1}) - \frac{1}{4}(2\theta_{n-1}^2 - \theta_{n-1}) - \frac{1}{4}\theta_{n-1}\right) + (\theta_{n-1}^2 - \theta_{n-1}) \\
&= \frac{1}{112}(9\theta_{n-1}^2 - 42\theta_{n-1} + 50) \geqslant 0\,.
\end{aligned}$$

Finally, the second entry for $n = 2$ is $-\frac{1}{4\theta_1^3}y'_1 + \frac{1}{2\theta_1^2} = \frac{1}{2\theta_1^3}\left(-\frac{1}{2}y'_1 + \theta_1\right)$ where

$$\begin{aligned}
-\frac{1}{2}y'_1 + \theta_1 &= -\frac{1}{2\varphi_2}(\theta_2 + 2\theta_1 - 2) + \theta_1 \\
&\geqslant -\frac{1}{3}(\theta_2 + 2\theta_1 - 2) + \theta_1 \\
&= \frac{1}{3}(\theta_1 - \theta_2 + 2) \\
&\geqslant \frac{1}{3}\left(\theta_1 - \frac{3}{2}\theta_1 - \frac{1}{2} + 2\right) \geqslant 0\,,
\end{aligned}$$

where in the second inequality we used $\theta_2 = \frac{1+\sqrt{1+8\theta_1^2}}{2} \leqslant \frac{1+3\theta_1}{2}$.

For item (ii), we can take $\xi' = 1 - \frac{\lambda'_{n-1,n}+\lambda'_{n,n-1}}{R'_n}$ where $\lambda'_{n-1,n} + \lambda'_{n,n-1} = \theta_n^2\left(\frac{1}{2\theta_0^2} + \frac{1}{2\theta_0^2} - \frac{1}{2\theta_1^2}\right) = \frac{\sqrt{5}+1}{4}\theta_n^2$ for $n \geqslant 2$. For $n = 1$, $\lambda'_{n-1,n} + \lambda'_{n,n-1} = \frac{1}{\theta_0^2} - \frac{1}{\theta_1^2} = \frac{3}{4}$.

□

We present the proofs of the technical lemmas within the previous parts as follows.

Proof of Lemma C.3. For notational convenience, let $i := n - j$ (which implies $1 \leqslant i \leqslant n-2$).

(a) Since $U(\varphi_n, \ldots, \varphi_1)$ is upper triangular, we can solve the equation iteratively starting from x'_n, which yields $x'_n = \cdots = x'_{j+2} = 0$. Also,

$$\begin{aligned}
x'_{j+1} &= \frac{\theta_{i+1}\theta_i}{\varphi_i} > 0, \\
x'_j &= -\frac{1}{\varphi_{i+1}}(2\theta_{i+1}^2 - \theta_{i+1} + x'_{j+1}) < 0, \\
x'_{j-1} &= \frac{1}{\varphi_{i+2}}(\theta_{i+2}\theta_{i+1} - \theta_{i+2} - \theta_{i+1} - x'_j - x'_{j+1}) \\
&= \frac{1}{\varphi_{i+2}}(\theta_{i+2}\theta_{i+1} - \theta_{i+2} - \theta_{i+1} + (2\theta_{i+1}^2 - \theta_{i+1}) + (\varphi_{i+1} - 1)x'_j) \\
&= \frac{1}{\varphi_{i+2}}(\theta_{i+2}(\theta_{i+1} - 1) + 2\theta_i^2 + (\varphi_{i+1} - 1)x'_j),
\end{aligned}$$

where from $(\varphi_{i+1}-1)x'_j = -\frac{\varphi_{i+1}-1}{\varphi_{i+1}}(2\theta_{i+1}^2 - \theta_{i+1} + x_{j+1}) > -\frac{1}{2}(2\theta_{i+1}^2 - \theta_{i+1} + x'_{j+1}) > -\frac{1}{2}(3\theta_{i+1}^2 - \theta_{i+1})$, $x'_{j-1} > \frac{1}{\varphi_{i+2}}\left(\left(\theta_{i+1} + \frac{1}{2}\right)(\theta_{i+1} - 1) + 2(\theta_{i+1}^2 - \theta_{i+1}) - \frac{1}{2}(3\theta_{i+1}^2 - \theta_{i+1})\right) = \frac{1}{2\varphi_{i+2}}(3\theta_{i+1}^2 - 4\theta_{i+1} - 1) > 0$.

(b) Solving the equation yields

$$x'_k = \frac{1}{\varphi_{n+1-k}}\left(-1 - \sum_{l=k+1}^{j+1} x'_l\right),\ k = j-2, \ldots, 1\,,$$

which implies $x'_k = \frac{1}{\varphi_{n+1-k}}(-x'_{k+1} + \varphi_{n-k}x'_{k+1}) = \frac{\varphi_{n-k}-1}{\varphi_{n+1-k}}x'_{k+1}$ for all $1 \leqslant k \leqslant j-3$. From $x'_{j-2} = \frac{1}{\varphi_{i+3}}(-1 - x'_{j-1} - x'_j - x'_{j+1})$, it suffices to show that $x'_{j-1} + x'_j + x'_{j+1} > -1$, which is equivalent to

$$\frac{1}{\varphi_{i+2}}(\theta_{i+2}\theta_{i+1} - \theta_{i+2} - \theta_{i+1}) - \left(1 - \frac{1}{\varphi_{i+2}}\right)\frac{1}{\varphi_{i+1}}(2\theta_{i+1}^2 - \theta_{i+1}) + \left(1 - \frac{1}{\varphi_{i+2}}\right)\left(1 - \frac{1}{\varphi_{i+1}}\right)\frac{1}{\varphi_i}\theta_{i+1}\theta_i > -1\,.$$

By Lemma C.2, $\frac{1}{\varphi_i} \geqslant \frac{2}{3}$ and $\frac{2}{3} \leqslant \frac{1}{\varphi_{i+1}}, \frac{1}{\varphi_{i+2}} \leqslant \frac{3}{4}$ (note that here, $i+2 < n$ as $j \geqslant 3$). Thus the left hand side is lower bounded by

$$\begin{aligned}
&\frac{2}{3}(\theta_{i+2}\theta_{i+1} - \theta_{i+2} - \theta_{i+1}) - \left(1 - \frac{2}{3}\right)\frac{3}{4}(2\theta_{i+1}^2 - \theta_{i+1}) + \left(1 - \frac{3}{4}\right)^2 \frac{2}{3}\theta_{i+1}\theta_i \\
&= \frac{1}{24}(16\theta_{i+2}(\theta_{i+1} - 1) - 12\theta_{i+1}^2 - 10\theta_{i+1} + \theta_{i+1}\theta_i) \\
&\geqslant \frac{1}{24}((16\theta_{i+1} + 8)(\theta_{i+1} - 1) - 12\theta_{i+1}^2 - 10\theta_{i+1} + \theta_i^2) \\
&= \frac{1}{24}(4\theta_{i+1}^2 - 18\theta_{i+1} + (\theta_{i+1}^2 - \theta_{i+1})) \\
&= \frac{1}{24}(5\theta_{i+1}^2 - 19\theta_{i+1}) > -1\,,
\end{aligned}$$

as desired.

(c) First, note that $x'_{j-1} > 0$ and $x'_j < 0$ from (a). Thus

$$\begin{aligned}
-\theta_{i+1}^2 + \frac{1}{2}x'_{j-1} - \frac{\theta_{i+1}}{2\theta_i}x'_j &\geqslant -\theta_{i+1}^2 + \frac{1}{2}(x'_{j-1} - x'_j) \\
&= -\theta_{i+1}^2 + \frac{1}{2}\frac{1}{\varphi_{i+2}}(\theta_{i+2}(\theta_{i+1} - 1) + 2\theta_i^2 + (\varphi_{i+1} - 1)x'_j) - \frac{1}{2}x'_j\,.
\end{aligned}$$

Consider first the case where $i \neq n-2$. Then $\frac{1}{\varphi_{i+2}} \geqslant \frac{2}{3}$, implying

$$\begin{aligned}
&-\theta_{i+1}^2 + \frac{1}{2}\frac{1}{\varphi_{i+2}}(\theta_{i+2}(\theta_{i+1} - 1) + 2\theta_i^2 + (\varphi_{i+1} - 1)x'_j) - \frac{1}{2}x'_j \\
&\geqslant -\theta_{i+1}^2 + \frac{1}{3}\left(\left(\theta_{i+1} + \frac{1}{2}\right)(\theta_{i+1} - 1) + 2(\theta_{i+1}^2 - \theta_{i+1}) + \frac{1}{2}x'_j\right) - \frac{1}{2}x'_j \\
&= -\frac{5}{6}\theta_{i+1} - \frac{1}{6} - \frac{1}{3}x'_j\,.
\end{aligned}$$

Since $-x'_j \geqslant \frac{2}{3}(2\theta_{i+1}^2 - \theta_{i+1} + x'_{j+1}) \geqslant \frac{2}{3}\left(2\theta_{i+1}^2 - \theta_{i+1} + \frac{2}{3}\theta_{i+1}\theta_i\right) \geqslant \frac{4}{3}\theta_{i+1}^2$ (from $\theta_i \geqslant \theta_1 \geqslant \frac{3}{2}$), the final term is lower bounded by $\frac{4}{9}\theta_{i+1}^2 - \frac{5}{6}\theta_{i+1} - \frac{1}{6} \geqslant 0$ (from $\theta_{i+1} \geqslant \theta_2 \geqslant 1 + \frac{\sqrt{5}}{2}$).

If $i = n-2$, then $\frac{1}{\varphi_{i+2}} \geqslant 2 - \sqrt{2}$ and $\theta_{i+2} \geqslant \sqrt{2}\theta_{i+1}$, implying

$$\begin{aligned}
&-\theta_{i+1}^2 + \frac{1}{2}\frac{1}{\varphi_{i+2}}(\theta_{i+2}(\theta_{i+1} - 1) + 2\theta_i^2 + (\varphi_{i+1} - 1)x'_j) - \frac{1}{2}x'_j \\
&\geqslant -\theta_{i+1}^2 + \frac{1}{2+\sqrt{2}}\left(\sqrt{2}\theta_{i+1}(\theta_{i+1} - 1) + 2(\theta_{i+1}^2 - \theta_{i+1}) + \frac{1}{2}x'_j\right) - \frac{1}{2}x'_j \\
&= -\theta_{i+1} - \frac{1}{2\sqrt{2}}x'_j \\
&\geqslant \frac{\sqrt{2}}{3}\theta_{i+1}^2 - \theta_{i+1} \geqslant 0\,,
\end{aligned}$$

where the last inequality is from $\theta_{i+1} \geqslant \theta_2 = \frac{1+\sqrt{7+2\sqrt{5}}}{2}$.

(d) First, note that $x'_{j-2} < 0$ and $x'_{j-1} > 0$ from (a) and (b). Thus

$$\begin{aligned}\theta_i^2 + \frac{\theta_{i+1}}{2\theta_{i+2}} x'_{j-2} - \frac{1}{2} x'_{j-1} &\geqslant \theta_i^2 + \frac{1}{2}(x'_{j-2} - x'_{j-1}) \\ &= \theta_i^2 + \frac{1}{2\varphi_{i+3}}(-1 - x'_{j-1} - x'_j - x'_{j+1}) - \frac{1}{2} x'_{j-1}\,.\end{aligned}$$

From Lemma C.2, this is lower bounded by

$$\begin{aligned}&\theta_i^2 + \frac{5}{14}(-1 - x'_{j-1} - x'_j - x'_{j+1}) - \frac{1}{2} x'_{j-1} \\ &= \theta_i^2 - \frac{5}{14} - \frac{6}{7} x'_{j-1} - \frac{5}{14}(\theta_{i+2}\theta_{i+1} - \theta_{i+2} - \theta_{i+1} - \varphi_{i+2} x'_{j-1}) \\ &= \theta_i^2 - \frac{5}{14}(\theta_{i+2} - 1)(\theta_{i+1} - 1) + \frac{5\varphi_{i+2} - 12}{14} x'_{j-1} \\ &\geqslant \theta_i^2 - \frac{5}{14}(\theta_{i+2} - 1)(\theta_{i+1} - 1) - \frac{19}{70}(\theta_{i+2}(\theta_{i+1} - 1) + 2\theta_i^2 + (\varphi_{i+1} - 1)x'_j) \\ &= \frac{16}{35}\theta_i^2 - \frac{5}{14}(\theta_{i+2} - 1)(\theta_{i+1} - 1) - \frac{19}{70}\theta_{i+2}(\theta_{i+1} - 1) + \frac{19}{70}\frac{\varphi_{i+1} - 1}{\varphi_{i+1}}(2\theta_{i+1}^2 - \theta_{i+1} + x'_{j+1}) \\ &\geqslant \frac{16}{35}\theta_i^2 - \frac{5}{14}(\theta_{i+2} - 1)(\theta_{i+1} - 1) - \frac{19}{70}\theta_{i+2}(\theta_{i+1} - 1) + \frac{19}{280}\left(2\theta_{i+1}^2 - \theta_{i+1} + \frac{2}{3}\theta_i^2\right) \\ &\geqslant \frac{16}{35}\theta_i^2 - \frac{5}{14}(\theta_{i+1} - \frac{1}{4})(\theta_{i+1} - 1) - \frac{19}{70}(\theta_{i+1} + \frac{3}{4})(\theta_{i+1} - 1) + \frac{19}{280}\left(2\theta_{i+1}^2 - \theta_{i+1} + \frac{2}{3}(\theta_{i+1}^2 - \theta_{i+1})\right) \\ &\geqslant \frac{1}{105}(\theta_{i+1}^2 - 6\theta_{i+1} + 12) \geqslant 0\,.\end{aligned}$$

□

Proof of Lemma C.4. (a) The formulae for y'_n, y'_{n-1}, y'_{n-2} can be derived straightforwardly by solving the equation, with $\theta_1^2 = \theta_1 + 1$ and $\varphi_1 = \frac{3+\sqrt{5}}{4} = \frac{\theta_1^2}{2}$. The only nontrivial inequality is $y'_{n-2} < 0$, which holds from $\theta_2 - 1 - (\varphi_2 - 1)y'_{n-1} = \theta_2 - 1 - \left(1 - \frac{1}{\varphi_2}\right)(\theta_2 + 2\theta_1 - 2) \geqslant \theta_2 - 1 - \frac{1}{3}(\theta_2 + 2\theta_1 - 2) = \frac{2}{3}\left(\theta_2 - \theta_1 - \frac{1}{2}\right) > 0$.

(b) The proof follows as in Lemma C.3 (b).

□