TRANSFORMERS WITH COT DISCOVER HEAVY-BALL ACCELERATION THROUGH TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern transformers often solve complex tasks by generating and refining intermediate reasoning steps. To develop a theoretical understanding of chain-of-thought (CoT) reasoning, a notable recent study by Huang et al. (2025) viewed transformers as algorithm executors, and showed that when transformers' CoT steps are supervised by a gradient-descent teacher, they can learn to implement gradient descent for in-context weight prediction in linear regression. In this paper, we study the same in-context weight prediction task but ask whether CoT reasoning can go beyond merely imitating the teacher algorithm. We show that, after initial supervision from a gradient-descent teacher, further trajectory-level supervision using the true regression weights leads the model to implement the Heavy-Ball method, thereby achieving acceleration over the original teacher. Notably, this accelerated algorithm emerges without supervision from any Heavy-Ball teacher. Our results provide a concrete theoretical example in which transformer reasoning can discover new algorithms through training, rather than merely imitate the algorithm used for supervision.

1 INTRODUCTION

Transformers have demonstrated remarkable success across a wide range of applications, including natural language processing (Vaswani et al., 2017) and computer vision (Dosovitskiy, 2020). These models have also exhibited complex reasoning abilities, particularly in mathematical problem solving and coding (Chowdhery et al., 2023). Chain-of-thought (CoT) reasoning has emerged as a key approach for enhancing such reasoning capabilities, enabling models to generate intermediate reasoning steps before producing the final answer (Wei et al., 2022; Kojima et al., 2022).

To theoretically understand the capabilities of transformers, recent works have viewed transformers as "algorithm executors." In particular, transformers have been shown to implement optimization algorithms for in-context learning tasks, where the context is modeled as the training dataset for a learning problem. Akyürek et al. (2022); Von Oswald et al. (2023) empirically demonstrated that transformers can perform in-context linear regression via gradient descent. Zhang et al. (2024) theoretically showed that one-layer transformers can be trained to perform in-context linear regression, while Ahn et al. (2023) showed that deep transformers can be trained to implement preconditioned gradient descent. Guo et al. (2024); Bai et al. (2023) further demonstrated the in-context learning capabilities of transformers by constructing multi-head ReLU transformers that implement a variety of learning algorithms. Gatmiry et al. (2024) showed that looped transformers can implement gradient descent for in-context linear regression. More recently, Huang et al. (2025) showed that, when supervised by a gradient

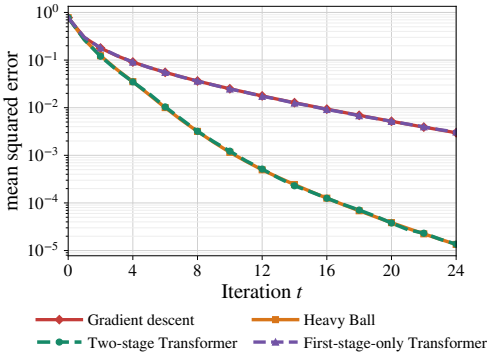


Figure 1: Comparison of gradient descent, Heavy-Ball, a transformer trained solely based on gradient-descent supervision, and a transformer trained first using the gradient-descent teacher and then using ground-truth regression weights.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

descent teacher, transformers can learn to use CoT reasoning for in-context weight prediction, with each reasoning step implementing one gradient descent update.

In this work, we consider a setting similar to that of Huang et al. (2025), where transformers use CoT to perform in-context weight prediction. However, rather than asking whether transformers can mimic a gradient-descent teacher, we ask a different question: *Can transformers with CoT go beyond mimicking the teacher algorithm and discover new algorithms through training?* To investigate this question, we consider a two-stage training procedure. In Stage 1, we supervise the transformer's CoT trajectory using a gradient descent teacher, following the setup of Huang et al. (2025). In Stage 2, we replace the teacher supervision with supervision from the true regression weights. We find that, when trained via this two-stage procedure, transformers with CoT can learn the Heavy-Ball method, achieving acceleration over the gradient descent teacher in Stage 1 (see Figure 1). We also establish rigorous theoretical guarantees to further demonstrate and explain this phenomenon.

The main contributions of our work are as follows:

- We identify an interesting phenomenon: as illustrated in Figure 1, a transformer trained solely with gradient-descent supervision learns to implement multi-step gradient descent, consistent with previous results (Huang et al., 2025). However, when further trained under trajectory supervision of the ground-truth regression weights, the transformer surprisingly learns to generate CoT trajectories that approximate the Heavy-Ball method, achieving acceleration. Notably, the Heavy-Ball method emerges without supervision from a Heavy-Ball teacher, demonstrating the capability of CoT to discover "new" algorithms unseen during training.
- We provide a theoretical explanation for how transformers with CoT learn the heavy-ball method. We show that, despite the nonconvex loss landscape, both stages of training succeed, with the corresponding objective functions attaining near-optimal values. We further identify the algorithm implemented by the trained transformer after the two-stage procedure and prove that its learned CoT iterates approximate the Heavy-Ball method. This establishes that the model not only optimizes the prescribed objectives, but also internalizes an accelerated optimization procedure. Our result therefore gives a rigorous demonstration of how transformers with CoT can go beyond mimicking a gradient descent teacher and discover a faster algorithm through training.
- The setting in this paper goes beyond classic in-context learning formulations. First, to distinguish the Heavy-Ball method from gradient descent, we consider a setting with a Marchenko–Pastur spectral limit, in contrast to prior settings with well-concentrated sample covariance matrices (Zhang et al., 2024; Huang et al., 2025). Second, the second stage training is without teacher forcing, and the loss lacks a clean analytical form due to the autoregressive nature of the outputs. To overcome these difficulties, we introduce auxiliary parameters to establish training convergence and develop recursively defined matrix polynomials to analyze the second-stage loss. The matrix polynomials also serve as a central tool for characterizing the connection between the learned algorithm and the Heavy-Ball method. These new proof techniques may be of independent interest.

2 Preliminaries

In this section, we introduce the in-context weight prediction task, the setting of one-layer linear self-attention with CoT, and the two-stage training procedure considered in this work.

2.1 In-Context Weight Prediction with CoT

In-context weight prediction. A line of recent works (Ahn et al., 2023; Bai et al., 2023; Zhang et al., 2024) has studied in-context learning under a theoretical setting where the goal is to predict an unknown response y_{query}, given a query $\mathbf{x}_{\text{query}}$ and a collection of context data pairs $\{(\mathbf{x}_i, y_i)\}_{i\in[n]}$. Beyond directly outputting the query response, Huang et al. (2025); Zhang & Cao (2026) further studied in-context weight prediction, where the goal is to predict the regression model weights.

In our work, we study the same in-context weight prediction task as in Huang et al. (2025). Specifically, the context is given as a training data set $\{(\mathbf{x}_i, y_i)\}_{i\in[n]}$ of a linear regression task. We assume that each context is associated with a ground-truth linear regression vector $\mathbf{w}^\star$, randomly generated according to the standard Gaussian distribution $\mathbf{w}^\star \sim N(0, \mathbf{I}_d)$. Given $\mathbf{w}^\star$, for each $i \in [n]$, we independently generate $\mathbf{x}_i \sim N(0, \mathbf{I}_d)$ and set $y_i = \mathbf{x}_i^\top \mathbf{w}^\star$. In the in-context weight prediction task, independent copies of $(\{(\mathbf{x}_i, y_i)\}_{i\in[n]}, \mathbf{w}^\star)$ are available for training, and the goal is to obtain a model that can predict the corresponding $\mathbf{w}^\star$ based on new contexts.

Linear self-attention architecture. We consider a simplified linear self-attention (LSA) model by removing the softmax nonlinearity, merging the projection and value matrices into $\mathbf{V} \in \mathbb{R}^{d_e \times d_e}$, and merging the key and query matrices into $\mathbf{W} \in \mathbb{R}^{d_e \times d_e}$, which follows the settings in recent works (Zhang et al., 2024; Ahn et al., 2023; Gatmiry et al., 2024; Huang et al., 2025). Here d_e denotes the dimension of the token embeddings. For an input matrix $\mathbf{Z}$, we denote

$$F_{\mathrm{LSA}}(\mathbf{Z}; \mathbf{V}, \mathbf{W}) = \mathbf{Z} + \mathbf{V}\mathbf{Z} \cdot \mathbf{Z}^\top \mathbf{W}\mathbf{Z}. \tag{1}$$

Chain-of-thought reasoning. In CoT reasoning, a transformer model autoregressively generates reasoning steps before reaching the final result. For in-context weight prediction, we follow a similar setup to that considered in Huang et al. (2025), where the reasoning steps correspond to iterations of an optimization algorithm, and consider the initial input of the transformer model as follows:

$$\mathbf{Z}_0 := \begin{bmatrix} \mathbf{X} & \mathbf{0}_{d\times 1} \\ \mathbf{y} & 0 \\ \mathbf{0}_{d\times n} & \mathbf{w}_0 \\ \mathbf{0}_{1\times n} & 1 \\ \mathbf{0}_{T\times n} & \mathbf{e}_0 \end{bmatrix} \in \mathbb{R}^{d_e \times (n+1)}, \qquad \begin{aligned} \mathbf{X} &:= \frac{1}{\sqrt{n}} [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_n] \in \mathbb{R}^{d\times n}, \\ \mathbf{y} &:= \frac{1}{\sqrt{n}} [y_1 \quad y_2 \quad \cdots \quad y_n] \in \mathbb{R}^{1\times n}. \end{aligned} \tag{2}$$

where $\mathbf{w}_0$ is the initialization of the optimization algorithm. Following Huang et al. (2025), we assume $\mathbf{w}_0 = \mathbf{0}_d$, and also assume that the training data set $\{(\mathbf{x}_i, y_i)\}_{i\in[n]}$ and the weight vectors are encoded in separate dimensions in the token embeddings. However, different from Huang et al. (2025), we include factors $1/\sqrt{n}$ in the definitions of $\mathbf{X}$ and $\mathbf{y}$. This difference is minor, since Huang et al. (2025) instead include an additional factor of $1/n$ in their definition of linear attention. Consequently, the well-trained model analyzed in their work is equivalent to one in which the $1/n$ factor in linear attention is omitted and the factors of $1/\sqrt{n}$ are included in $\mathbf{X}$ and $\mathbf{y}$. A more important difference is that we introduce a one-hot positional encoding $\mathbf{e}_0 = [1, 0, \ldots, 0]^\top \in \mathbb{R}^T$ for the initialization token, where T denotes the maximum number of CoT steps. Notably, incorporating these positional encodings requires us to specify a maximum number of reasoning steps before defining the model. Accordingly, each token has total embedding dimension $d_e := 2d + T + 2$.

Now we can define the input of the t-th reasoning step recursively. Specifically, we denote the model prediction as $\mathbf{w}_{t+1} := F_{\mathrm{LSA}}(\mathbf{Z}_t; \mathbf{V}, \mathbf{W})_{[d+2:2d+1,-1]}$ and append it to $\mathbf{Z}_t$ so that

$$\mathbf{Z}_{t+1} = \left[\mathbf{Z}_t, (\mathbf{0}_d^\top, 0, \mathbf{w}_{t+1}^\top, 1, \mathbf{e}_{t+1}^\top)^\top\right], \quad 0 \le t < T-1,$$

where $\mathbf{e}_t = (\underbrace{0, \ldots, 0}_{t}, 1, \underbrace{0, \ldots, 0}_{T-t-1})^\top \in \mathbb{R}^T$ is the one-hot positional encoding.

2.2 Two-Stage Training

We consider a two-stage training framework. Stage 1 introduces supervision from a gradient-descent teacher, while in Stage 2, we supervise the CoT trajectory with the true regression weights.

Stage 1. In Stage 1, following Huang et al. (2025), we consider supervision from a gradient-descent teacher with teacher forcing. The input to the transformer at iteration t is given as follows:

$$\mathbf{Z}_t^{\mathrm{GD}} = \begin{bmatrix} \mathbf{X} & \mathbf{0}_{d\times 1} & \mathbf{0}_{d\times 1} & \cdots & \mathbf{0}_{d\times 1} \\ \mathbf{y} & 0 & 0 & \cdots & 0 \\ \mathbf{0}_{d\times n} & \mathbf{w}_0 & \mathbf{w}_1^{\mathrm{GD}} & \cdots & \mathbf{w}_t^{\mathrm{GD}} \\ \mathbf{0}_{1\times n} & 1 & 1 & \cdots & 1 \\ \mathbf{0}_{T\times n} & \mathbf{e}_0 & \mathbf{e}_1 & \cdots & \mathbf{e}_t \end{bmatrix} \in \mathbb{R}^{d_e \times (n+t+1)}, \qquad 0 \le t < T,$$

where, for $t \ge 1$, $\mathbf{w}_t^{\mathrm{GD}} = \mathbf{w}_{t-1}^{\mathrm{GD}} - \eta_0 \mathbf{X}(\mathbf{X}^\top \mathbf{w}_{t-1}^{\mathrm{GD}} - \mathbf{y}^\top)$ is the teacher's iterate after t gradient steps. The learning rate η_0 is fixed across all steps and tasks. Given $\mathbf{Z}_t^{\mathrm{GD}}$, we aim to train the transformer to predict the next gradient-descent iterate. We consider gradient flow over the population loss:

$$\mathcal{L}_{\mathrm{GD}}(\mathbf{V}, \mathbf{W}) = \frac{1}{2T}\mathbb{E}\left[\sum_{t=0}^{T-1} \left\| F_{\mathrm{LSA}}(\mathbf{Z}_t^{\mathrm{GD}}; \mathbf{V}, \mathbf{W})_{[d+2:2d+1,-1]} - \mathbf{w}_{t+1}^{\mathrm{GD}} \right\|^2\right]. \tag{3}$$

The gradient flow dynamics of the parameters are

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\tau} = -\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{GD}}(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} := (\mathbf{V}, \mathbf{W}). \tag{4}$$

We train the model until $\mathcal{L}_{\mathrm{GD}}$ is minimized below a threshold.

Stage 2. We then consider further training the transformer on ground-truth solutions $\mathbf{w}^\star$ without supervision from any classical algorithm. At the t-th iteration, the transformer makes predictions based on its own autoregressively generated sequence $\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t$. Therefore, we define

$$\mathbf{Z}_t = \begin{bmatrix} \mathbf{X} & \mathbf{0}_{d\times 1} & \mathbf{0}_{d\times 1} & \cdots & \mathbf{0}_{d\times 1} \\ \mathbf{y} & 0 & 0 & \cdots & 0 \\ \mathbf{0}_{d\times n} & \mathbf{w}_0 & \mathbf{w}_1 & \cdots & \mathbf{w}_t \\ \mathbf{0}_{1\times n} & 1 & 1 & \cdots & 1 \\ \mathbf{0}_{T\times n} & \mathbf{e}_0 & \mathbf{e}_1 & \cdots & \mathbf{e}_t \end{bmatrix} \in \mathbb{R}^{d_e\times(n+t+1)}, \qquad 0 \le t < T,$$

and define the training loss in Stage 2 as

$$\mathcal{L}(\mathbf{V}, \mathbf{W}) = \frac{1}{2T}\mathbb{E}\left[\sum_{t=0}^{T-1} \left\| F_{\mathrm{LSA}}(\mathbf{Z}_t; \mathbf{V}, \mathbf{W})_{[d+2:2d+1,-1]} - \mathbf{w}^\star \right\|^2\right]. \tag{5}$$

We use exact population gradient flow on the raw parameters:

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\tau} = -\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}). \tag{6}$$

Notably, without teacher forcing, Equation (5) does not have a clean analytic form because $\mathbf{Z}_t$ is generated autoregressively. During training, $\mathbf{w}_1, \ldots, \mathbf{w}_T$ can be very complex, which differs substantially from Stage 1, as well as the setting of Huang et al. (2025).

We note that Stage 1 is analogous to pretraining with next-token prediction and to supervised fine-tuning: the input contains teacher weights rather than the model's own predictions. The loss function in Stage 2 can be viewed as an objective resembling reinforcement learning post-training with discount factor 1. However, because our setting of in-context linear regression is simplified, these analogies serve only as high-level motivation for considering such a training procedure. They do not imply that our analysis directly yields conclusions about general pretraining or post-training.

3 Main Result

In this section, we prove the convergence of the two-stage training procedure and show that the final solution approximates the time-varying Heavy-Ball method, defined as follows.

Let $\{(\alpha_t, \beta_t)\}_{t=0}^{T-1}$ be any sequence with $\beta_0 = 0$. Initialize $\mathbf{w}_0^{\mathrm{HB}} = \mathbf{0}_d$ and for $1 \le t < T$, define

$$\mathbf{w}_1^{\mathrm{HB}} = \alpha_0 \mathbf{X}\mathbf{y}^\top, \quad \mathbf{w}_{t+1}^{\mathrm{HB}} = \mathbf{w}_t^{\mathrm{HB}} - \alpha_t(\mathbf{X}\mathbf{X}^\top \mathbf{w}_t^{\mathrm{HB}} - \mathbf{X}\mathbf{y}^\top) + \beta_t(\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}_{t-1}^{\mathrm{HB}}). \tag{7}$$

We consider the best coefficients under the same loss objective:

$$\{(\alpha_t^\star, \beta_t^\star)\}_{t=0}^{T-1} \in \operatorname*{arg\,min}_{\{(\alpha_t,\beta_t)\}_{t=0}^{T-1}, \beta_0=0} \frac{1}{2T}\sum_{t=1}^{T} \mathbb{E}\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star\|^2.$$

In the following, we introduce two technical assumptions. We first note that the transformer input (2) is given as a block matrix, motivating us to consider the weight matrices as block matrices:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} & \mathbf{V}_{14} & \mathbf{V}_{15} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \mathbf{V}_{23} & \mathbf{V}_{24} & \mathbf{V}_{25} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} & \mathbf{V}_{34} & \mathbf{V}_{35} \\ \mathbf{V}_{41} & \mathbf{V}_{42} & \mathbf{V}_{43} & \mathbf{V}_{44} & \mathbf{V}_{45} \\ \mathbf{V}_{51} & \mathbf{V}_{52} & \mathbf{V}_{53} & \mathbf{V}_{54} & \mathbf{V}_{55} \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \mathbf{W}_{13} & \mathbf{W}_{14} & \mathbf{W}_{15} \\ \mathbf{W}_{21} & \mathbf{W}_{22} & \mathbf{W}_{23} & \mathbf{W}_{24} & \mathbf{W}_{25} \\ \mathbf{W}_{31} & \mathbf{W}_{32} & \mathbf{W}_{33} & \mathbf{W}_{34} & \mathbf{W}_{35} \\ \mathbf{W}_{41} & \mathbf{W}_{42} & \mathbf{W}_{43} & \mathbf{W}_{44} & \mathbf{W}_{45} \\ \mathbf{W}_{51} & \mathbf{W}_{52} & \mathbf{W}_{53} & \mathbf{W}_{54} & \mathbf{W}_{55} \end{bmatrix}.$$

where the block row and column sizes are $(d, 1, d, 1, T)$. We make the following assumptions.

Assumption 3.1. *We assume that the initialization of the parameters satisfies*

$$\mathbf{V}(0) = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & v_{33}(0)\mathbf{I}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \qquad \mathbf{W}(0) = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & \sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{8}$$

Here $v_{33}(0)$ and σ are any constants satisfying $v_{33}(0) > 0$ and $\sigma < 0$, independent of d, n, T. Further, we fix $\mathbf{W}_{33} = \mathbf{0}_{d\times d}$ for all $\tau > 0$.

Assumption 3.1 is a technical assumption that poses constraints on the initialization and on freezing one of the blocks in the $\mathbf{W}$ matrix. We note that similar assumptions have been frequently considered in theoretical studies of the training dynamics of transformer models. For example, Huang et al. (2025); Wu et al. (2026a) considered almost zero initializations with certain non-zero blocks. Huang et al. (2025); Gatmiry et al. (2024) considered freezing certain blocks in the weight matrices.

We also make the following assumption specifying hyperparameter scalings considered in this paper.

Assumption 3.2. *Suppose that $T = \Theta((\log d)^{1/4})$, $\eta_0 \in [0.4, 0.9]$, $n = \Theta(d)$, and $d/n \le 0.1$.*

We emphasize that Assumption 3.2 introduces a key difference compared with related work (Huang et al., 2025), where the authors assumed that $n = \Theta(d \log^5 d) \gg d$, implying that the sample covariance matrix $\mathbf{S} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$ concentrates around $\mathbf{I}_d$. Notably, $\mathbf{S} \approx \mathbf{I}_d$ is a setting where the Heavy-Ball method and gradient descent achieve the same convergence rate. Hence, to convincingly demonstrate the learning of the Heavy-Ball method, we consider a different regime where n and d are proportional to each other. In this setting, the eigenvalue distribution of $\mathbf{S}$ is asymptotically characterized by the Marchenko-Pastur (MP) law (Marčenko & Pastur, 1967).

We now show that, under the initialization above and the two-stage training procedure, our transformer model discovers the Heavy-Ball method.

Theorem 3.1. *For any $0 < \epsilon < 1$, suppose that $n, d = \Omega(\exp(\log^4(1/\epsilon)))$. Further suppose that the training with gradient flow (4) in Stage 1 stops after $\tau_1 = d^{o(1)} + O(T \log d)$, and the follow-up Stage 2 training with gradient flow (6) runs for $\tau - \tau_1 = O(d^{1/\alpha + o(1)} \epsilon^{-1/\alpha})$, where $\alpha > 0$ is a constant independent of d, n, T. Then under Assumptions 3.1 and 3.2, the following results hold:*

- *The training loss in Stage 1 reaches $\mathcal{L}_{\mathrm{GD}} \le d^{-1/4}$.*
- *The training loss in Stage 2 reaches $\mathcal{L} - \min \mathcal{L} \le \epsilon$.*
- *Uniformly for $T/2 \le t < T$, it holds that*

$$\frac{\mathbb{E}\|\mathbf{w}_{t+1}(\tau) - \mathbf{w}_t^{\mathrm{HB}}\|^2}{\mathbb{E}\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star\|^2} \le \epsilon.$$

Here $\mathbf{w}_t^{\mathrm{HB}}$ denotes the t-th iterate of the Heavy-Ball method with parameters $\{(\alpha_t^\star, \beta_t^\star)\}_{t=0}^{T-1}$.

Theorem 3.1 shows that the training in both stage 1 and stage 2 converges, despite the nonconvex loss landscape. Moreover, the last conclusion of Theorem 3.1 shows that the transformer produces CoT steps whose $t+1$-th step is close to the t-th iteration of the Heavy-Ball method. This demonstrates that the algorithm learned by the transformer is essentially equivalent to the Heavy-Ball method. Notably, the two-stage training procedure never involves supervision from any Heavy-Ball teacher. Our result thus indicates that transformers with CoT are capable of discovering unseen algorithms through proper training, providing complementary findings to Huang et al. (2025) where the authors demonstrated the capability of transformers to recover the gradient-descent teacher algorithm.

Theorem 3.1 is under the assumption that $n, d = \Omega(\exp(\log^4(1/\epsilon)))$. This condition is super-polynomial but subexponential in $1/\epsilon$. While such a super-polynomial assumption is relatively strong, the last conclusion in Theorem 3.1 with $\epsilon = 0.5$ can already imply that $\mathbb{E}\|\mathbf{w}_{t+1}(\tau) - \mathbf{w}^\star\|^2 = O(\mathbb{E}\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star\|^2)$. Therefore, Theorem 3.1 demonstrates that transformers with CoT can discover an algorithm with an accelerated convergence rate under mild assumptions.

4 Proof Sketch

In this section, we outline the proof of Theorem 3.1. Since we consider the setting where $n = \Theta(d)$ and the eigenvalues of $\mathbf{S}$ follow the MP law, even in Stage 1, our analysis is significantly different from the prior work (Huang et al., 2025). Stage 2 is also more challenging without teacher forcing. To address these challenges, we develop a proof technique based on auxiliary parameters.

We first identify the parameter structure preserved throughout training.

Lemma 4.1. *Throughout the two-stage training, the parameters $\mathbf{V}(\tau), \mathbf{W}(\tau)$ retain the form*

$$\mathbf{V}(\tau) = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & 0 & \mathbf{0} \\ v_{31}(\tau)\mathbf{I}_d & \mathbf{0} & v_{33}(\tau)\mathbf{I}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{W}(\tau) = \begin{pmatrix} \mathbf{0} & \mathbf{0} & w_{13}(\tau)\mathbf{I}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & w_{24}(\tau) & \mathbf{w}_{25}(\tau) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & w_{44}(\tau) & \mathbf{w}_{45}(\tau) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{w}_{54}(\tau) & \mathbf{W}_{55}(\tau) \end{pmatrix}.$$

Lemma 4.1 shows that our parameters remain on a manifold throughout two-stage training. By this lemma, we can reduce the analysis of the full matrices $\mathbf{V}, \mathbf{W}$ to that of

$$\bar{\boldsymbol{\theta}}(\tau) = \big(\mathbf{V}_{31}(\tau), \mathbf{W}_{13}(\tau), \mathbf{V}_{33}(\tau), w_{24}(\tau), \mathbf{w}_{25}(\tau), w_{44}(\tau), \mathbf{w}_{45}(\tau), \mathbf{w}_{54}(\tau), \mathbf{W}_{55}(\tau)\big).$$

This conclusion also allows us to establish the following lemma.

Lemma 4.2. *Throughout training in Stages 1 and 2, the t-th CoT iterate $\mathbf{w}_t(\tau)$ has the form $\mathbf{w}_t(\tau) = p_{t,\bar{\boldsymbol{\theta}}(\tau)}(\mathbf{S})\mathbf{w}^\star$, where $p_{t,\bar{\boldsymbol{\theta}}(\tau)}$ is a polynomial of degree at most t with zero constant term.*

Lemma 4.2 is central to our analysis. Since we consider the challenging setting where $n = \Theta(d)$, the sample covariance matrix $\mathbf{S}$ cannot be approximated by the identity matrix, and the study of the training dynamics requires precise characterization of their dependence on $\mathbf{S}$. Lemma 4.2 provides this characterization. Recall that the losses in both stages are defined by the distance from $\mathbf{w}_t$ to $\mathbf{w}^\star$ or $\mathbf{w}_t^{\mathrm{GD}}$. By Lemma 4.2, we can express the losses in the two stages as functions of matrix polynomials in $\mathbf{S}$. We now give the two-stage analysis.

Stage 1. Thanks to Lemma 4.2, we can give the following decomposition of the training loss.

Lemma 4.3. *The population loss can be written as*

$$\mathcal{L}_{\mathrm{GD}}(\tau) = \mathbb{E}\left[\begin{array}{l} A(\mathbf{S})(\widetilde{\theta}_1(\tau) - \eta_0)^2 + B(\mathbf{S})(\widetilde{\theta}_2(\tau) + \eta_0)^2 \\ \quad + \Big\|\mathbf{C}_1(\mathbf{S})(\widetilde{\theta}_2(\tau) + \eta_0) + \mathbf{C}_2(\mathbf{S})\,\mathrm{vec}(\widetilde{\boldsymbol{\Theta}}_3(\tau) - \eta_0\mathbf{e}_1(\mathbf{1}_T - \mathbf{e}_0)^\mathsf{T})\Big\|^2 \end{array}\right],$$

where $A, B \in \mathbb{R}$, $\mathbf{C}_1 \in \mathbb{R}^{(T-1)d^2\times 1}$, $\mathbf{C}_2 \in \mathbb{R}^{(T-1)d^2\times T^2}$ are explicit functions of $\mathbf{S}$, and

$$\begin{aligned}
&\widetilde{\theta}_1 := v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0), \qquad \widetilde{\theta}_2 := v_{31}w_{13}, \qquad \widetilde{\boldsymbol{\Theta}}_3 \in \mathbb{R}^{T\times T},\\
&(\widetilde{\boldsymbol{\Theta}}_3)_{1t} := v_{31}\big(w_{24} + \mathbf{w}_{25}\mathbf{e}_t\big) + \eta_0 v_{33}\big(w_{44} + \mathbf{w}_{45}\mathbf{e}_t + \mathbf{e}_1^\mathsf{T}\mathbf{w}_{54} + \mathbf{e}_1^\mathsf{T}\mathbf{W}_{55}\mathbf{e}_t\big), \quad 1 \le t < T,\\
&(\widetilde{\boldsymbol{\Theta}}_3)_{\ell t} := v_{33}\big(w_{44} + \mathbf{w}_{45}\mathbf{e}_t + \mathbf{e}_\ell^\mathsf{T}\mathbf{w}_{54} + \mathbf{e}_\ell^\mathsf{T}\mathbf{W}_{55}\mathbf{e}_t\big), \qquad\qquad 2 \le \ell \le t < T.
\end{aligned}$$

Lemma 4.3 implies that, as gradient flow minimizes $\mathcal{L}_{\mathrm{GD}}(\tau)$, we have $\widetilde{\theta}_1(\tau) \to \eta_0$, $\widetilde{\theta}_2(\tau) \to -\eta_0$, and $\widetilde{\boldsymbol{\Theta}}_3(\tau) \to \eta_0\mathbf{e}_1(\mathbf{1}_T - \mathbf{e}_0)^\mathsf{T}$. This further motivates us to use $\widetilde{\theta}_1, \widetilde{\theta}_2, \widetilde{\boldsymbol{\Theta}}_3$ as auxiliary parameters to study convergence. We emphasize that the training dynamics we analyze are still gradient flow in the original parameters; the auxiliary parameters are only a tool for the proof.

The next lemma gives a PL inequality in these auxiliary parameters.

Lemma 4.4. *There exists a constant $c_1 > 0$ such that*

$$\Big|\partial_{\widetilde{\theta}_1(\tau)}\mathcal{L}_{\mathrm{GD}}(\tau)\Big|^2 + \frac{1}{d}\Big|\partial_{\widetilde{\theta}_2(\tau)}\mathcal{L}_{\mathrm{GD}}(\tau)\Big|^2 + \Big\|\nabla_{\widetilde{\boldsymbol{\Theta}}_3(\tau)}\mathcal{L}_{\mathrm{GD}}(\tau)\Big\|_F^2 \ge c_1\frac{d/n}{T}\mathcal{L}_{\mathrm{GD}}(\tau).$$

Notably, this inequality concerns auxiliary parameters, whereas convergence requires a PL inequality for the original parameters. Therefore, we define

$$\widetilde{\boldsymbol{\theta}}(\tau) := \big(\widetilde{\theta}_1(\tau), \widetilde{\theta}_2(\tau), \widetilde{\boldsymbol{\Theta}}_3(\tau)\big), \qquad \mathbf{J}_1(\bar{\boldsymbol{\theta}}(\tau)) := \frac{\partial\widetilde{\boldsymbol{\theta}}(\tau)}{\partial\bar{\boldsymbol{\theta}}(\tau)}.$$

Then we have the following lemma.

Lemma 4.5. *There is a time $\tau_0 = d^{o(1)}$ and a constant $c_2 > 0$ such that, for every $\tau \ge \tau_0$,*

$$\mathbf{J}_1(\bar{\boldsymbol{\theta}}(\tau))\mathbf{J}_1(\bar{\boldsymbol{\theta}}(\tau))^\mathsf{T} \succeq c_2\ \mathrm{diag}(1, d^{-1}, \mathbf{I}).$$

Note that the inequality in Lemma 4.5 holds only after a time $\tau_0 = d^{o(1)}$. This highlights the challenge of the problem: the PL condition is established only in the later phase of training.

By the chain rule, $\nabla_{\bar{\boldsymbol{\theta}}(\tau)}\mathcal{L}_{\mathrm{GD}}(\tau) = \mathbf{J}_1(\bar{\boldsymbol{\theta}}(\tau))^\mathsf{T}\nabla_{\widetilde{\boldsymbol{\theta}}(\tau)}\mathcal{L}_{\mathrm{GD}}(\tau)$. Combining Lemmas 4.4 and 4.5 therefore gives the PL inequality with respect to $\bar{\boldsymbol{\theta}}$. This further leads to the following inequality:

$$-\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{L}_{\mathrm{GD}}(\tau) = \Big\|\nabla_{\bar{\boldsymbol{\theta}}(\tau)}\mathcal{L}_{\mathrm{GD}}(\tau)\Big\|^2 \ge c_1c_2\frac{d/n}{T}\mathcal{L}_{\mathrm{GD}}(\tau), \qquad \tau \ge \tau_0.$$

Solving this inequality proves the first conclusion of Theorem 3.1: it takes a training time $\tau = d^{o(1)} + O((nT/d)\log d)$ to reach $\mathcal{L}_{\mathrm{GD}} \le d^{-1/4}$.

Stage 2. Our Stage 2 analysis has two parts: (i) convergence analysis and (ii) comparison with the Heavy-Ball method. We first discuss part (i) convergence analysis. We have the following lemma.

Lemma 4.6. *In Stage 2, the training loss can be written as*

$$\mathcal{L}(\tau) = \frac{1}{2T}\mathbb{E}\sum_{t=1}^{T}\|p_{t,\bar{\boldsymbol{\theta}}(\tau)}(\mathbf{S}) - \mathbf{I}_d\|_F^2,$$

where $p_{t,\bar{\boldsymbol{\theta}}(\tau)}(\mathbf{S})$ is a polynomial defined recursively by

$$p_{0,\bar{\boldsymbol{\theta}}(\tau)}(\mathbf{S}) = \mathbf{0}_{d\times d}, \qquad p_{1,\bar{\boldsymbol{\theta}}(\tau)}(\mathbf{S}) = \widetilde{\theta}_1(\tau)\mathbf{S},$$
$$p_{t+1,\bar{\boldsymbol{\theta}}(\tau)}(\mathbf{S}) = (\mathbf{I}_d + \widetilde{\theta}_2(\tau)\mathbf{S})p_{t,\bar{\boldsymbol{\theta}}(\tau)}(\mathbf{S}) + \sum_{1\le\ell\le t} p_{\ell,\bar{\boldsymbol{\theta}}(\tau)}(\mathbf{S})\,(\widetilde{\boldsymbol{\Theta}}_4(\tau))_{\ell t}, \quad 1\le t<T.$$

Here, $\widetilde{\theta}_1(\tau)$ and $\widetilde{\theta}_2(\tau)$ are defined in Lemma 4.3, and $\widetilde{\boldsymbol{\Theta}}_4 \in \mathbb{R}^{T\times T}$ is given as

$$(\widetilde{\boldsymbol{\Theta}}_4)_{1t} := \widetilde{\theta}_1^{-1}v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t) + v_{33}\big(w_{44} + \mathbf{w}_{45}\mathbf{e}_t + \mathbf{e}_1^\top\mathbf{w}_{54} + \mathbf{e}_1^\top\mathbf{W}_{55}\mathbf{e}_t\big), \quad 1\le t<T,$$
$$(\widetilde{\boldsymbol{\Theta}}_4)_{\ell t} := v_{33}\big(w_{44} + \mathbf{w}_{45}\mathbf{e}_t + \mathbf{e}_\ell^\top\mathbf{w}_{54} + \mathbf{e}_\ell^\top\mathbf{W}_{55}\mathbf{e}_t\big), \qquad 2\le\ell\le t<T.$$

Lemma 4.6 gives the loss function based on the matrix polynomials $p_{t,\bar{\boldsymbol{\theta}}(\tau)}(\mathbf{S})$. Compared with Lemma 4.3 in Stage 1, this result is significantly more complicated, as $p_{t,\bar{\boldsymbol{\theta}}(\tau)}(\mathbf{S})$ is defined recursively. This is due to the fact that the transformer takes its own outputs in the previous reasoning steps in Stage 2. Despite this more challenging setting, we can still show the PL inequality for these auxiliary parameters $\widetilde{\theta}_1, \widetilde{\theta}_2, \widetilde{\boldsymbol{\Theta}}_4$. Write $\mathcal{L}^*$ for the minimum of $\mathcal{L}$. Then we have the following result.

Lemma 4.7. *In Stage 2, there exists a constant $c_3 > 0$ such that the following PL inequality holds:*

$$\frac{1}{d}\left|\partial_{\widetilde{\theta}_1(\tau)}\mathcal{L}(\tau)\right|^2 + \frac{1}{d}\left|\partial_{\widetilde{\theta}_2(\tau)}\mathcal{L}(\tau)\right|^2 + \frac{1}{\sqrt{d}}\left\|\nabla_{\widetilde{\boldsymbol{\Theta}}_4(\tau)}\mathcal{L}(\tau)\right\|_F^2 \ge c_3\frac{d/n}{T}(\mathcal{L}(\tau) - \mathcal{L}^*).$$

Similar to Stage 1, we aim to translate the PL condition from the auxiliary parameters to $\bar{\boldsymbol{\theta}}$. Define

$$\widetilde{\boldsymbol{\theta}}'(\tau) := \big(\widetilde{\theta}_1(\tau), \widetilde{\theta}_2(\tau), \widetilde{\boldsymbol{\Theta}}_4(\tau)\big), \qquad \mathbf{J}_2(\bar{\boldsymbol{\theta}}(\tau)) := \frac{\partial\widetilde{\boldsymbol{\theta}}'(\tau)}{\partial\bar{\boldsymbol{\theta}}(\tau)}.$$

Then we have the following lemma.

Lemma 4.8. *There exists a constant $c_4 > 0$ such that for all $\tau \ge \tau_1$,*

$$\mathbf{J}_2(\bar{\boldsymbol{\theta}}(\tau))\mathbf{J}_2(\bar{\boldsymbol{\theta}}(\tau))^\top \succeq \frac{c_4}{d^{o(1)} + \frac{d/n}{T}\tau}\operatorname{diag}\Big(d^{-1}, d^{-1}, d^{-1/2}\mathbf{I}\Big).$$

Combining Lemmas 4.7 and 4.8 gives

$$-\frac{\mathrm{d}}{\mathrm{d}\tau}(\mathcal{L}(\tau) - \mathcal{L}^*) = \|\nabla_{\bar{\boldsymbol{\theta}}(\tau)}\mathcal{L}(\tau)\|^2 \ge \frac{c_3c_4(d/n)}{T\left(d^{o(1)} + \frac{d/n}{T}\tau\right)}(\mathcal{L}(\tau) - \mathcal{L}^*).$$

Solving this inequality gives, for some constant $\alpha > 0$ and every $0 < \epsilon < 1$, $\mathcal{L}(\tau) - \mathcal{L}^* \le \epsilon$ after time $O\big(d^{o(1)}(\epsilon/d)^{-1/\alpha}\big)$. This proves the second conclusion of Theorem 3.1.

We now explain the proof of the last conclusion of Theorem 3.1 and discuss the comparison with the Heavy-Ball method. By Lemma 4.2, the iterates given by the trained transformer can be expressed by a matrix polynomial of $\mathbf{S}$. This motivates us to give the following definition. Define

$$p_t^\star := \operatorname*{arg\,min}_{\substack{p(0)=0\\ \deg p\le t}} \mathbb{E}\left\|p(\mathbf{S})\mathbf{w}^\star - \mathbf{w}^\star\right\|^2, \qquad 0\le t<T.$$

In fact, recent works have pointed out that $p_t^\star$ is related to the Heavy-Ball method under the MP law (Pedregosa & Scieur, 2020; Scieur & Pedregosa, 2020). However, these results mostly focus on the asymptotic case when $n, d \to \infty$. Here we give the following lemma under the finite n, d setting.

Lemma 4.9. *For every $0 \le t < T$,*

$$p_t^\star(\mathbf{S})\mathbf{w}^\star = \mathbf{w}_t^{\mathrm{HB}}.$$

Moreover, there exists a constant $c_5 > 0$ such that, uniformly for $T/2 \le t < T$,

$$\mathbb{E}\left\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star\right\|^2 \ge c_5 d\left(\frac{d}{n}\right)^t.$$

Now we relate the algorithm learned by the transformer model to the Heavy-Ball method by comparing $p_{t,\bar{\boldsymbol{\theta}}(\tau)}$ with $p_t^\star$. Based on Lemma 4.7, we can further show that $\widetilde{\boldsymbol{\theta}}'(\tau) \to \widetilde{\boldsymbol{\theta}}'^\star$ and $|\widetilde{\theta}_2^\star| < 1$. By the definition of $p_{t,\bar{\boldsymbol{\theta}}(\tau)}$ in Lemma 4.6, we see that $p_{t,\bar{\boldsymbol{\theta}}(\tau)}$ can also be equivalently defined by $\widetilde{\boldsymbol{\theta}}'(\tau)$. Therefore, we introduce a slight abuse of notation and denote by $p_{t,\widetilde{\boldsymbol{\theta}}'^\star}$ the training limit of $p_{t,\bar{\boldsymbol{\theta}}(\tau)}$. By Lemma 4.2, $p_{t,\widetilde{\boldsymbol{\theta}}'^\star}$ further determines a unique $\mathbf{w}_t^\star$ by

$$\mathbf{w}_t^\star := p_{t,\widetilde{\boldsymbol{\theta}}'^\star}(\mathbf{S})\mathbf{w}^\star, \qquad 0 \le t \le T.$$

It follows from Lemma 4.6 that the highest-degree coefficient of $p_{t+1,\widetilde{\boldsymbol{\theta}}'^\star}$ is $\widetilde{\theta}_1^\star(\widetilde{\theta}_2^\star)^t$. For fixed $\widetilde{\theta}_1^\star$ and $\widetilde{\theta}_2^\star$, the coefficients of degrees $1, \dots, t$ can be freely adjusted through $\widetilde{\boldsymbol{\Theta}}_4$. Hence $p_{t+1,\widetilde{\boldsymbol{\theta}}'^\star}$ is the optimal degree-$(t+1)$ polynomial with a certain highest-degree coefficient. Since $|\widetilde{\theta}_2^\star| < 1$, the highest-degree coefficient decays exponentially, suggesting that for late t, $p_{t+1,\widetilde{\boldsymbol{\theta}}'^\star}$ approaches $p_t^\star$. Motivated by this, we have the following lemma.

Lemma 4.10. *There exists a constant $c_6 > 0$ such that, uniformly for $T/2 \le t < T$,*

$$\mathbb{E}\left\|\mathbf{w}_{t+1}^\star - \mathbf{w}_t^{\mathrm{HB}}\right\|^2 = \mathbb{E}\left\|[p_{t+1,\widetilde{\boldsymbol{\theta}}'^\star}(\mathbf{S}) - p_t^\star(\mathbf{S})]\mathbf{w}^\star\right\|^2 \le c_6 d \exp\left[-2(\log d)^{1/4}\right]\left(\frac{d}{n}\right)^t.$$

Note that $\mathbf{w}_t(\tau) \to \mathbf{w}_t^\star$ as $\mathcal{L}(\tau) \to \mathcal{L}^*$. By choosing a sufficiently large τ, we can make $\mathbf{w}_t(\tau)$ sufficiently close to $\mathbf{w}_t^\star$ so that, uniformly for $T/2 \le t < T$,

$$\mathbb{E}\left\|\mathbf{w}_{t+1}(\tau) - \mathbf{w}_{t+1}^\star\right\|^2 \le \frac{c_5\epsilon}{4} d\left(\frac{d}{n}\right)^T.$$

Combining Lemmas 4.9 and 4.10 with the triangle inequality gives, uniformly for $T/2 \le t < T$,

$$\frac{\mathbb{E}\left\|\mathbf{w}_{t+1}(\tau) - \mathbf{w}_t^{\mathrm{HB}}\right\|^2}{\mathbb{E}\left\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star\right\|^2} \le \epsilon/2 + \frac{2c_6}{c_5}\exp\left[-2(\log d)^{1/4}\right] \le \epsilon$$

for $n, d = \Omega(\exp(\log^4(1/\epsilon)))$ and $0 < \epsilon < 1$, the preceding estimates prove Theorem 3.1.

5 Experiment

In this section, we introduce our experimental setup to empirically examine our theoretical results. In all the experiments, we test two settings with $(d, n) = (16, 32)$ and $(12, 24)$ respectively. We set the maximum CoT steps as $T = 24$, and let the learning rate of the GD teacher in Stage 1 be $\eta_0 = 0.5$. The entries of $\mathbf{V}, \mathbf{W}$ are initialized independently from $N(\mathbf{0}, 10^{-4}\mathbf{I})$. We use float64 online Adam(W) with zero weight decay, learning rate 10^{-6}, and batch size 6144 to train the models. The in-context weight prediction data are generated exactly following the definitions in Subsection 2.1.

Several aspects of our experiments differ from the theoretical setting. First of all, we consider Gaussian random initialization for the weight matrices $\mathbf{V}$ and $\mathbf{W}$, without freezing any blocks or entries. Moreover, in Stage 1, we remove teacher forcing and define the loss using iterates generated autoregressively by the transformer itself. In addition, we train the model using online stochastic Adam rather than gradient descent/flow. Finally, while our theory assumes $d/n < 0.1$, this condition is not satisfied in our experimental setup. We remark that strictly following the theoretical setting yields similar experimental results. The above modifications are included mainly to empirically show that our theoretical findings can capture more general training behavior in practice.

We evaluate several methods for in-context weight prediction. Besides transformers trained with the two-stage procedure, we consider transformers trained using only Stage 1 or only Stage 2. We also examine a transformer variant without positional encodings. We compare these models with a well-tuned time-varying heavy-ball method and the gradient-descent teacher used in Stage 1.

Figure 2 presents experiment results in two settings, $d = 16, n = 32$ and $d = 12, n = 24$. In both settings, the transformer trained with the two-stage procedure closely matches the Heavy-Ball method, consistent with Theorem 3.1. In contrast, both (i) the transformer without positional encodings, trained under the same two-stage procedure, and (ii) the transformer trained only in the first

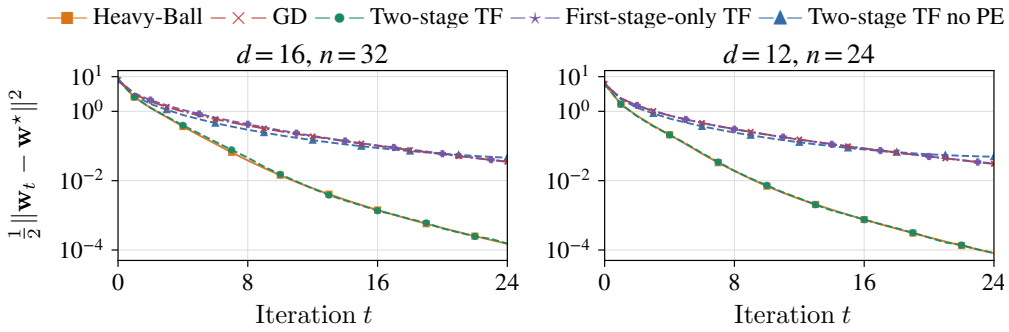


Figure 2: In-context weight prediction performance of several models and baseline algorithms.

stage with gradient-descent supervision perform similarly to gradient descent. This further demonstrates the necessity of the two-stage training procedure together with positional encodings for the CoT steps. The second-stage-only transformer is not stable and therefore is not included in Figure 2.

Figure 3 further shows example heat maps of the weight matrices learned via two-stage training, along with evidence that training a randomly initialized transformer directly in the second stage is unstable. The learned parameter structure is consistent with Lemma 4.1: the dominant blocks are $\mathbf{W}_{13}, \mathbf{V}_{31}, \mathbf{V}_{33}, w_{24}, \mathbf{W}_{55}$, while the remaining entries are small. $\mathbf{V}_{31}, \mathbf{W}_{13}, \mathbf{V}_{33}$ are approximately diagonal, whereas $\mathbf{W}_{55}$ has clear nonzero off-diagonal entries. By contrast, direct second-stage training from random initialization, tested for $d \in \{12, 16\}$ with $n = 32$, consistently leads to nonfinite gradients, with smaller learning rates only delaying failure.

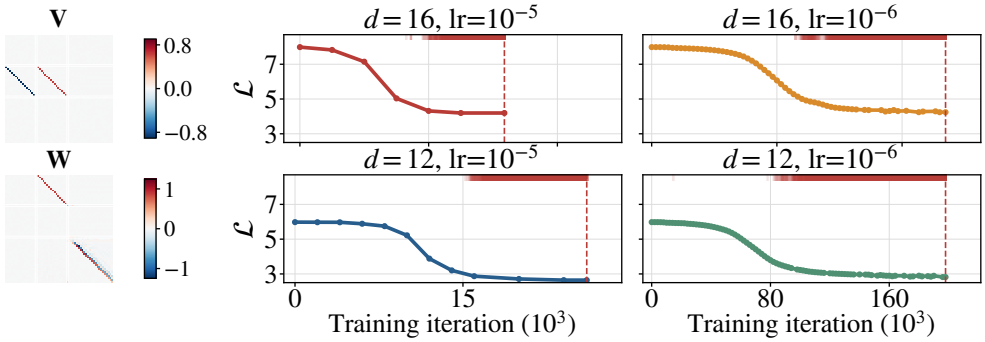


Figure 3: Examples of the heatmaps of learned weight matrices after the two-stage training for $d = 16, n = 32$ (left), together with the progress of second-stage-only training for $n = 32$ under four different combinations of dimension d and learning rate (right). For second-stage-only training, nonfinite gradient values are marked in red. When nonfinite gradients occur, we do not stop training immediately; instead, we allow the model to retry using a newly generated independent in-context dataset. Training is terminated only after nonfinite gradients are observed for 20 consecutive retries. Dashed lines indicate termination.

6 Conclusion

In this paper, we study in-context weight prediction with transformers using CoT reasoning. We focus on whether transformers with CoT can go beyond mimicking a teacher algorithm and discover new algorithms. We show that a two-stage training paradigm enables such discovery: the first stage supervises the CoT trajectory with a gradient-descent teacher, while the second stage trains the model using only the ground-truth regression weights. The learned model discovers the Heavy-Ball method, achieving acceleration over the gradient-descent teacher used in the first stage.

Our current analysis is limited to linear self-attention for in-context linear regression. Extending the findings to more practical softmax attention and more general learning tasks are important future work directions. Moreover, our current setup considers training with gradient flow. Extending the training methods to other training algorithms, including stochastic gradient descent, AdamW, and reinforcement learning-based training techniques, is also interesting future work directions. It is also interesting to extend our findings to other types of positional encodings.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with theorem proving and verification, experiment code generation, notation consistency, and presentation polishing. We did not use generative AI tools to generate research ideas. The authors reviewed and verified all AI-assisted work and take responsibility for the final content of this work.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *International Conference on Learning Representations*, 2022.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

A. N. Bishop, P. Del Moral, and A. Niclas. An introduction to Wishart matrix moments. *Foundations and Trends in Machine Learning*, 11(2):97–218, 2018.

Włodzimierz Bryc. Compound real wishart and q-wishart matrices. *International Mathematics Research Notices*, 2008(9):rnn079–rnn079, 2008.

Yuan Cao, Yihan He, Dennis Wu, Hong-Yu Chen, Jianqing Fan, and Han Liu. Transformers simulate mle for sequence generation in bayesian networks. *arXiv preprint arXiv:2501.02547*, 2025.

Sitan Chen and Yuanzhi Li. Provably learning a multi-head attention layer. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1744–1754, 2025.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. *Advances in Neural Information Processing Systems*, 37:66479–66567, 2024.

Xingwu Chen and Difan Zou. What can transformer learn with varying depth? case studies on sequence learning tasks. In *International Conference on Machine Learning*, 2024.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in neural information processing systems*, 36:70757–70798, 2023.

K. Gatmiry, N. Saunshi, S. J. Reddi, S. Jegelka, and S. Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, 2024.

Halil Alperen Gozeten, M Emrullah Ildiz, Xuechen Zhang, Mahdi Soltanolkotabi, Marco Mondelli, and Samet Oymak. Test-time training provably improves transformers as in-context learners. *Proceedings of machine learning research*, 267:20266, 2025.

Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. How do transformers learn in-context beyond simple functions? a case study on learning with representations. In *International Conference on Learning Representations*, volume 2024, pp. 46014–46042, 2024.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4902–4919, 2023.

Xinyang Hu, Fengzhuo Zhang, Siyu Chen, and Zhuoran Yang. Unveiling the statistical foundations of chain-of-thought prompting methods. *Journal of Machine Learning Research*, 27(202):1–104, 2026.

Jianhao Huang, Zixuan Wang, and Jason Lee. Transformers learn to implement multi-step gradient descent with chain of thought. In *International Conference on Learning Representations*, 2025.

Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Timothy Kusalik, James A Mingo, and Roland Speicher. Orthogonal polynomials and fluctuations of random matrices. *arXiv preprint math/0503169*, 2005.

Tianle Li, Chenyang Zhang, Xingwu Chen, Yuan Cao, and Difan Zou. On the robustness of transformers against context hijacking for linear classification. *Advances in Neural Information Processing Systems*, 38:69373–69411, 2026a.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International conference on machine learning*, pp. 19565–19594. PMLR, 2023.

Yingcong Li, Xiangyu Chang, Muti Kara, Xiaofeng Liu, Amit Roy-Chowdhury, and Samet Oymak. When and how unlabeled data provably improve in-context learning. *Advances in Neural Information Processing Systems*, 38:31831–31866, 2026b.

Zihao Li, Yuan Cao, Cheng Gao, Yihan He, Han Liu, Jason M Klusowski, Jianqing Fan, and Mengdi Wang. One-layer transformer provably learns one-nearest neighbor in context. *Advances in Neural Information Processing Systems*, 37:82166–82204, 2024.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations*, volume 2024, pp. 7690–7706, 2024.

Alireza Mousavi-Hosseini, Clayton Sanford, Denny Wu, and Murat Erdogdu. When do transformers outperform feedforward and recurrent networks? a statistical perspective. *Advances in Neural Information Processing Systems*, 38:64629–64676, 2026.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. *International Conference on Machine Learning*, 2024.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Fabian Pedregosa and Damien Scieur. Acceleration through spectral density estimation. In *International Conference on Machine Learning*, pp. 7553–7562. PMLR, 2020.

C Emily I Redelmeier. Genus expansion for real wishart matrices. *Journal of Theoretical Probability*, 24(4):1044–1062, 2011.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *International Conference on Learning Representations*, 2022.

Damien Scieur and Fabian Pedregosa. Universal average-case optimality of polyak momentum. In *International conference on machine learning*, pp. 8565–8572. PMLR, 2020.

Wei Shi and Yuan Cao. Towards understanding transformers in learning random walks. *Advances in Neural Information Processing Systems*, 38:66411–66454, 2026.

Zhongjie Shi, Zhiying Fang, and Yuan Cao. Approximation and estimation capability of vision transformers for hierarchical compositional models. *Applied and Computational Harmonic Analysis*, pp. 101849, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2717–2739, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. In *International Conference on Machine Learning*, 2024.

Zixuan Wang, Eshaan Nichani, Alberto Bietti, Alex Damian, Daniel Hsu, Jason D Lee, and Denny Wu. Learning compositional functions with transformers from easy-to-hard data. *arXiv preprint arXiv:2505.23683*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Lyumin Wu, Chenyang Zhang, and Yuan Cao. Looped transformers with layer normalization provably learn the power method. *arXiv preprint arXiv:2606.00605*, 2026a.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *arXiv preprint arXiv:2307.13339*, 2023.

Yu-Hsuan Wu, Yihan He, Yuan Cao, Jianqing Fan, and Han Liu. Understanding transformers for time series forecasting: A case study on moirai. In *International Conference on Learning Representations*, volume 2026, pp. 94518–94554, 2026b.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. Multi-head transformers provably learn symbolic multi-step reasoning via gradient descent. *Advances in Neural Information Processing Systems*, 38:116194–116285, 2026a.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. Agentic transformers provably learn to search via reinforcement learning. *arXiv preprint arXiv:2606.00183*, 2026b.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Chenyang Zhang and Yuan Cao. Transformers efficiently perform in-context logistic regression via normalized gradient descent. *arXiv preprint arXiv:2605.06609*, 2026.

Chenyang Zhang, Qingyue Zhao, Quanquan Gu, and Yuan Cao. Transformers trained via gradient descent can provably learn a class of teacher models. *International Conference on Learning Representations*, 2026.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

A Additional Related Works

Expressive power and generalization of transformers as algorithm executors. A line of recent works studies transformers as algorithm executors, aiming to characterize what algorithms can be implicitly implemented by transformers through in-context learning. Li et al. (2023) provide a theoretical framework for understanding transformers as algorithm executors and establish generalization and stability properties for in-context learning. Chen & Zou (2024) investigate how transformer depth affects sequence learning capabilities, showing that while shallow transformers can achieve memorization, more sophisticated reasoning and generalization behaviors require deeper architectures. Cao et al. (2025) study transformer expressivity in Bayesian network sequence modeling and show that transformers can implement in-context maximum likelihood estimation and autoregressive sampling. Shi et al. (2025) characterize the approximation and estimation capabilities of vision transformers for hierarchical compositional models.

Recent works further explore the connection between transformer computation and classical statistical algorithms. Mousavi-Hosseini et al. (2026) study the statistical advantages of transformers over feedforward and recurrent architectures by analyzing their sample complexity under dynamic sparse retrieval settings. Li et al. (2026a) analyze the robustness of transformers in the presence of context hijacking for linear classification tasks. Li et al. (2026b) study when unlabeled data can provably improve in-context learning performance and show that deeper or looped transformers can implicitly construct iterative estimators related to classical semi-supervised learning algorithms. Wu et al. (2026b) investigate transformer-based time-series forecasting through the case study of Moirai and provide theoretical insights into how transformers perform sequential prediction tasks.

Training guarantees of transformers. A complementary line of work studies whether gradient-based training can enable transformers to recover nontrivial algorithmic structures and computational mechanisms. Jelassi et al. (2022) show that vision transformers can provably learn spatial structures under appropriate training settings. Nichani et al. (2024) investigate how gradient descent enables transformers to recover latent causal structures, while Wang et al. (2024) demonstrate that transformers can learn sparse token selection mechanisms that are difficult for fully connected networks to recover. Chen et al. (2024) characterize the emergence and training dynamics of induction heads, revealing how attention mechanisms learn retrieval-like behaviors from data. Li et al. (2024) prove that a one-layer transformer can learn one-nearest-neighbor prediction in context, and Wang et al. (2025) study the learning of compositional functions with transformers under easy-to-hard data curricula.

Recent works further extend these training guarantees to broader algorithmic settings. Chen & Li (2025) establish provable learning guarantees for multi-head attention layers. Gozeten et al. (2025) show that test-time training can improve transformers as in-context learners. Shi & Cao (2026) analyze how transformers learn random walk structures, while Zhang et al. (2026) study teacher-student settings and prove that gradient-trained transformers can learn classes of teacher models. Wu et al. (2026a) and Zhang & Cao (2026) further investigate looped transformers and normalized gradient descent, providing theoretical guarantees for iterative optimization procedures implemented by transformers.

Studies on chain-of-thought reasoning. Recent works also study chain-of-thought (CoT) reasoning as a mechanism that enables transformers to perform multi-step computation. Wei et al. (2022) introduced the CoT prompting method to enhance the reasoning capability of language models by generating intermediate reasoning steps. Before the formalization of CoT, Nye et al. (2021) demonstrated that allowing models to produce intermediate computations through scratchpads can substantially improve multi-step problem solving. Subsequent works proposed various extensions of CoT and scratchpad prompting to improve reasoning efficiency and reliability (Wang et al., 2022; Yao et al., 2023).

Recent studies have further investigated the mechanisms underlying CoT from empirical, interpretability, and theoretical perspectives. Wang et al. (2023); Saparov & He (2022) empirically analyzed the reasoning behaviors induced by CoT, while Wu et al. (2023); Hou et al. (2023) examined how intermediate reasoning emerges from transformer computation through mechanistic analysis. On the theoretical side, Merrill & Sabharwal (2024); Feng et al. (2023) characterized the expressive power of transformers with CoT, showing that intermediate reasoning steps can substantially extend their computational capabilities. Hu et al. (2026) further investigated the statistical foundations of CoT prompting.

Going beyond these studies, Huang et al. (2025) provided the first theoretical analysis of CoT training dynamics in an in-context learning setting, showing that transformers trained with CoT supervision can learn to implement multi-step gradient descent. Yang et al. (2026a) showed that transformers with CoT can learn to implement path finding algorithms in trees. More recently, Yang et al. (2026b) showed that transformers with CoT can be trained by policy gradient under a depth-wise curriculum to perform randomized depth-first search. Our work builds upon this line of research by studying whether CoT trajectory-level training can enable transformers to discover accelerated optimization algorithms beyond the supervised teacher algorithm.

B Proofs of Lemmas 4.1–4.2

We use lower-case unbolded letters for scalars, lower-case bold letters for vectors, and upper-case bold letters for matrices. For a positive integer m, let $[m] = \{1, \dots, m\}$. We use $\mathbf{I}_m$ for the $m \times m$ identity matrix, $\mathbf{0}_m$ and $\mathbf{1}_m$ for the zero and all-ones vectors in $\mathbb{R}^m$, and $\mathbf{0}_{m\times k}$ for the $m \times k$ zero matrix. For a matrix $\mathbf{A}$, $\mathbf{A}_{[:,-1]}$ denotes its last column, $\mathbf{A}_{[i:j,-1]}$ denotes rows i through j of that column, and $\mathrm{vec}(\mathbf{A})$ stacks its columns into a vector. For an effective array, entries outside the displayed active index ranges are defined to be zero. The arrays retain their stated ambient dimensions; sums and derivatives below are restricted to the displayed active ranges. We write $\mathbf{V}_{ij}$ and $\mathbf{W}_{ij}$ for the blocks of the parameter matrices $\mathbf{V}$ and $\mathbf{W}$, with the block sizes specified below. We use $\|\mathbf{v}\| = \|\mathbf{v}\|_2$ for the Euclidean norm of a vector, and $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_F$ for the spectral and Frobenius norms of a matrix, respectively. The notation $\mathbb{1}\{\cdot\}$ denotes the indicator.

B.1 Proof of Lemma 4.1

By Equation (1), the retained output is $\mathbf{w}_{t+1} := F_{\mathrm{LSA}}(\mathbf{Z}_t; \mathbf{V}, \mathbf{W})_{[d+2:2d+1,-1]}$. Before imposing any zero-gradient identities, direct multiplication with the single block $\mathbf{W}_{33}$ fixed to zero gives

$$
\begin{aligned}
\mathbf{w}_{t+1} = \mathbf{w}_t &+ \mathbf{V}_{31} \sum_{i=1}^{n} \frac{\mathbf{x}_i}{\sqrt{n}} \left(\begin{pmatrix} \mathbf{x}_i/\sqrt{n} \\ y_i/\sqrt{n} \\ \mathbf{0}_d \\ 0 \\ \mathbf{0}_T \end{pmatrix}^T \mathbf{W} \begin{pmatrix} \mathbf{0}_d \\ 0 \\ \mathbf{w}_t \\ 1 \\ \mathbf{e}_t \end{pmatrix} \right) \\
&+ \mathbf{V}_{32} \sum_{i=1}^{n} \frac{y_i}{\sqrt{n}} \left(\begin{pmatrix} \mathbf{x}_i/\sqrt{n} \\ y_i/\sqrt{n} \\ \mathbf{0}_d \\ 0 \\ \mathbf{0}_T \end{pmatrix}^T \mathbf{W} \begin{pmatrix} \mathbf{0}_d \\ 0 \\ \mathbf{w}_t \\ 1 \\ \mathbf{e}_t \end{pmatrix} \right) \\
&+ \sum_{\ell=0}^{t} (\mathbf{V}_{33}\mathbf{w}_\ell + \mathbf{V}_{34} + \mathbf{V}_{35}\mathbf{e}_\ell) \left(\begin{pmatrix} \mathbf{0}_d \\ 0 \\ \mathbf{w}_\ell \\ 1 \\ \mathbf{e}_\ell \end{pmatrix}^T \mathbf{W} \begin{pmatrix} \mathbf{0}_d \\ 0 \\ \mathbf{w}_t \\ 1 \\ \mathbf{e}_t \end{pmatrix} \right). \qquad (9)
\end{aligned}
$$

In particular, the factors appearing in this calculation are

$$\begin{pmatrix} \mathbf{x}_i/\sqrt{n} \\ y_i/\sqrt{n} \\ \mathbf{0}_d \\ 0 \\ \mathbf{0}_T \end{pmatrix}^T \mathbf{W} \begin{pmatrix} \mathbf{0}_d \\ 0 \\ \mathbf{w}_t \\ 1 \\ \mathbf{e}_t \end{pmatrix} = \frac{1}{\sqrt{n}} \left[\mathbf{x}_i^T \{ \mathbf{W}_{13}\mathbf{w}_t + \mathbf{W}_{14} + \mathbf{W}_{15}\mathbf{e}_t \} + y_i \{ \mathbf{W}_{23}\mathbf{w}_t + w_{24} + \mathbf{w}_{25}\mathbf{e}_t \} \right],$$

$$\begin{pmatrix} \mathbf{0}_d \\ 0 \\ \mathbf{w}_\ell \\ 1 \\ \mathbf{e}_\ell \end{pmatrix}^T \mathbf{W} \begin{pmatrix} \mathbf{0}_d \\ 0 \\ \mathbf{w}_t \\ 1 \\ \mathbf{e}_t \end{pmatrix} = \mathbf{w}_\ell^\top \mathbf{W}_{34} + \mathbf{w}_\ell^\top \mathbf{W}_{35}\mathbf{e}_t + \mathbf{W}_{43}\mathbf{w}_t + w_{44} + \mathbf{w}_{45}\mathbf{e}_t$$

$$+ \mathbf{e}_\ell^\top \mathbf{W}_{53}\mathbf{w}_t + \mathbf{e}_\ell^\top \mathbf{W}_{54} + \mathbf{e}_\ell^\top \mathbf{W}_{55}\mathbf{e}_t. \tag{10}$$

Proof of Lemma 4.1. Under Assumption 3.1, the population gradient flow preserves the form (11) for all $\tau \geq 0$:

$$\mathbf{V} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & 0 & \mathbf{0} \\ v_{31}\mathbf{I}_d & \mathbf{0} & v_{33}\mathbf{I}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \qquad \mathbf{W} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & w_{13}\mathbf{I}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & w_{24} & \mathbf{w}_{25} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} & w_{44} & \mathbf{w}_{45} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{w}_{54} & \mathbf{W}_{55} \end{pmatrix} \tag{11}$$

For a vector valued function $\mathbf{f}(\mathbf{z})$, we denote by $\mathbf{f}(\mathbf{z})[\mathbf{h}]$ the Fréchet derivative of $\mathbf{f}(\mathbf{z})$ at $\mathbf{z}$ in the direction of $\mathbf{h}$ satisfying

$$\lim_{\|\mathbf{h}\| \to 0} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})[\mathbf{h}]\|}{\|\mathbf{h}\|} = 0.$$

Then direct differentiation of $\{\text{RHS (9)}\}$ gives

$$\begin{aligned}
&D_{\mathbf{V}_{32}}\{\text{RHS (9)}\}[\delta\mathbf{V}_{32}] \\
&\quad = \delta\mathbf{V}_{32} \left\{ (\mathbf{w}^\star)^\mathsf{T}\mathbf{S}\mathbf{W}_{13}\mathbf{w}_t + (w_{24} + \mathbf{w}_{25}\mathbf{e}_t)(\mathbf{w}^\star)^\mathsf{T}\mathbf{S}\mathbf{w}^\star \right\}, \\
&D_{(\mathbf{V}_{34},\mathbf{V}_{35})}\{\text{RHS (9)}\}[\delta\mathbf{V}_{34}, \delta\mathbf{V}_{35}] \\
&\quad = \sum_{\ell=0}^{t} (\delta\mathbf{V}_{34} + \delta\mathbf{V}_{35}\mathbf{e}_\ell) \left(w_{44} + \mathbf{w}_{45}\mathbf{e}_t + \mathbf{e}_\ell^\mathsf{T}\mathbf{w}_{54} + \mathbf{e}_\ell^\mathsf{T}\mathbf{W}_{55}\mathbf{e}_t \right), \\
&D_{(\mathbf{W}_{14},\mathbf{W}_{15})}\{\text{RHS (9)}\}[\delta\mathbf{W}_{14}, \delta\mathbf{W}_{15}] = \mathbf{V}_{31}\mathbf{S}(\delta\mathbf{W}_{14} + \delta\mathbf{W}_{15}\mathbf{e}_t), \\
&D_{\mathbf{W}_{23}}\{\text{RHS (9)}\}[\delta\mathbf{W}_{23}] = \mathbf{V}_{31}\mathbf{S}\mathbf{w}^\star\, \delta\mathbf{W}_{23}\mathbf{w}_t, \\
&D_{(\mathbf{W}_{34},\mathbf{W}_{35})}\{\text{RHS (9)}\}[\delta\mathbf{W}_{34}, \delta\mathbf{W}_{35}] = \mathbf{V}_{33} \sum_{\ell=0}^{t} \mathbf{w}_\ell \mathbf{w}_\ell^\mathsf{T} (\delta\mathbf{W}_{34} + \delta\mathbf{W}_{35}\mathbf{e}_t), \\
&D_{(\mathbf{W}_{43},\mathbf{W}_{53})}\{\text{RHS (9)}\}[\delta\mathbf{W}_{43}, \delta\mathbf{W}_{53}] = \mathbf{V}_{33} \sum_{\ell=0}^{t} \mathbf{w}_\ell (\delta\mathbf{W}_{43} + \mathbf{e}_\ell^\mathsf{T}\delta\mathbf{W}_{53})\mathbf{w}_t.
\end{aligned}$$

Moreover,

$$\begin{aligned}
D_{\mathbf{w}_\ell}\{\text{RHS (9)}\} &= \mathbb{1}\{\ell = t\}(\mathbf{I}_d + \mathbf{V}_{31}\mathbf{S}\mathbf{W}_{13}) \\
&\quad + \left(w_{44} + \mathbf{w}_{45}\mathbf{e}_t + \mathbf{e}_\ell^\mathsf{T}\mathbf{w}_{54} + \mathbf{e}_\ell^\mathsf{T}\mathbf{W}_{55}\mathbf{e}_t \right) \mathbf{V}_{33}.
\end{aligned}$$

Since $(\mathbf{X}, \mathbf{w}^\star) \overset{\mathrm{d}}{=} (\mathbf{X}, -\mathbf{w}^\star)$ and $\mathbf{w}_t(\mathbf{X}, -\mathbf{w}^\star) = -\mathbf{w}_t(\mathbf{X}, \mathbf{w}^\star)$, these partial derivatives are even in $\mathbf{w}^\star$. For every displayed block B,

$$\begin{aligned}
D_B\mathbf{w}_0[\delta B] &= 0, \\
D_B\mathbf{w}_{t+1}[\delta B] &= \sum_{\ell=0}^{t} D_{\mathbf{w}_\ell}\{\text{RHS (9)}\} D_B\mathbf{w}_\ell[\delta B] + D_B\{\text{RHS (9)}\}[\delta B], \qquad (12) \\
D_B\mathbf{w}_t(\mathbf{X}, -\mathbf{w}^\star)[\delta B] &= D_B\mathbf{w}_t(\mathbf{X}, \mathbf{w}^\star)[\delta B].
\end{aligned}$$

Changing $\mathbf{w}^\star$ to $-\mathbf{w}^\star$ therefore gives

$$\mathbb{E}\langle \mathbf{w}_t - \mathbf{w}^\star, D_B \mathbf{w}_t[\delta B]\rangle = -\mathbb{E}\langle \mathbf{w}_t - \mathbf{w}^\star, D_B \mathbf{w}_t[\delta B]\rangle = 0,$$
$$D_B\mathcal{L}[\delta B] = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\langle \mathbf{w}_t - \mathbf{w}^\star, D_B \mathbf{w}_t[\delta B]\rangle = 0. \tag{13}$$

The remaining trainable zero blocks satisfy

$$D_{\mathbf{V}_{ij}}\mathbf{w}_t = 0 \quad (i \neq 3), \qquad D_{\mathbf{W}_{i1}}\mathbf{w}_t = D_{\mathbf{W}_{i2}}\mathbf{w}_t = 0 \quad (1 \le i \le 5).$$

For the Stage-1 inputs,

$$\mathbf{w}_t^{\mathrm{GD}}(\mathbf{X}, -\mathbf{w}^\star) = -\mathbf{w}_t^{\mathrm{GD}}(\mathbf{X}, \mathbf{w}^\star), \qquad D_B\mathbf{w}_t^{\mathrm{GD}} = 0, \qquad D_B\mathcal{L}_{\mathrm{GD}} = 0.$$

The output rows $i \neq 3$ have zero residual. Thus, for every trainable zero block in (11),

$$B(0) = 0, \qquad D_B\mathcal{L}_{\mathrm{GD}} = D_B\mathcal{L} = 0 \implies B(\tau) = 0.$$

□

Rotational equivariance and uniqueness of the same ODE give

$$\mathbf{V}_{31} = v_{31}\mathbf{I}_d, \qquad \mathbf{W}_{13} = w_{13}\mathbf{I}_d, \qquad \mathbf{V}_{33} = v_{33}\mathbf{I}_d.$$

For $0 \le \ell \le t < T$, define directly from the displayed $\mathbf{W}$-blocks

$$(\widetilde{\mathbf{W}}_{55})_{\ell t} := v_{33}\big(w_{44} + \mathbf{w}_{45}\mathbf{e}_t + \mathbf{e}_\ell^\top \mathbf{W}_{54} + \mathbf{e}_\ell^\top \mathbf{W}_{55}\mathbf{e}_t\big). \tag{14}$$

The retained update is now

$$\mathbf{w}_{t+1} = \mathbf{w}_t + v_{31}w_{13}\mathbf{S}\mathbf{w}_t + v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t)\mathbf{X}\mathbf{y}^\top + \sum_{\ell=0}^{t}(\widetilde{\mathbf{W}}_{55})_{\ell t}\mathbf{w}_\ell, \tag{15}$$

$$p_0(\lambda) = 0, \qquad p_{t+1}(\lambda) = p_t(\lambda) + \lambda\{\widetilde{\theta}_2 p_t(\lambda) + v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t)\} + \sum_{\ell=1}^{t}(\widetilde{\mathbf{W}}_{55})_{\ell t}p_\ell(\lambda). \tag{16}$$

The $\ell = 0$ term vanishes because $\mathbf{w}_0 = 0$.

Rotational reduction. For every $\mathbf{U} \in O(d)$,

$$(\mathbf{x}_i, \mathbf{w}^\star, \mathbf{w}_t) \mapsto (\mathbf{U}\mathbf{x}_i, \mathbf{U}\mathbf{w}^\star, \mathbf{U}\mathbf{w}_t),$$

and

$$(\mathbf{V}_{31}, \mathbf{W}_{13}, \mathbf{V}_{33}) \mapsto (\mathbf{U}\mathbf{V}_{31}\mathbf{U}^\top, \mathbf{U}\mathbf{W}_{13}\mathbf{U}^\top, \mathbf{U}\mathbf{V}_{33}\mathbf{U}^\top),$$
$$\mathcal{L}(\mathbf{U}\mathbf{V}_{31}\mathbf{U}^\top, \mathbf{U}\mathbf{W}_{13}\mathbf{U}^\top, \mathbf{U}\mathbf{V}_{33}\mathbf{U}^\top) = \mathcal{L}(\mathbf{V}_{31}, \mathbf{W}_{13}, \mathbf{V}_{33}),$$
$$\nabla_A\mathcal{L}|_{(\mathbf{U}\mathbf{V}_{31}\mathbf{U}^\top, \mathbf{U}\mathbf{W}_{13}\mathbf{U}^\top, \mathbf{U}\mathbf{V}_{33}\mathbf{U}^\top)} = \mathbf{U}\,\nabla_A\mathcal{L}|_{(\mathbf{V}_{31}, \mathbf{W}_{13}, \mathbf{V}_{33})}\,\mathbf{U}^\top, \qquad A \in \{\mathbf{V}_{31}, \mathbf{W}_{13}, \mathbf{V}_{33}\}.$$

Since $[\mathbf{V}_{31}(0), \mathbf{U}] = [\mathbf{W}_{13}(0), \mathbf{U}] = [\mathbf{V}_{33}(0), \mathbf{U}] = 0$, uniqueness of the gradient-flow ODE gives

$$[\mathbf{V}_{31}(\tau), \mathbf{U}] = [\mathbf{W}_{13}(\tau), \mathbf{U}] = [\mathbf{V}_{33}(\tau), \mathbf{U}] = 0 \qquad (\mathbf{U} \in O(d),\ \tau \ge 0).$$

Coordinate sign flips force every such matrix to be diagonal, and coordinate permutations force equal diagonal entries. Hence

$$\mathbf{V}_{31}(\tau) = v_{31}(\tau)\mathbf{I}_d, \qquad \mathbf{W}_{13}(\tau) = w_{13}(\tau)\mathbf{I}_d, \qquad \mathbf{V}_{33}(\tau) = v_{33}(\tau)\mathbf{I}_d.$$

B.2 Proof of Lemma 4.2

Proof of Lemma 4.2. Apply the conclusion of Lemma 4.1 to (9) and we have that in both stages, the predicted iterate has the form $\mathbf{w}_t = p_{t,\boldsymbol{\theta}}(\mathbf{S})\mathbf{w}^\star$, with $p_{t,\boldsymbol{\theta}}(0) = 0$ and $\deg p_{t,\boldsymbol{\theta}} \le t$. In Stage 1, each teacher iterate is a zero-constant polynomial in $\mathbf{S}$, and the retained update increases its degree by at most one. In Stage 2, the same conclusion follows by induction in (16), starting from $p_0 = 0$. □

C PROOFS OF LEMMAS 4.3–4.5

Throughout the appendices, we use the same scaling assumptions:

$$n = \Theta(d), \qquad \frac{d}{n} \le \frac{1}{10}, \qquad T = \Theta\Big((\log d)^{1/4}\Big), \tag{17}$$

The horizon constant is fixed so that, for all sufficiently large d,

$$\left(\frac{d}{n}\right)^2 T \ge 1024(\log d)^{1/4}. \tag{18}$$

This is compatible with (17) and is used only in the late-time spectral comparison. Equations (17)–(18) specify the fixed constants implicit in the Θ-scalings of Assumption 3.2: the horizon constants may depend on the fixed comparison bounds in $n = \Theta(d)$, but are independent of d, n, T and the target accuracy ϵ.

The uniform teacher-polynomial estimates used below are proved in Section G.3.

Throughout this appendix $w_{24}(0) = \sigma < 0$, $\mathbf{w}_{25}(0) = \mathbf{0}_{1\times T}$, $w_{44}(0) = 0$, $\mathbf{w}_{45}(0) = \mathbf{0}_{1\times T}$, $\mathbf{w}_{54}(0) = \mathbf{0}_{T\times 1}$, and $\mathbf{W}_{55}(0) = \mathbf{0}_{T\times T}$; all these quantities are trainable from time zero. The only fixed block is $\mathbf{W}_{33} = \mathbf{0}_{d\times d}$. The scalar initialization is

$$v_{31}(0) = w_{13}(0) = 0, \qquad v_{33}(0) > 0, \qquad v_{33}(0) = \Theta(1).$$

C.1 PROOF OF LEMMA 4.3

Lemma C.1. *Under the Stage-1 initialization, every omitted block remains zero and*

$$\mathbf{V}_{31} = v_{31}\mathbf{I}_d, \qquad \mathbf{W}_{13} = w_{13}\mathbf{I}_d, \qquad \mathbf{V}_{33} = v_{33}\mathbf{I}_d.$$

For

$$(\widetilde{\mathbf{W}}_{55})_{\ell t} = v_{33}\big(w_{44} + \mathbf{w}_{45}\mathbf{e}_t + \mathbf{e}_\ell^\top \mathbf{w}_{54} + \mathbf{e}_\ell^\top \mathbf{W}_{55}\mathbf{e}_t\big), \tag{19}$$

The Stage-1 auxiliary array in the proof sketch is

$$\begin{aligned} (\widetilde{\mathbf{\Theta}}_3)_{1t} &:= v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t) + \eta_0(\widetilde{\mathbf{W}}_{55})_{1t}, && 1 \le t < T, \\ (\widetilde{\mathbf{\Theta}}_3)_{\ell t} &:= (\widetilde{\mathbf{W}}_{55})_{\ell t}, && 2 \le \ell \le t < T. \end{aligned} \tag{20}$$

$$(\widetilde{\mathbf{\Theta}}_3)_{\ell t} = \mathbf{e}_\ell^\top \widetilde{\mathbf{\Theta}}_3 \mathbf{e}_t, \qquad 0 \le \ell, t < T,$$

$$\widetilde{\mathbf{\Theta}}_3 = \sum_{t=1}^{T-1}\sum_{\ell=1}^{t} (\widetilde{\mathbf{\Theta}}_3)_{\ell t}\mathbf{e}_\ell \mathbf{e}_t^\top.$$

The exact normalized loss is

$$\frac{\mathcal{L}_{\mathrm{GD}}}{d} = \frac{1}{2Td}\sum_{t=0}^{T-1} \mathbb{E}\ \mathrm{tr}\left(\left[\begin{array}{l} (v_{31}w_{13} + \eta_0)\mathbf{S}\{\mathbf{I} - (\mathbf{I} - \eta_0\mathbf{S})^t\} \\ + \{v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t) - \eta_0\}\mathbf{S} \\ + \displaystyle\sum_{\ell=1}^{t} (\widetilde{\mathbf{W}}_{55})_{\ell t}\{\mathbf{I} - (\mathbf{I} - \eta_0\mathbf{S})^\ell\} \end{array}\right]^2\right).$$

Proof. Replacing $\mathbf{w}^\star$ and every teacher iterate $\{\mathbf{I} - (\mathbf{I} - \eta_0\mathbf{S})^\ell\}\mathbf{w}^\star$ by its negative preserves their joint law. The third-row prediction error is odd under this transformation, whereas the directional derivative produced by any omitted block is even or lies in an output row orthogonal to the error. Its population inner product with the error is therefore zero. Rotational equivariance and ODE uniqueness then force the three square $d \times d$ blocks to remain scalar multiples of the identity.

The full-column target agrees with the input in every output row except row three. Direct multiplication of that row, using the teacher iterates, gives

$$\begin{aligned} &\{\mathbf{I} - (\mathbf{I} - \eta_0\mathbf{S})^t\}\mathbf{w}^\star + v_{31}w_{13}\mathbf{S}\{\mathbf{I} - (\mathbf{I} - \eta_0\mathbf{S})^t\}\mathbf{w}^\star + v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t)\mathbf{X}\mathbf{y}^\top \\ &\qquad + \sum_{\ell=1}^{t} (\widetilde{\mathbf{W}}_{55})_{\ell t}\{\mathbf{I} - (\mathbf{I} - \eta_0\mathbf{S})^\ell\}\mathbf{w}^\star. \end{aligned}$$

Subtracting the $(t+1)$-st GD iterate and using

$$\{\mathbf{I} - (\mathbf{I} - \eta_0 \mathbf{S})^{t+1}\} - \{\mathbf{I} - (\mathbf{I} - \eta_0 \mathbf{S})^{t}\} = \eta_0 \mathbf{S}(\mathbf{I} - \eta_0 \mathbf{S})^t$$

and $\mathbb{E}[\mathbf{w}^\star (\mathbf{w}^\star)^\mathsf{T} \mid \mathbf{S}] = \mathbf{I}_d$ gives (21). □

For the spectral form of the loss $\mathcal{L}_{\mathrm{GD}}$, let $\nu_{d,n}$ be the expected spectral law of $\mathbf{S}$, and let $\mu_{d/n}$ be the Marchenko-Pastur law. We use the polynomial inner products and norms

$$\begin{aligned}
\mathbb{E}_{\lambda \sim \nu_{d,n}} f(\lambda) &= \frac{1}{d} \mathbb{E}\ \mathrm{tr}\, f(\mathbf{S}), \\
\langle f, g \rangle_{d,n} &= \frac{1}{d} \mathbb{E}\ \mathrm{tr}(f(\mathbf{S}) g(\mathbf{S}))\,, & \|R\|_{d,n}^2 &= \langle R, R \rangle_{d,n}, \\
\langle f, g \rangle_{\mathrm{MP}} &= \mathbb{E}_{\lambda \sim \mu_{d/n}}[f(\lambda) g(\lambda)], & \|R\|_{\mathrm{MP}}^2 &= \langle R, R \rangle_{\mathrm{MP}}.
\end{aligned}$$

Proof of Lemma 4.3. For $p_t(\lambda) = 1 - (1 - \eta_0 \lambda)^t$, write

$$\begin{aligned}
A_0(\lambda) &= (\widetilde{\theta}_1 - \eta_0)\lambda, \\
A_t(\lambda) &= (\widetilde{\theta}_2 + \eta_0)\lambda p_t(\lambda) + \big((\widetilde{\mathbf{\Theta}}_3)_{1t} - \eta_0\big)\lambda + \sum_{\ell=2}^{t} (\widetilde{\mathbf{\Theta}}_3)_{\ell t} p_\ell(\lambda), \qquad 1 \le t < T, \\
\frac{\mathcal{L}_{\mathrm{GD}}}{d} &= \frac{1}{2T} \sum_{t=0}^{T-1} \|A_t\|_{d,n}^2.
\end{aligned} \tag{22}$$

Let P_t be the orthogonal projection in $\langle \cdot, \cdot \rangle_{d,n}$ onto $\mathrm{span}\{\lambda, p_2, \ldots, p_t\}$. Then

$$A_t = (\widetilde{\theta}_2 + \eta_0)(\lambda p_t - P_t[\lambda p_t]) + P_t A_t, \qquad \langle \lambda p_t - P_t[\lambda p_t], P_t A_t \rangle_{d,n} = 0. \tag{23}$$

The functions in Lemma 4.3 are

$$\begin{aligned}
A(\mathbf{S}) &= \frac{\|\mathbf{S}\|_F^2}{2T}, \\
B(\mathbf{S}) &= \frac{1}{2T} \sum_{t=1}^{T-1} \|(\lambda p_t - P_t[\lambda p_t])(\mathbf{S})\|_F^2, \\
[\mathbf{C}_1(\mathbf{S})]_t &= \frac{1}{\sqrt{2T}} \mathrm{vec}\big((P_t[\lambda p_t])(\mathbf{S})\big), \\
[\mathbf{C}_2(\mathbf{S})]_{t,(\ell,s)} &= [\mathbf{C}_2(\mathbf{S})]_{t,sT+\ell+1} \\
&= \frac{\mathbb{1}\{t = s\}}{\sqrt{2T}} \begin{cases} \mathrm{vec}(\mathbf{S}), & \ell = 1,\ 1 \le s < T, \\ \mathrm{vec}(p_\ell(\mathbf{S})), & 2 \le \ell \le s < T, \\ \mathbf{0}_{d^2}, & \text{otherwise.} \end{cases} \qquad 0 \le \ell, s < T.
\end{aligned} \tag{24}$$

Here $t = 1, \ldots, T-1$ indexes blocks of d^2 rows. Substitution in (23) gives

$$\begin{aligned}
\mathcal{L}_{\mathrm{GD}} = \mathbb{E}\Big[& A(\mathbf{S})(\widetilde{\theta}_1 - \eta_0)^2 + B(\mathbf{S})(\widetilde{\theta}_2 + \eta_0)^2 \\
& + \|\mathbf{C}_1(\mathbf{S})(\widetilde{\theta}_2 + \eta_0) + \mathbf{C}_2(\mathbf{S})\, \mathrm{vec}(\widetilde{\mathbf{\Theta}}_3 - \eta_0 \mathbf{e}_1 (\mathbf{1}_T - \mathbf{e}_0)^\mathsf{T})\|^2 \Big].
\end{aligned}$$

□

Lemma C.2. *The loss $\mathcal{L}_{\mathrm{GD}}$ is zero if and only if*

$$v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0) = \eta_0, \qquad -v_{31} w_{13} = \eta_0, \tag{25}$$

$$(\widetilde{\mathbf{\Theta}}_3)_{1t} = \eta_0 \qquad (1 \le t < T), \tag{26}$$

$$(\widetilde{\mathbf{\Theta}}_3)_{\ell t} = 0 \qquad (2 \le \ell \le t < T). \tag{27}$$

Proof. At $t = 0$, the residual in (21) is $\{v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0) - \eta_0\}\mathbf{S}$, so positive definiteness of the finite moment Gram matrix gives the first identity in (25). For $t \geq 1$, only

$$(v_{31}w_{13} + \eta_0)\lambda\{1 - (1 - \eta_0\lambda)^t\}$$

has degree $t + 1$. Its leading coefficient is nonzero unless $-v_{31}w_{13} = \eta_0$. After imposing that identity, the remaining polynomials

$$\lambda, \quad 1 - (1 - \eta_0\lambda)^2, \quad \ldots, \quad 1 - (1 - \eta_0\lambda)^t$$

have successive degrees $1, \ldots, t$ and nonzero leading coefficients. Their coefficients must therefore vanish successively, giving (26)-(27). The converse follows by direct substitution. □

When calculating gradients, the scalar identity blocks have metric $d\{(\mathrm{d}v_{31})^2 + (\mathrm{d}w_{13})^2 + (\mathrm{d}v_{33})^2\}$, while every remaining retained block has its ordinary Frobenius metric. Equivalently, for $\mathcal{L}_{\mathrm{GD}}/d$, the raw Euclidean flow restricted to this invariant subspace is

$$\begin{aligned} \dot{v}_{31} &= -D_{v_{31}}(\mathcal{L}_{\mathrm{GD}}/d), & \dot{w}_{13} &= -D_{w_{13}}(\mathcal{L}_{\mathrm{GD}}/d), \\ \dot{v}_{33} &= -D_{v_{33}}(\mathcal{L}_{\mathrm{GD}}/d), & \dot{w}_{24} &= -d\, D_{w_{24}}(\mathcal{L}_{\mathrm{GD}}/d), \\ \dot{\mathbf{w}}_{25} &= -d\, \nabla_{\mathbf{w}_{25}}(\mathcal{L}_{\mathrm{GD}}/d), & \dot{\mathbf{W}}_B &= -d\, \nabla_{\mathbf{W}_B}(\mathcal{L}_{\mathrm{GD}}/d), \\ & & B &\in \{44, 45, 54, 55\}. \end{aligned} \tag{28}$$

For $B \in \{44, 45, 54, 55\}$,

$$-d\, D_{\mathbf{W}_B}(\mathcal{L}_{\mathrm{GD}}/d) = -D_{\mathbf{W}_B}\mathcal{L}_{\mathrm{GD}}.$$

Hence

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathcal{L}_{\mathrm{GD}}}{d} = &\left|D_{v_{31}}\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right|^2 + \left|D_{w_{13}}\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right|^2 + \left|D_{v_{33}}\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right|^2 \\ &+ d\left|D_{w_{24}}\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right|^2 + d\left\|\nabla_{\mathbf{w}_{25}}\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right\|^2 + d\sum_{B\in\{44,45,54,55\}}\left\|\nabla_{\mathbf{W}_B}\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right\|_F^2. \end{aligned} \tag{29}$$

Lemma C.3. *Along the Stage-1 flow,*

$$v_{31}^2 - w_{13}^2 - \frac{w_{24}^2 + \|\mathbf{w}_{25}\|^2}{d} = -\frac{\sigma^2}{d}, \tag{30}$$

$$dv_{33}^2 - \|w_{44}\|_F^2 - \|\mathbf{w}_{45}\|_F^2 - \|\mathbf{w}_{54}\|_F^2 - \|\mathbf{W}_{55}\|_F^2 = dv_{33}(0)^2. \tag{31}$$

$$\sum_{t=0}^{T-1}(\mathbf{w}_{25})_t - w_{24} = -\sigma. \tag{32}$$

Consequently, $v_{33}(\tau) \geq v_{33}(0) > 0$.

Proof. The loss in (21) is invariant under

$$(v_{31}, w_{13}, w_{24}, \mathbf{w}_{25}) \mapsto (e^{\zeta}v_{31}, e^{-\zeta}w_{13}, e^{-\zeta}w_{24}, e^{-\zeta}\mathbf{w}_{25}).$$

Differentiating at $\zeta = 0$ gives the Euler identity

$$v_{31}D_{v_{31}}(\mathcal{L}_{\mathrm{GD}}/d) - w_{13}D_{w_{13}}(\mathcal{L}_{\mathrm{GD}}/d) - w_{24}D_{w_{24}}(\mathcal{L}_{\mathrm{GD}}/d) - \langle \mathbf{w}_{25}, \nabla_{\mathbf{w}_{25}}(\mathcal{L}_{\mathrm{GD}}/d)\rangle = 0.$$

Substitution of the flow in (28) therefore gives the derivative of the left side of (30) equal to zero (the factor d in the ordinary-block velocities is exactly cancelled by the $1/d$ in the invariant). Its initial value is $-\sigma^2/d$. The chain rule gives $D_{w_{24}}(\mathcal{L}_{\mathrm{GD}}/d) = \sum_{t=0}^{T-1} D_{w_{24}+(\mathbf{w}_{25})_t}(\mathcal{L}_{\mathrm{GD}}/d)$ and $(\nabla_{\mathbf{w}_{25}}(\mathcal{L}_{\mathrm{GD}}/d))_t = D_{w_{24}+(\mathbf{w}_{25})_t}(\mathcal{L}_{\mathrm{GD}}/d)$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\sum_{t=0}^{T-1}(\mathbf{w}_{25})_t - w_{24}\right) = 0,$$

and its initial value is $-\sigma$. The same calculation for

$$(v_{33}, w_{44}, \mathbf{w}_{45}, \mathbf{w}_{54}, \mathbf{W}_{55}) \mapsto (e^{\zeta}v_{33}, e^{-\zeta}w_{44}, e^{-\zeta}\mathbf{w}_{45}, e^{-\zeta}\mathbf{w}_{54}, e^{-\zeta}\mathbf{W}_{55})$$

gives (31). Hence

$$v_{33}^2 = v_{33}(0)^2 + \frac{1}{d}\sum_{B\in\{44,45,54,55\}}\|\mathbf{W}_B\|_F^2 \geq v_{33}(0)^2 > 0, \qquad v_{33}(\tau) > 0.$$

□

In the next three lemmas, we use $C \geq 1$ to denote constants depending only on the fixed constants $v_{33}(0)$, σ, and the fixed regime constants.

Lemma C.4. *Let A_t, p_t, P_t be as in (22)–(23). Put*

$$M = \|\lambda\|_{d,n}^2, \qquad \mathsf{G} = \sum_{t=1}^{T-1} \left\| \lambda\{1-(1-\eta_0\lambda)^t\} - P_t[\lambda\{1-(1-\eta_0\lambda)^t\}] \right\|_{d,n}^2,$$

and set $\mathsf{z} = \sum_{t=1}^{T-1} \|P_t A_t\|_{d,n}^2$. Then, at every Stage-1 point,

$$\frac{\mathcal{L}_{\mathrm{GD}}}{d} = \frac{M}{2T}\{v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0) - \eta_0\}^2 + \frac{\mathsf{G}}{2T}(v_{31}w_{13} + \eta_0)^2 + \frac{\mathsf{z}}{2T}. \tag{33}$$

Uniformly in the stated regime,

$$M \asymp 1, \qquad C^{-1}\frac{d}{n} \leq \mathsf{G} \leq 2\frac{d}{n}. \tag{34}$$

Along the flow with $v_{33} \geq v_{33}(0) > 0$,

$$-\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathcal{L}_{\mathrm{GD}}}{d} \geq \frac{d}{T}\exp\{-CT\log(e+T)\}(d/n)^{CT}\min\{v_{31}^2, 1\}\left[\frac{\mathcal{L}_{\mathrm{GD}}}{d} - \frac{\mathsf{G}}{2T}(v_{31}w_{13}+\eta_0)^2\right]. \tag{35}$$

For $1 \leq \ell \leq t < T$, choose the fixed coefficient variations by

$$P_t[\lambda\{1-(1-\eta_0\lambda)^t\}] = \sum_{\ell=1}^{t} \delta(\widetilde{\mathbf{W}}_{55})_{\ell t}\{1-(1-\eta_0\lambda)^\ell\}. \tag{36}$$

They depend only on the fixed spectral inner product and teacher step, not on the current parameters, and satisfy $\sum_{t=1}^{T-1}\sum_{\ell=1}^{t} |\delta(\widetilde{\mathbf{W}}_{55})_{\ell t}|^2 \leq \exp\{CT\log(e+T)\}(n/d)^{CT}$.

Lemma C.5. *Under (8), $\mathcal{L}_{\mathrm{GD}}(0)/d \leq C/T$, the raw flow exists for all finite times, and $v_{33} \geq v_{33}(0) > 0$. On every deterministic interval $[0, H_d]$ with $H_d = d^{o(1)}$, the exact identity*

$$\dot{w}_{13} = -\frac{v_{31}\mathsf{G}}{T}(v_{31}w_{13}+\eta_0) + \frac{v_{31}}{dv_{33}}\sum_{t=1}^{T-1}\sum_{\ell=1}^{t}\delta(\widetilde{\mathbf{W}}_{55})_{\ell t}\frac{\mathrm{d}}{\mathrm{d}\tau}(\mathbf{W}_{55})_{\ell t} \tag{37}$$

implies, uniformly for $0 \leq s \leq \tau \leq H_d$,

$$\begin{aligned} &\left| w_{13}(\tau) - w_{13}(s) + \frac{\mathsf{G}}{T}\int_s^\tau v_{31}(r)\{v_{31}(r)w_{13}(r)+\eta_0\}\,\mathrm{d}r \right| \\ &\qquad \leq \frac{\exp\{CT\log(e+T)\}(n/d)^{CT}}{\sqrt{d}}(1+H_d/T)^{3/2} = d^{-1/2+o(1)}. \end{aligned} \tag{38}$$

Lemma C.6. *Under (8), there is a time $\tau_0 = d^{o(1)}$ such that*

$$v_{31} < 0, \qquad w_{13} > 0, \qquad v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0) \geq \frac{\eta_0}{2}, \qquad -v_{31}w_{13} \geq \frac{\eta_0}{2}. \tag{39}$$

At that time the normalized guide loss is strictly less than $9\eta_0^2\mathsf{G}/(32T)$. Consequently, for all $\tau \geq \tau_0$,

$$-v_{31}w_{13} > \frac{\eta_0}{4}, \qquad |v_{31}| \geq C^{-1} > 0. \tag{40}$$

C.2 Proof of Lemma 4.4

Proof. Equations (22)–(23) give

$$\frac{\mathcal{L}_{\mathrm{GD}}}{d} = \frac{M}{2T}(\widetilde{\theta}_1 - \eta_0)^2 + \frac{\mathsf{G}}{2T}(\widetilde{\theta}_2+\eta_0)^2 + \frac{1}{2T}\sum_{t=1}^{T-1}\|P_tA_t\|_{d,n}^2, \qquad M \asymp 1, \quad \mathsf{G} \asymp d/n.$$

The basis $\lambda, p_2, \ldots, p_t$, Lemma G.5, and (160) imply

$$\begin{aligned} &|D_{\widetilde{\theta}_1}(\mathcal{L}_{\mathrm{GD}}/d)|^2 + \|D_{\widetilde{\mathbf{\Theta}}_3}(\mathcal{L}_{\mathrm{GD}}/d)\|_F^2 \\ &\qquad \geq \frac{\exp\{-CT\log(e+T)\}(d/n)^{CT}}{T}\left[\frac{M}{2T}(\widetilde{\theta}_1-\eta_0)^2 + \frac{1}{2T}\sum_{t=1}^{T-1}\|P_tA_t\|_{d,n}^2\right]. \end{aligned} \tag{41}$$

For the remaining term, choose

$$\delta\widetilde{\theta}_1 = 0, \qquad \delta\widetilde{\theta}_2 = 1,$$
$$\lambda\,\delta(\widetilde{\mathbf{\Theta}}_3)_{1t} + \sum_{\ell=2}^{t} p_\ell\,\delta(\widetilde{\mathbf{\Theta}}_3)_{\ell t} = -P_t[\lambda p_t],$$
$$\|\delta\widetilde{\mathbf{\Theta}}_3\|_F^2 \le \exp\{CT\log(e+T)\}(n/d)^{CT} = d^{o(1)}.$$

Then

$$D(\mathcal{L}_{\mathrm{GD}}/d)[\delta\widetilde{\theta}_1, \delta\widetilde{\theta}_2, \delta\widetilde{\mathbf{\Theta}}_3] = \frac{\mathsf{G}}{T}(\widetilde{\theta}_2 + \eta_0),$$

and weighted Cauchy–Schwarz gives

$$|D_{\widetilde{\theta}_2}(\mathcal{L}_{\mathrm{GD}}/d)|^2 + d\|D_{\widetilde{\mathbf{\Theta}}_3}(\mathcal{L}_{\mathrm{GD}}/d)\|_F^2 \ge \frac{(\mathsf{G}/T)^2(\widetilde{\theta}_2+\eta_0)^2}{1 + d^{-1}\|\delta\widetilde{\mathbf{\Theta}}_3\|_F^2} \ge c\frac{d/n}{T}\frac{\mathsf{G}}{2T}(\widetilde{\theta}_2+\eta_0)^2.$$

Since $d\exp\{-CT\log(e+T)\}(d/n)^{CT} \ge c\,d/n$,

$$d|D_{\widetilde{\theta}_1}(\mathcal{L}_{\mathrm{GD}}/d)|^2 + |D_{\widetilde{\theta}_2}(\mathcal{L}_{\mathrm{GD}}/d)|^2 + d\|D_{\widetilde{\mathbf{\Theta}}_3}(\mathcal{L}_{\mathrm{GD}}/d)\|_F^2 \ge c\frac{d/n}{T}\frac{\mathcal{L}_{\mathrm{GD}}}{d}.$$

Multiplication by d proves Lemma 4.4. □

C.3 Proof of Lemma 4.5

Proof. Lemma C.6 and (31) give $|v_{31}| \ge c$ and $v_{33} \ge c$ for $\tau \ge \tau_0$. For arbitrary auxiliary variations, set

$$\begin{aligned}
&\delta v_{31} = \delta v_{33} = \delta w_{24} = \delta w_{44} = 0, \qquad \delta\mathbf{w}_{45} = \mathbf{0}, \quad \delta\mathbf{w}_{54} = \mathbf{0},\\
&\delta w_{13} = \frac{\delta\widetilde{\theta}_2}{v_{31}}, \qquad \delta(\mathbf{w}_{25})_0 = \frac{\delta\widetilde{\theta}_1}{v_{31}},\\
&\delta(\mathbf{w}_{25})_t = \frac{\delta(\widetilde{\mathbf{\Theta}}_3)_{1t}}{v_{31}}, \qquad 1 \le t < T,\\
&\delta(\mathbf{W}_{55})_{\ell t} = \begin{cases} \delta(\widetilde{\mathbf{\Theta}}_3)_{\ell t}/v_{33}, & 2 \le \ell \le t < T,\\ 0, & \text{otherwise.}\end{cases}
\end{aligned} \tag{42}$$

Direct differentiation of (20) gives

$$\begin{aligned}
&D\widetilde{\theta}_1[\delta\bar{\boldsymbol{\theta}}] = \delta\widetilde{\theta}_1, \qquad D\widetilde{\theta}_2[\delta\bar{\boldsymbol{\theta}}] = \delta\widetilde{\theta}_2, \quad D\widetilde{\mathbf{\Theta}}_3[\delta\bar{\boldsymbol{\theta}}] = \delta\widetilde{\mathbf{\Theta}}_3,\\
&\|\delta\bar{\boldsymbol{\theta}}\|^2 = d(\delta w_{13})^2 + \|\delta\mathbf{w}_{25}\|^2 + \|\delta\mathbf{W}_{55}\|_F^2\\
&\qquad \le C\left[|\delta\widetilde{\theta}_1|^2 + d|\delta\widetilde{\theta}_2|^2 + \|\delta\widetilde{\mathbf{\Theta}}_3\|_F^2\right].
\end{aligned}$$

Duality yields

$$\mathbf{J}_1\mathbf{J}_1^{\mathsf{T}} \succeq c\,\mathrm{diag}(1, d^{-1}, \mathbf{I}).$$

□

Lemma C.7. *For $\tau \ge \tau_0$,*

$$-\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{L}_{\mathrm{GD}} \ge c\frac{d/n}{T}\mathcal{L}_{\mathrm{GD}}. \tag{43}$$

Proof. By the chain rule and Lemmas 4.4–4.5,

$$-\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{L}_{\mathrm{GD}} = \|D_{\bar{\boldsymbol{\theta}}}\mathcal{L}_{\mathrm{GD}}\|^2 \ge c\left[|D_{\widetilde{\theta}_1}\mathcal{L}_{\mathrm{GD}}|^2 + d^{-1}|D_{\widetilde{\theta}_2}\mathcal{L}_{\mathrm{GD}}|^2 + \|D_{\widetilde{\mathbf{\Theta}}_3}\mathcal{L}_{\mathrm{GD}}\|_F^2\right] \ge c\frac{d/n}{T}\mathcal{L}_{\mathrm{GD}}.$$

□

Proof of Theorem 3.1. At initialization, $v_{31} = 0$ and $w_{44} = \mathbf{w}_{45} = \mathbf{w}_{54} = \mathbf{W}_{55} = 0$, so (21) becomes

$$\frac{\eta_0^2}{2Td} \sum_{t=0}^{T-1} \mathbb{E}\ \mathrm{tr}\{\mathbf{S}^2(\mathbf{I} - \eta_0 \mathbf{S})^{2t}\}.$$

To make this bound uniform in the whole step interval, first replace the finite-Wishart moments in each summand by the corresponding MP moments. The support estimate (187), together with $\eta_0 \le 1$, gives

$$\frac{\eta_0^2}{2T} \sum_{t=0}^{T-1} \mathbb{E}_{\lambda \sim \mu_{d/n}} \left[\lambda^2 (1 - \eta_0 \lambda)^{2t}\right] \le \frac{C}{T} \sum_{t=0}^{T-1} (17/20)^{2t} \le C/T.$$

For each t, expanding the polynomial and applying (188) gives, uniformly for $0 \le t < T$ and for every $\eta_0 \in [1/3, 9/10]$,

$$\left| \frac{1}{d} \mathbb{E}\, \mathrm{tr}[\mathbf{S}^2(\mathbf{I} - \eta_0 \mathbf{S})^{2t}] - \mathbb{E}_{\lambda \sim \mu_{d/n}}[\lambda^2 (1 - \eta_0 \lambda)^{2t}] \right| \le d^{-1+o(1)}.$$

Averaging this error over t, and using $\eta_0 \le 1$, proves

$$\frac{\mathcal{L}_{\mathrm{GD}}(0)}{d} \le C/T$$

with an absolute constant independent of the selected step. No compact support assertion is being made for the finite-Wishart law.

The solution is global. Indeed, integrating (29) over any finite interval $[0, s]$ bounds

$$\int_0^s \Big[\dot{v}_{31}^2 + \dot{w}_{13}^2 + \dot{v}_{33}^2 + \frac{1}{d} \{ \dot{w}_{24}^2 + \|\dot{\mathbf{w}}_{25}\|^2 + \|\dot{w}_{44}\|_F^2 + \|\dot{\mathbf{w}}_{45}\|_F^2 + \|\dot{\mathbf{w}}_{54}\|_F^2 + \|\dot{\mathbf{W}}_{55}\|_F^2 \} \Big] \, \mathrm{d}\tau \le \frac{\mathcal{L}_{\mathrm{GD}}(0)}{d}.$$

Cauchy-Schwarz bounds every raw displacement on $[0, s]$. A finite maximal existence time would therefore have a finite raw limit, and local existence for the polynomial vector field would extend the solution past that time, a contradiction. By Lemma C.6, let $\tau_0 \le d^{o(1)}$ be the time supplied by (39). Lemma C.7 and Gronwall's inequality then give

$$\mathcal{L}_{\mathrm{GD}}(\tau) \le Cd \exp\left[-c \frac{d/n}{T} (\tau - \tau_0) \right], \qquad \tau \ge \tau_0.$$

Taking $\epsilon_{\mathrm{GD}} = d^{-1/4}$ gives the Stage-1 time bound $d^{o(1)} + O(T \log d)$ used in Theorem 3.1. □

D Proofs of Lemmas 4.6–4.8

D.1 Proof of Lemma 4.6

On the rotationally invariant subspace write $\mathbf{V}_{31} = v_{31}\mathbf{I}_d$, $\mathbf{W}_{13} = w_{13}\mathbf{I}_d$, and $\mathbf{V}_{33} = v_{33}\mathbf{I}_d$. For $B \in \{44, 45, 54, 55\}$, let $\mathbf{W}_B$ denote the corresponding submatrix. The coefficient contributed by $w_{44}, \mathbf{w}_{45}, \mathbf{w}_{54}, \mathbf{W}_{55}$ is

$$(\widetilde{\mathbf{W}}_{55})_{\ell t} = v_{33} \big(w_{44} + \mathbf{w}_{45} \mathbf{e}_t + \mathbf{e}_\ell^\top \mathbf{w}_{54} + \mathbf{e}_\ell^\top \mathbf{W}_{55} \mathbf{e}_t \big). \tag{44}$$

Set

$$\widetilde{\theta}_1 = v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0), \qquad \widetilde{\theta}_2 = v_{31} w_{13}. \tag{45}$$

For $1 \le t < T$ and $\widetilde{\theta}_1 \ne 0$, define the Stage-2 coefficient array by

$$\begin{aligned} (\widetilde{\mathbf{\Theta}}_4)_{1t} &:= \frac{v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t)}{\widetilde{\theta}_1} + (\widetilde{\mathbf{W}}_{55})_{1t}, \\ (\widetilde{\mathbf{\Theta}}_4)_{\ell t} &:= (\widetilde{\mathbf{W}}_{55})_{\ell t}, \qquad 2 \le \ell \le t < T. \end{aligned} \tag{46}$$

$$(\widetilde{\mathbf{\Theta}}_4)_{\ell t} = \mathbf{e}_\ell^\mathsf{T} \widetilde{\mathbf{\Theta}}_4 \mathbf{e}_t, \qquad 0 \le \ell, t < T,$$
$$\widetilde{\mathbf{\Theta}}_4 = \sum_{t=1}^{T-1} \sum_{\ell=1}^{t} (\widetilde{\mathbf{\Theta}}_4)_{\ell t} \mathbf{e}_\ell \mathbf{e}_t^\mathsf{T}.$$

For a polynomial $R(\lambda) = \sum_{j=0}^{m} r_j \lambda^j$, write

$$[\lambda^j] R = r_j, \qquad \mathrm{coeff}_{1:t}(R) = (r_1, \dots, r_t)^\mathsf{T},$$
$$\mathrm{coeff}_m(R) = (r_0, \dots, r_m)^\mathsf{T}, \qquad \|R\|_{\mathrm{coeff}} = \|\,\mathrm{coeff}_m(R)\|_2.$$

Coefficient vectors are padded by zeros when needed.

Lemma D.1. *At every raw parameter value with $\widetilde{\theta}_1 \neq 0$, including vanishing values of w_{13} or v_{33},*

$$p_0 = 0, \quad p_1 = \widetilde{\theta}_1 \lambda, \qquad p_{t+1} = (1 + \widetilde{\theta}_2 \lambda) p_t + \sum_{\ell=1}^{t} p_\ell \, (\widetilde{\mathbf{\Theta}}_4)_{\ell t}. \tag{47}$$

Consequently the normalized population loss is a polynomial function of the displayed scalar and coefficient arrays. If $\widetilde{\theta}_1 \widetilde{\theta}_2 \neq 0$, then

$$[\lambda^t] p_t = \widetilde{\theta}_1 \widetilde{\theta}_2^{t-1}. \tag{48}$$

For fixed nonzero $\widetilde{\theta}_1, \widetilde{\theta}_2$, the map from the coefficients

$$(\widetilde{\mathbf{\Theta}}_4)_{1t}, \quad (\widetilde{\mathbf{\Theta}}_4)_{2t}, \dots, (\widetilde{\mathbf{\Theta}}_4)_{tt}$$

in the t-th column of $\widetilde{\mathbf{\Theta}}_4$ to the lower coefficients of p_{t+1} is an affine bijection.

Proof. The block multiplication gives the recurrence

$$p_{t+1} = (1 + \widetilde{\theta}_2 \lambda) p_t + v_{31}(w_{24} + \mathbf{w}_{25} \mathbf{e}_t) \lambda + (\widetilde{\mathbf{W}}_{55})_{1t} p_1 + \sum_{\ell=2}^{t} (\widetilde{\mathbf{W}}_{55})_{\ell t} p_\ell.$$

Since $p_1 = \widetilde{\theta}_1 \lambda$, $v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t)\lambda + (\widetilde{\mathbf{W}}_{55})_{1t} p_1 = \left[\frac{v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t)}{\widetilde{\theta}_1} + (\widetilde{\mathbf{W}}_{55})_{1t} \right] p_1$, proving (47). Only $\widetilde{\theta}_2 \lambda p_t$ creates the next degree, which proves (48). The polynomials $p_1, p_2, \dots, p_t$ have successive nonzero leading coefficients and form a triangular basis of all zero-constant polynomials of degree at most t. This proves the last statement. □

The normalized raw Frobenius metric on the invariant subspace is represented directly by

$$\begin{aligned} &(\delta v_{31})^2 + (\delta w_{13})^2 + (\delta v_{33})^2 \\ &\quad + \frac{1}{d} \big\{ (\delta w_{24})^2 + \|\delta \mathbf{w}_{25}\|^2 + \|\delta w_{44}\|_F^2 + \|\delta \mathbf{w}_{45}\|_F^2 \\ &\qquad + \|\delta \mathbf{w}_{54}\|_F^2 + \|\delta \mathbf{W}_{55}\|_F^2 \big\}. \end{aligned} \tag{49}$$

Consequently, gradient flow for $\mathcal{L}/d$ is the coordinate system

$$\begin{aligned} &\dot{v}_{31} = -D_{v_{31}}(\mathcal{L}/d), && \dot{w}_{13} = -D_{w_{13}}(\mathcal{L}/d), && \dot{v}_{33} = -D_{v_{33}}(\mathcal{L}/d), \\ &\dot{w}_{24} = -d\, D_{w_{24}}(\mathcal{L}/d), && \dot{\mathbf{w}}_{25} = -d\, \nabla_{\mathbf{w}_{25}}(\mathcal{L}/d), \\ &\dot{w}_{44} = -d\, \nabla_{w_{44}}(\mathcal{L}/d), && \dot{\mathbf{w}}_{45} = -d\, \nabla_{\mathbf{w}_{45}}(\mathcal{L}/d), \\ &\dot{\mathbf{w}}_{54} = -d\, \nabla_{\mathbf{w}_{54}}(\mathcal{L}/d), && \dot{\mathbf{W}}_{55} = -d\, \nabla_{\mathbf{W}_{55}}(\mathcal{L}/d). \end{aligned} \tag{50}$$

Lemma D.2. *Along (50),*

$$\begin{aligned} &v_{31}^2 - w_{13}^2 - \frac{1}{d} \left(w_{24}^2 + \|\mathbf{w}_{25}\|^2 \right) \\ &\quad = v_{31}(0)^2 - w_{13}(0)^2 - \frac{1}{d} \left(w_{24}(0)^2 + \|\mathbf{w}_{25}(0)\|^2 \right), \end{aligned} \tag{51}$$

$$\begin{aligned} &d v_{33}^2 - \|w_{44}\|_F^2 - \|\mathbf{w}_{45}\|_F^2 - \|\mathbf{w}_{54}\|_F^2 - \|\mathbf{W}_{55}\|_F^2 \\ &\quad = d v_{33}(0)^2 - \sum_{B \in \{44,45,54,55\}} \|\mathbf{W}_B(0)\|_F^2 = d (v_{33}^{\mathrm{init}})^2. \end{aligned} \tag{52}$$

$$\sum_{t=0}^{T-1} (\mathbf{w}_{25})_t - w_{24} = -\sigma. \tag{53}$$

Here time zero is the Stage-2 handoff, and $v_{33}^{\mathrm{init}} > 0$ is the scalar at the original initialization (8). In particular, $v_{33}(\tau) \geq v_{33}^{\mathrm{init}} > 0$ throughout Stage 2.

Proof. The transformation

$$v_{31} \mapsto e^{\zeta} v_{31}, \qquad w_{13} \mapsto e^{-\zeta} w_{13}, \qquad w_{24} \mapsto e^{-\zeta} w_{24}, \qquad \mathbf{w}_{25} \mapsto e^{-\zeta} \mathbf{w}_{25}$$

leaves all products $v_{31} w_{13}, v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t)$, hence every rollout and the loss, unchanged. Differentiating at $\zeta = 0$ and substituting the gradient-flow equations gives

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(v_{31}^2 - w_{13}^2 - \frac{1}{d}(w_{24}^2 + \|\mathbf{w}_{25}\|^2)\right) = 0. \tag{54}$$

This is (51). It is continuous across the Stage-1/Stage-2 switch because the raw parameters are unchanged, so its right side is the fixed value determined at initialization.

The Stage-2 objective depends on $(w_{24}, \mathbf{w}_{25})$ only through the vector $(w_{24} + (\mathbf{w}_{25})_t)_{t=0}^{T-1}$. Consequently

$$D_{w_{24}}(\mathcal{L}/d) = \sum_{t=0}^{T-1} D_{w_{24}+(\mathbf{w}_{25})_t}(\mathcal{L}/d), \qquad D_{(\mathbf{w}_{25})_t}(\mathcal{L}/d) = D_{w_{24}+(\mathbf{w}_{25})_t}(\mathcal{L}/d),$$

and the normalized flow gives

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\sum_{t=0}^{T-1}(\mathbf{w}_{25})_t - w_{24}\right) = 0.$$

Its value is $-\sigma$ at the original initialization and therefore remains $-\sigma$ through the Stage-1/Stage-2 switch.

The quantities $w_{44}, \mathbf{w}_{45}, \mathbf{w}_{54}, \mathbf{W}_{55}$ occur in (44) only through the coefficients $(\widetilde{\mathbf{W}}_{55})_{\ell t}$, each of which contains a factor v_{33}. Differentiating the loss with respect to each entry and using (50) gives

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\tau}\left[d v_{33}^2 - \|w_{44}\|_F^2 - \|\mathbf{w}_{45}\|_F^2 - \|\mathbf{w}_{54}\|_F^2 - \|\mathbf{W}_{55}\|_F^2\right] \\
&\quad = -2d v_{33} D_{v_{33}}(\mathcal{L}/d) + 2d\langle w_{44}, \nabla_{w_{44}}(\mathcal{L}/d)\rangle_F + 2d\langle \mathbf{w}_{45}, \nabla_{\mathbf{w}_{45}}(\mathcal{L}/d)\rangle_F \\
&\qquad + 2d\langle \mathbf{w}_{54}, \nabla_{\mathbf{w}_{54}}(\mathcal{L}/d)\rangle_F + 2d\langle \mathbf{W}_{55}, \nabla_{\mathbf{W}_{55}}(\mathcal{L}/d)\rangle_F = 0,
\end{aligned}$$

because each term in (44) is homogeneous of degree one in v_{33} and degree one in exactly one of the four blocks $w_{44}, \mathbf{w}_{45}, \mathbf{w}_{54}, \mathbf{W}_{55}$. The invariant is inherited continuously from Stage 1, where these four blocks start at zero, so its right side is $d(v_{33}^{\mathrm{init}})^2$ without any reset at the switch. □

D.2 Proof of Lemma 4.7

Lemma D.3. *Set*

$$\widetilde{\theta}_1 := v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0), \qquad \widetilde{\theta}_2 := v_{31} w_{13}. \tag{55}$$

Assume $\widetilde{\theta}_1 \widetilde{\theta}_2 \neq 0$. Then

$$p_t(0) = 0, \qquad \deg p_t = t, \qquad [\lambda^t] p_t = \widetilde{\theta}_1 \widetilde{\theta}_2^{t-1}. \tag{56}$$

Conversely, for fixed nonzero $\widetilde{\theta}_1, \widetilde{\theta}_2$, every polynomial sequence satisfying (56) is realized by a unique array of the auxiliary recurrence coefficients

$$(\widetilde{\mathbf{\Theta}}_4)_{1t}, \quad (\widetilde{\mathbf{\Theta}}_4)_{2t}, \ldots, (\widetilde{\mathbf{\Theta}}_4)_{tt} \qquad (1 \leq t < T).$$

Proof. **(i)** The first step is $p_1 = \widetilde{\theta}_1 \lambda$. In (16), only $\widetilde{\theta}_2 \lambda p_t$ can create degree $t+1$, since

$$\deg\left(\sum_{\ell=1}^{t}(\widetilde{\mathbf{W}}_{55})_{\ell t} p_\ell\right) \leq t.$$

This proves (56) by induction.

(ii) For the converse, nonzero leading coefficients and increasing degrees imply

$$\operatorname{span}\{p_1, \dots, p_t\} = \{q : q(0) = 0,\ \deg q \le t\}.$$

By (56), the unique expansion

$$p_{t+1} - (1 + \widetilde{\theta}_2\lambda)p_t = \sum_{\ell=1}^{t} p_\ell \, (\widetilde{\mathbf{\Theta}}_4)_{\ell t}$$

determines the t-th coefficient column. Induction over t proves the claim. □

For $1 \le t < T$, define the triangular coefficient array $\widetilde{\theta}_5$ by

$$p_{t+1}(\lambda) = \widetilde{\theta}_1 \widetilde{\theta}_2^t \lambda^{t+1} + \sum_{j=1}^{t} (\widetilde{\theta}_5)_{jt} \lambda^j. \tag{57}$$

The change of basis from $p_1(\lambda), p_2(\lambda), \dots, p_t(\lambda)$ to $\lambda, \lambda^2, \dots, \lambda^t$ is triangular, with diagonal $\widetilde{\theta}_1, \widetilde{\theta}_1\widetilde{\theta}_2, \dots, \widetilde{\theta}_1\widetilde{\theta}_2^{t-1}$. Equivalently,

$$(\widetilde{\theta}_5)_{jt} = [\lambda^j] \left\{ (1 + \widetilde{\theta}_2\lambda)p_t(\lambda) + \sum_{\ell=1}^{t} p_\ell(\lambda) \, (\widetilde{\mathbf{\Theta}}_4)_{\ell t} \right\}, \qquad 1 \le j \le t < T. \tag{58}$$

Let

$$\begin{aligned} \mathbf{H}_t &:= \left[\left\langle \lambda^i, \lambda^j \right\rangle_{d,n} \right]_{1 \le i,j \le t} \succ 0, \qquad 1 \le t < T, \\ \mathbf{H} &:= \operatorname{diag}(\mathbf{H}_1, \dots, \mathbf{H}_{T-1}), \\ \widetilde{\theta}_5^\star &:= \arg\min_{\widetilde{\theta}_5} \left\{ \mathcal{L} : \widetilde{\theta}_1, \widetilde{\theta}_2 \text{ fixed} \right\}. \end{aligned} \tag{59}$$

For fixed $(\widetilde{\theta}_1, \widetilde{\theta}_2)$, completing the square gives

$$\begin{aligned} \frac{\mathcal{L} - \mathcal{L}^\star}{d} = \frac{1}{2T} \Bigg[& (\widetilde{\theta}_5 - \widetilde{\theta}_5^\star)^\top \mathbf{H} (\widetilde{\theta}_5 - \widetilde{\theta}_5^\star) \\ & + C_{\nu_{d,n}}(-\widetilde{\theta}_2) \left(\widetilde{\theta}_1 - \frac{B_{\nu_{d,n}}(-\widetilde{\theta}_2)}{C_{\nu_{d,n}}(-\widetilde{\theta}_2)} \right)^2 \\ & + \frac{B_{\nu_{d,n}}(-\widetilde{\theta}_2^\star)^2}{C_{\nu_{d,n}}(-\widetilde{\theta}_2^\star)} - \frac{B_{\nu_{d,n}}(-\widetilde{\theta}_2)^2}{C_{\nu_{d,n}}(-\widetilde{\theta}_2)} \Bigg]. \end{aligned} \tag{60}$$

The MP expression is obtained by replacing $B_{\nu_{d,n}}, C_{\nu_{d,n}}$ with B_T, C_T. It is not the finite-Wishart objective: already the first-step second moments are $1 + d/n$ and $1 + d/n + 1/n$, respectively.

$$\begin{aligned} \frac{1}{2T} \sum_{t=1}^{T} \left[\begin{array}{l} \dfrac{\left(\frac{d}{n}\right)^{t-1}}{\sum_{r=0}^{t-1} \left(\frac{d}{n}\right)^r} \left(1 - \widetilde{\theta}_1(-\widetilde{\theta}_2)^{t-1}\right)^2 \\ + \left(\dfrac{d}{n}\right)^t \widetilde{\theta}_1^2 (-\widetilde{\theta}_2)^{2t-2} \end{array} \right] &= \frac{C_T(-\widetilde{\theta}_2)}{2T} \left(\widetilde{\theta}_1 - \frac{B_T(-\widetilde{\theta}_2)}{C_T(-\widetilde{\theta}_2)} \right)^2 \\ &\quad + \frac{1}{2T} \sum_{k=0}^{T-1} \frac{\left(\frac{d}{n}\right)^k}{\sum_{r=0}^{k} \left(\frac{d}{n}\right)^r} - \frac{1}{2T} \frac{B_T(-\widetilde{\theta}_2)^2}{C_T(-\widetilde{\theta}_2)}. \end{aligned} \tag{61}$$

For the rollout polynomials, write

$$R_t(\lambda) = p_t(\lambda) - 1, \qquad \mathbf{w}_t - \mathbf{w}^\star = R_t(\mathbf{S})\mathbf{w}^\star.$$

For each t, R_t^* denotes the degree-at-most-t polynomial with $R_t^*(0) = -1$ minimizing $\|R_t^*\|_{\mathrm{MP}}$.

Lemma D.4. *After minimizing every coefficient column under the MP law, the normalized objective is the expression displayed in (61). For fixed $(-\widetilde{\theta}_2)$, its unique minimizing $\widetilde{\theta}_1$ is*

$$\widetilde{\theta}_1^{\mathrm{MP}}(-\widetilde{\theta}_2) = \frac{\displaystyle\sum_{k=0}^{T-1} \frac{\left(\frac{d}{n}\right)^k}{\sum_{r=0}^{k}\left(\frac{d}{n}\right)^r}(-\widetilde{\theta}_2)^k}{\displaystyle\sum_{k=0}^{T-1}\left(\frac{\left(\frac{d}{n}\right)^k}{\sum_{r=0}^{k}\left(\frac{d}{n}\right)^r} + \left(\frac{d}{n}\right)^{k+1}\right)(-\widetilde{\theta}_2)^{2k}}. \tag{62}$$

Proof. Insert $\ell_t = (-1)^{t-1}\widetilde{\theta}_1(-\widetilde{\theta}_2)^{t-1}$ from (56) into Lemma G.7. At each step,

$$R_t(\lambda) = (1 - \widetilde{\theta}_1(-\widetilde{\theta}_2)^{t-1})R^*_{t-1}(\lambda) + (-1)^{t-1}\widetilde{\theta}_1(-\widetilde{\theta}_2)^{t-1}\pi_t(\lambda). \tag{63}$$

Substitution into (165) and summation over $t = 1, \dots, T$ give (61). The coefficient of $\widetilde{\theta}_1^2$ in that expression is at least $(1 + \left(\frac{d}{n}\right))/(2T) > 0$. Hence

$$\begin{aligned} 0 = &-\sum_{k=0}^{T-1} \frac{\left(\frac{d}{n}\right)^k}{\sum_{r=0}^{k}\left(\frac{d}{n}\right)^r}(-\widetilde{\theta}_2)^k \\ &+ \widetilde{\theta}_1 \sum_{k=0}^{T-1}\left(\frac{\left(\frac{d}{n}\right)^k}{\sum_{r=0}^{k}\left(\frac{d}{n}\right)^r} + \left(\frac{d}{n}\right)^{k+1}\right)(-\widetilde{\theta}_2)^{2k}. \end{aligned}$$

Solving this stationarity equation gives (62) and proves uniqueness. □

Define

$$B_T(-\widetilde{\theta}_2) = \sum_{k=0}^{T-1} \frac{\left(\frac{d}{n}\right)^k}{\sum_{r=0}^{k}\left(\frac{d}{n}\right)^r}(-\widetilde{\theta}_2)^k, \qquad C_T(-\widetilde{\theta}_2) = \sum_{k=0}^{T-1} \frac{\left(\frac{d}{n}\right)^k \sum_{r=0}^{k+1}\left(\frac{d}{n}\right)^r}{\sum_{r=0}^{k}\left(\frac{d}{n}\right)^r}(-\widetilde{\theta}_2)^{2k}. \tag{64}$$

Expanding (61) and minimizing the strictly convex $\widetilde{\theta}_1$-coordinate gives

$$\begin{aligned} \min_{\widetilde{\theta}_1} \frac{1}{2T}\sum_{t=1}^{T}\left[\begin{array}{c} \frac{\left(\frac{d}{n}\right)^{t-1}}{\sum_{r=0}^{t-1}\left(\frac{d}{n}\right)^r}\left(1 - \widetilde{\theta}_1(-\widetilde{\theta}_2)^{t-1}\right)^2 \\ + \left(\frac{d}{n}\right)^t \widetilde{\theta}_1^2(-\widetilde{\theta}_2)^{2t-2} \end{array}\right] &= \frac{1}{2T}\sum_{k=0}^{T-1}\frac{\left(\frac{d}{n}\right)^k}{\sum_{r=0}^{k}\left(\frac{d}{n}\right)^r} - \frac{1}{2T}\frac{B_T(-\widetilde{\theta}_2)^2}{C_T(-\widetilde{\theta}_2)}, \\ \widetilde{\theta}_1^{\mathrm{MP}}(-\widetilde{\theta}_2) &= \frac{B_T(-\widetilde{\theta}_2)}{C_T(-\widetilde{\theta}_2)}. \end{aligned} \tag{65}$$

Lemma D.5. *For all sufficiently large d in the fixed-aspect-ratio regime (17),*

$$\frac{B_T(-\widetilde{\theta}_2)^2}{C_T(-\widetilde{\theta}_2)}$$

has exactly one critical point on $(-\widetilde{\theta}_2) > 0$. It lies in

$$\frac{1 + d/n}{1 + d/n + (d/n)^2} < (-\widetilde{\theta}_2^{\mathrm{MP}}) < 1, \qquad \frac{\mathrm{d}^2}{\mathrm{d}(-\widetilde{\theta}_2)^2}\frac{B_T(-\widetilde{\theta}_2)^2}{C_T(-\widetilde{\theta}_2)}\bigg|_{(-\widetilde{\theta}_2)=(-\widetilde{\theta}_2^{\mathrm{MP}})} < 0.$$

Consequently $(\widetilde{\theta}_1^{\mathrm{MP}}, (-\widetilde{\theta}_2^{\mathrm{MP}}))$ is the unique positive critical point and global minimizer of the normalized objective in (61), and

$$(-\widetilde{\theta}_2^{\mathrm{MP}}) \le 1 - \frac{(d/n)^2}{128}.$$

Proof. The MP law is the μ-case of Lemma G.8. Its coefficient formula (168) is exactly (64), so that lemma gives the unique nondegenerate maximum and the upper bound in the statement. For the lower bound, direct pairing of the (i,j) and (j,i) terms gives

$$\begin{aligned}&\frac{(-\widetilde{\theta}_2)}{2}\{2B_T'(-\widetilde{\theta}_2)C_T(-\widetilde{\theta}_2) - B_T(-\widetilde{\theta}_2)C_T'(-\widetilde{\theta}_2)\}\\ &= \sum_{0\le j<i<T}(i-j)\frac{(d/n)^{i+j}(-\widetilde{\theta}_2)^{i+2j}}{\left(\sum_{s=0}^{i}(d/n)^s\right)\left(\sum_{s=0}^{j}(d/n)^s\right)}\\ &\quad\cdot\left\{\sum_{s=0}^{j+1}(d/n)^s - (-\widetilde{\theta}_2)^{i-j}\sum_{s=0}^{i+1}(d/n)^s\right\}. \end{aligned} \tag{66}$$

For every $k \ge 1$,

$$\frac{\sum_{s=0}^{k}(d/n)^s}{\sum_{s=0}^{k+1}(d/n)^s} \ge \frac{1+d/n}{1+d/n+(d/n)^2},$$

because, after cross multiplication, the difference is

$$\left(\frac{d}{n}\right)^2\left\{\sum_{s=0}^{k}(d/n)^s - (d/n)^{k-1}(1+d/n)\right\} \ge 0,$$

with equality only for $k=1$. Hence, at $(-\widetilde{\theta}_2) = (1+d/n)/(1+d/n+(d/n)^2)$,

$$\frac{\sum_{s=0}^{j+1}(d/n)^s}{\sum_{s=0}^{i+1}(d/n)^s} = \prod_{k=j+1}^{i}\frac{\sum_{s=0}^{k}(d/n)^s}{\sum_{s=0}^{k+1}(d/n)^s} \ge (-\widetilde{\theta}_2)^{i-j}.$$

Every summand in (66) is nonnegative, and the $(i,j)=(2,0)$ summand is positive. Thus

$$\frac{\mathrm{d}}{\mathrm{d}(-\widetilde{\theta}_2)}\frac{B_T(-\widetilde{\theta}_2)^2}{C_T(-\widetilde{\theta}_2)}\bigg|_{(-\widetilde{\theta}_2)=(1+d/n)/(1+d/n+(d/n)^2)} > 0,$$

which places the unique critical point strictly to the right of the displayed endpoint. Finally, the strict convexity in $\widetilde{\theta}_1$ from (61) lifts it to the stated positive minimizer of the right-hand side of (61). To exclude the negative branch, note that C_T is even and the even and odd parts of B_T are both nonzero with nonnegative coefficients. Hence, by strict triangle inequality,

$$|B_T(-(-\widetilde{\theta}_2))| < B_T(-\widetilde{\theta}_2) \qquad \text{for } (-\widetilde{\theta}_2) > 0.$$

Thus the minimized function at a negative argument is strictly larger than at the corresponding positive argument, so the positive minimizer is global. □

Proof of Lemma 4.7. The notation $D_{\widetilde{\theta}_i}\big|_{\widetilde{\theta}_5}$ holds $\widetilde{\theta}_5$ fixed; the notation $D_{\widetilde{\theta}_i}\big|_{\widetilde{\mathbf{\Theta}}_4}$ holds $\widetilde{\mathbf{\Theta}}_4$ fixed. Lemmas E.2 and F.5 give, throughout Stage 2,

$$\frac{1}{100} < \widetilde{\theta}_1 < 2, \qquad \frac{1}{14} < -\widetilde{\theta}_2 < \frac{38}{13}, \qquad \frac{d}{n}(-\widetilde{\theta}_2)^2 < \frac{722}{845}, \qquad \frac{\mathcal{L}}{d} \le \frac{C}{T}.$$

The coefficient estimates below require only these scalar and loss bounds.

(i) On this interval, (169) and the geometric MP series give

$$C_{\nu_{d,n}}(-\widetilde{\theta}_2) \ge c, \qquad \sum_{j=0}^{2}\left(|B_{\nu_{d,n}}^{(j)}(-\widetilde{\theta}_2)| + |C_{\nu_{d,n}}^{(j)}(-\widetilde{\theta}_2)|\right) \le C.$$

The ratio d/n ranges over a fixed compact subset of $(0, 1/10]$. Lemma G.2, uniform convergence of the first two derivatives, and Lemma G.8 imply

$$\left|\left(\frac{B_{\nu_{d,n}}^2}{C_{\nu_{d,n}}}\right)'(-\widetilde{\theta}_2)\right|^2 \ge c\left\{\frac{B_{\nu_{d,n}}(-\widetilde{\theta}_2^\star)^2}{C_{\nu_{d,n}}(-\widetilde{\theta}_2^\star)} - \frac{B_{\nu_{d,n}}(-\widetilde{\theta}_2)^2}{C_{\nu_{d,n}}(-\widetilde{\theta}_2)}\right\}.$$

Here nondegeneracy gives the bound near the unique maximum, and the derivative is uniformly bounded away from zero on the remaining compact set. Differentiating the last two terms of (60) gives

$$TD_{\widetilde{\theta}_1}\frac{\min_{\widetilde{\theta}_5}\mathcal{L}}{d} = C_{\nu_{d,n}}(-\widetilde{\theta}_2)\widetilde{\theta}_1 - B_{\nu_{d,n}}(-\widetilde{\theta}_2),$$
$$2TD_{-\widetilde{\theta}_2}\frac{\min_{\widetilde{\theta}_5}\mathcal{L}}{d} = C'_{\nu_{d,n}}(-\widetilde{\theta}_2)\widetilde{\theta}_1^2 - 2B'_{\nu_{d,n}}(-\widetilde{\theta}_2)\widetilde{\theta}_1,$$

and hence

$$\left|\widetilde{\theta}_1 - \frac{B_{\nu_{d,n}}(-\widetilde{\theta}_2)}{C_{\nu_{d,n}}(-\widetilde{\theta}_2)}\right|^2 + \left|\left(\frac{B^2_{\nu_{d,n}}}{C_{\nu_{d,n}}}\right)'(-\widetilde{\theta}_2)\right|^2 \le CT^2\sum_{i=1}^{2}\left|D_{\widetilde{\theta}_i}\frac{\min_{\widetilde{\theta}_5}\mathcal{L}}{d}\right|^2.$$

Consequently,

$$\sum_{i=1}^{2}\left|D_{\widetilde{\theta}_i}\frac{\min_{\widetilde{\theta}_5}\mathcal{L}}{d}\right|^2 \ge c\frac{d/n}{T}\frac{\min_{\widetilde{\theta}_5}\mathcal{L}-\mathcal{L}^\star}{d}. \tag{67}$$

(ii) The normal equations and their derivatives are, for $1 \le t < T$ and $i = 1, 2$,

$$\begin{aligned}\mathbf{H}_t(\widetilde{\theta}^\star_5)_{1:t,t} &= \left(\left\langle \lambda^j, 1-\widetilde{\theta}_1\widetilde{\theta}_2^t\lambda^{t+1}\right\rangle_{d,n}\right)_{j=1}^{t},\\ \mathbf{H}_t D_{\widetilde{\theta}_i}(\widetilde{\theta}^\star_5)_{1:t,t} &= -D_{\widetilde{\theta}_i}(\widetilde{\theta}_1\widetilde{\theta}_2^t)\left(\langle\lambda^j,\lambda^{t+1}\rangle_{d,n}\right)_{j=1}^{t}.\end{aligned} \tag{68}$$

Thus the exact derivatives of (60) satisfy

$$\begin{aligned}D_{\widetilde{\theta}_5}\frac{\mathcal{L}}{d} &= \frac{1}{T}\mathbf{H}(\widetilde{\theta}_5-\widetilde{\theta}^\star_5),\\ D_{\widetilde{\theta}_i}\frac{\mathcal{L}}{d}\Big|_{\widetilde{\theta}_5} &= D_{\widetilde{\theta}_i}\frac{\min_{\widetilde{\theta}_5}\mathcal{L}}{d} - (D_{\widetilde{\theta}_i}\widetilde{\theta}^\star_5)^{\mathsf{T}}D_{\widetilde{\theta}_5}\frac{\mathcal{L}}{d}, \qquad i=1,2.\end{aligned} \tag{69}$$

For a sufficiently large fixed C, put

$$\Lambda_T = \exp\{CT^2\log(e+T)\}(n/d)^{CT^2} = d^{o(1)}. \tag{70}$$

The equality follows from (17). Lemma G.5, (160), and the normal equations give

$$\|\mathbf{H}^{-1}\|_2 + \sum_{i=1}^{2}\|D_{\widetilde{\theta}_i}\widetilde{\theta}^\star_5\|^2 \le \exp\{CT\log(e+T)\}(n/d)^{CT} \le \Lambda_T.$$

It follows that

$$\sum_{i=1}^{2}\left|D_{\widetilde{\theta}_i}\frac{\min_{\widetilde{\theta}_5}\mathcal{L}}{d}\right|^2 \le 2\sum_{i=1}^{2}\left|D_{\widetilde{\theta}_i}\frac{\mathcal{L}}{d}\Big|_{\widetilde{\theta}_5}\right|^2 + 2\Lambda_T\left\|D_{\widetilde{\theta}_5}\frac{\mathcal{L}}{d}\right\|^2,$$
$$\left\|D_{\widetilde{\theta}_5}\frac{\mathcal{L}}{d}\right\|^2 \ge \frac{2}{T\Lambda_T}\frac{\mathcal{L}-\min_{\widetilde{\theta}_5}\mathcal{L}}{d}.$$

Combining these bounds with (67) yields

$$\sum_{i=1}^{2}\left|D_{\widetilde{\theta}_i}\frac{\mathcal{L}}{d}\Big|_{\widetilde{\theta}_5}\right|^2 + 4\Lambda_T\left\|D_{\widetilde{\theta}_5}\frac{\mathcal{L}}{d}\right\|^2 \ge c\frac{d/n}{T}\frac{\mathcal{L}-\mathcal{L}^\star}{d}. \tag{71}$$

(iii) Let $\mathbf{M}_t$ be the monomial coefficient matrix of $p_1, p_2, \ldots, p_t$, and let $\mathbf{D}_t$ be its diagonal. Then

$$\mathbf{D}_t = \operatorname{diag}(\widetilde{\theta}_1, \widetilde{\theta}_1\widetilde{\theta}_2, \ldots, \widetilde{\theta}_1\widetilde{\theta}_2^{t-1}), \qquad \mathbf{N}_t = \mathbf{D}_t^{-1}\mathbf{M}_t - \mathbf{I}, \qquad \mathbf{N}_t^t = 0,$$
$$\|\mathbf{M}_t\|_2 + \|\mathbf{D}_t^{-1}\|_2 \le \exp\{CT\log(e+T)\}(n/d)^{CT}.$$

The coefficient bound follows from $\mathcal{L}/d \le C/T$ and (160). Hence

$$\mathbf{M}_t^{-1} = \sum_{j=0}^{t-1} (-\mathbf{N}_t)^j \mathbf{D}_t^{-1}, \qquad \|\mathbf{M}_t^{-1}\|_2 \le \exp\{CT^2 \log(e+T)\}(n/d)^{CT^2}. \tag{72}$$

Applying this inverse to $p_{t+1} - (1+\widetilde{\theta}_2\lambda)p_t$ bounds $\|\widetilde{\mathbf{\Theta}}_4\|_F^2$ by Λ_T, after increasing its fixed constant. For arbitrary $\delta\widetilde{\theta}_1, \delta\widetilde{\theta}_2, \delta\widetilde{\theta}_5$, prescribe

$$\begin{aligned} \delta p_1 &= \lambda\, \delta\widetilde{\theta}_1, \\ \delta p_{t+1} &= \delta(\widetilde{\theta}_1 \widetilde{\theta}_2^t)\lambda^{t+1} + \sum_{j=1}^{t} (\delta\widetilde{\theta}_5)_{jt} \lambda^j. \end{aligned}$$

Differentiating (47) gives the full inverse

$$\begin{aligned} (\delta\widetilde{\mathbf{\Theta}}_4)_{1:t,t} = \mathbf{M}_t^{-1} \operatorname{coeff}_{1:t} \Big[& \delta p_{t+1} - (1+\widetilde{\theta}_2\lambda)\delta p_t - \delta\widetilde{\theta}_2\, \lambda p_t \\ & - \sum_{\ell=1}^{t} \delta p_\ell\, (\widetilde{\mathbf{\Theta}}_4)_{\ell t} \Big]. \end{aligned} \tag{73}$$

The degree-$(t+1)$ terms cancel. Since all δp_ℓ are prescribed, the preceding bounds give

$$\|\delta\widetilde{\mathbf{\Theta}}_4\|_F^2 \le \Lambda_T \left(|\delta\widetilde{\theta}_1|^2 + |\delta\widetilde{\theta}_2|^2 + \|\delta\widetilde{\theta}_5\|^2 \right),$$

after one fixed enlargement of C in (70). Taking the dual norms gives

$$\begin{aligned} \left\| D_{\widetilde{\theta}_5} \frac{\mathcal{L}}{d} \right\|^2 &\le \Lambda_T \left\| D_{\widetilde{\mathbf{\Theta}}_4} \frac{\mathcal{L}}{d} \right\|_F^2, \\ \sum_{i=1}^{2} \left| D_{\widetilde{\theta}_i} \frac{\mathcal{L}}{d} \Big|_{\widetilde{\theta}_5} \right|^2 &\le 2 \sum_{i=1}^{2} \left| D_{\widetilde{\theta}_i} \frac{\mathcal{L}}{d} \Big|_{\widetilde{\mathbf{\Theta}}_4} \right|^2 + 2\Lambda_T \left\| D_{\widetilde{\mathbf{\Theta}}_4} \frac{\mathcal{L}}{d} \right\|_F^2. \end{aligned}$$

Under (17),

$$(2 + 4\Lambda_T)\Lambda_T = d^{o(1)} \le 2\sqrt{d}$$

for all sufficiently large d. Substitution in (71) and multiplication by d therefore give

$$\frac{1}{d} \sum_{i=1}^{2} \left| D_{\widetilde{\theta}_i} \mathcal{L} \Big|_{\widetilde{\mathbf{\Theta}}_4} \right|^2 + \frac{1}{\sqrt{d}} \|D_{\widetilde{\mathbf{\Theta}}_4} \mathcal{L}\|_F^2 \ge c \frac{d/n}{T} (\mathcal{L} - \mathcal{L}^\star). \tag{74}$$

□

D.3 Proof of Lemma 4.8

Proof. By (112), every variation of the active auxiliary parameters has a raw lift with

$$\begin{aligned} \mathbf{J}_2\, \delta\bar{\boldsymbol{\theta}} &= (\delta\widetilde{\theta}_1, \delta\widetilde{\theta}_2, \big((\delta\widetilde{\mathbf{\Theta}}_4)_{\ell t}\big)_{1 \le \ell \le t < T}^{\mathsf{T}})^{\mathsf{T}}, \\ \|\delta\bar{\boldsymbol{\theta}}\|^2 &\le C\big(1 + (d/n)^{-1/2} + v_{31}^2 + v_{33}^2\big) \left\{ d|\delta\widetilde{\theta}_1|^2 + d|\delta\widetilde{\theta}_2|^2 + \|\delta\widetilde{\mathbf{\Theta}}_4\|_F^2 \right\}. \end{aligned}$$

The factors d are from $\|\delta v_{31}\mathbf{I}_d\|_F^2 = d|\delta v_{31}|^2$ and $\|\delta w_{13}\mathbf{I}_d\|_F^2 = d|\delta w_{13}|^2$. The minimum norm of a raw lift gives

$$\begin{aligned} (\mathbf{J}_2\mathbf{J}_2^{\mathsf{T}})^{-1} &\preceq C\big(1 + (d/n)^{-1/2} + v_{31}^2 + v_{33}^2\big)\, \operatorname{diag}(d, d, \mathbf{I}), \\ \mathbf{J}_2\mathbf{J}_2^{\mathsf{T}} &\succeq \frac{c}{1 + (d/n)^{-1/2} + v_{31}^2 + v_{33}^2} \operatorname{diag}(d^{-1}, d^{-1}, \mathbf{I}) \\ &\succeq \frac{c}{1 + (d/n)^{-1/2} + v_{31}^2 + v_{33}^2} \operatorname{diag}(d^{-1}, d^{-1}, d^{-1/2}\mathbf{I}). \end{aligned}$$

Here the matrices are restricted to the active index ranges, as in Appendix B. With Stage-2 time zero at the handoff, the energy identity gives

$$\int_0^\tau (\dot{v}_{31}^2 + \dot{v}_{33}^2)\,\mathrm{d}s \le \frac{\mathcal{L}(0) - \mathcal{L}(\tau)}{d} \le \frac{\mathcal{L}(0)}{d} \le C\frac{d/n}{T}.$$

The last bound follows from Lemma E.2 and $n = \Theta(d)$. Lemma G.15 and Cauchy–Schwarz yield

$$\begin{aligned} 1 + (d/n)^{-1/2} + v_{31}(\tau)^2 + v_{33}(\tau)^2 &\le 1 + (d/n)^{-1/2} + 2v_{31}(0)^2 + 2v_{33}(0)^2 + 2\tau\,\frac{\mathcal{L}(0)}{d} \\ &\le d^{o(1)} + C\frac{d/n}{T}\tau. \end{aligned}$$

Consequently,

$$\mathbf{J}_2\mathbf{J}_2^\top \succeq \frac{c}{d^{o(1)} + \frac{d/n}{T}\tau}\,\mathrm{diag}(d^{-1}, d^{-1}, d^{-1/2}\mathbf{I}).$$

In the time convention of Lemma 4.8, replace the elapsed time by $\tau - \tau_1 \le \tau$. □

E Evolution of $\mathbf{w}_{t+1}(\tau)$

Lemma E.1. *Uniformly for* $1/3 \le \eta_0 \le 9/10$ *and* $0 < d/n \le 1/10$,

$$\sum_{t\ge 1} \mathbb{E}_{\lambda\sim\mu_{d/n}}(1 - \eta_0\lambda)^{2t} < \frac{49}{50}. \tag{75}$$

Uniformly for $1/3 \le (-\widetilde{\theta}_2) \le 9/10$, *the MP conditional function satisfies*

$$2T \min_{\widetilde{\theta}_1} \frac{1}{2T}\sum_{t=1}^{T}\left[\begin{array}{c}\frac{\left(\frac{d}{n}\right)^{t-1}}{\sum_{r=0}^{t-1}\left(\frac{d}{n}\right)^r}\left(1 - \widetilde{\theta}_1(-\widetilde{\theta}_2)^{t-1}\right)^2 \\ + \left(\frac{d}{n}\right)^t \widetilde{\theta}_1^2(-\widetilde{\theta}_2)^{2t-2}\end{array}\right] < \frac{149}{100}\frac{d}{n}. \tag{76}$$

The same bound holds for the finite function up to $d^{-1+o(1)}$.

Proof. Geometric summation gives

$$\sum_{t\ge 1}\mathbb{E}_{\mu_{d/n}}(1 - \eta_0\lambda)^{2t} = -1 + \frac{1}{2\eta_0(1 - d/n)} + \frac{1}{2}\mathbb{E}_{\mu_{d/n}}\frac{1}{2 - \eta_0\lambda}.$$

This is convex in η_0: each summand before summation is convex. For fixed η_0, it is increasing in d/n; expand the last resolvent in powers of $\eta_0\lambda/2$ and use the nonnegative coefficients of the MP moment polynomials. The expansion converges uniformly by (187). Thus it suffices to check the two step endpoints at $d/n = 1/10$. Direct evaluation gives, respectively,

$$\frac{169}{24} - \frac{\sqrt{2361}}{8} < \frac{49}{50}, \qquad \frac{823}{648} - \frac{\sqrt{6961}}{72} < \frac{49}{50}.$$

For the conditional bound, Lemma D.5 makes the minimized function decreasing throughout $[1/3, 9/10]$, since its minimizing $(-\widetilde{\theta}_2)$ is greater than $(1 + d/n)/(1 + d/n + (d/n)^2) > 9/10$. At $(-\widetilde{\theta}_2) = 1/3$, take $\widetilde{\theta}_1 = (1 + d/n)^{-1}$ in (61). Keep its first two terms and bound the remaining terms by $(1 + d/n)\sum_{k\ge 2}(d/n)^k$. This gives

$$\begin{aligned} &2T \min_{\widetilde{\theta}_1} \frac{1}{2T}\sum_{t=1}^{T}\left[\begin{array}{c}\frac{\left(\frac{d}{n}\right)^{t-1}}{\sum_{r=0}^{t-1}\left(\frac{d}{n}\right)^r}\left(1 - \widetilde{\theta}_1(-\widetilde{\theta}_2)^{t-1}\right)^2 \\ + \left(\frac{d}{n}\right)^t \widetilde{\theta}_1^2(-\widetilde{\theta}_2)^{2t-2}\end{array}\right]\Bigg|_{(-\widetilde{\theta}_2)=1/3} \\ &\le \frac{d/n}{1 + d/n} + \frac{d/n}{1 + d/n}\left(1 - \frac{1}{3(1 + d/n)}\right)^2 + \frac{(d/n)^2}{9(1 + d/n)^2} + \frac{(1 + d/n)(d/n)^2}{1 - d/n} < \frac{149}{100}\frac{d}{n}. \end{aligned}$$

After clearing the positive denominator, the last inequality is

$$41 - 18\frac{d}{n} - 2400\left(\frac{d}{n}\right)^2 - 6182\left(\frac{d}{n}\right)^3 - 4941\left(\frac{d}{n}\right)^4 - 900\left(\frac{d}{n}\right)^5 > 0.$$

Its left side is decreasing and is $85149/10000 > 0$ at $d/n = 1/10$. The finite-function coefficient transfer in Lemma G.8 gives the stated perturbation. □

Lemma E.2. *From the unchanged endpoint of Lemma F.3, the Stage-2 raw flow reaches*

$$\frac{2T\mathcal{L}}{d} \le \frac{3}{2}\frac{d}{n} \tag{77}$$

within time $d^{-1+o(1)}$. Throughout this initial interval,

$$\frac{1}{100} < \widetilde{\theta}_1 < 2, \qquad -\widetilde{\theta}_2 = \eta_0 + o(1), \qquad c \le |v_{31}|, w_{13}, v_{33} \le C,$$

and every ordinary raw block has norm at most $d^{o(1)}$. No parameter is frozen or reset during this interval.

Proof. The energy identity implies global existence by the finite-interval path-length argument used in Stage 1. Lemma E.1 and (101) give

$$2T\mathcal{L}(0)/d < 49/50.$$

On the interval $0 \le \tau \le d^{-1/2}$, the scalar energy bound and (207) therefore give

$$\begin{aligned} &\int_0^{d^{-1/2}} (|\dot{v}_{31}| + |\dot{w}_{13}| + |\dot{v}_{33}|)\,\mathrm{d}\tau \le Cd^{-1/4}/\sqrt{T}, \\ &-\widetilde{\theta}_2 = \eta_0 + O(d^{-1/4}), \qquad c \le |v_{31}|, w_{13}, v_{33} \le C. \end{aligned} \tag{78}$$

The first residual has squared norm $(1 + d/n + 1/n)\widetilde{\theta}_1^2 - 2\widetilde{\theta}_1 + 1 \ge (\widetilde{\theta}_1 - 1)^2$. Loss monotonicity thus gives $1/100 < \widetilde{\theta}_1 < 2$. In particular $8/25 < (-\widetilde{\theta}_2) < 23/25$ for sufficiently large d. The finite Gram bound controls all rollout coefficients by $d^{o(1)}$. Their leading coefficients are $\widetilde{\theta}_1\widetilde{\theta}_2^{t-1}$, whose inverses are $d^{o(1)}$. Triangular inversion in Lemma D.1 consequently bounds the auxiliary recurrence coefficients and their inverse Jacobian by

$$\exp\{CT^2\log(e+T)\}(n/d)^{CT^2} = d^{o(1)}.$$

This argument uses only the preceding bounds, not the later compactness lemma.

(i) For $1 \le s < T$, differentiating (47) gives

$$\begin{aligned} \delta p_1 &= \lambda\,\delta\widetilde{\theta}_1, \\ \delta p_{s+1} &= (1 + \widetilde{\theta}_2\lambda)\delta p_s + \lambda p_s\,\delta\widetilde{\theta}_2 \\ &\quad + \delta\Big[v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_s) + \widetilde{\theta}_1(\widetilde{\mathbf{W}}_{55})_{1s}\Big]\lambda \\ &\quad + \sum_{\ell=2}^{s}\Big[(\widetilde{\mathbf{W}}_{55})_{\ell s}\delta p_\ell + \delta(\widetilde{\mathbf{W}}_{55})_{\ell s}p_\ell\Big]. \end{aligned} \tag{79}$$

Consequently,

$$\begin{aligned} D\left(\frac{\mathcal{L}}{d}\right)[\delta\widetilde{\theta}_1, \delta\widetilde{\theta}_2, \delta\widetilde{\mathbf{\Theta}}_4] &= D_{\widetilde{\theta}_1}\frac{\mathcal{L}}{d}\,\delta\widetilde{\theta}_1 + D_{\widetilde{\theta}_2}\frac{\mathcal{L}}{d}\,\delta\widetilde{\theta}_2 \\ &\quad + \sum_{t=1}^{T-1} D_{(\widetilde{\mathbf{\Theta}}_4)_{1t}}\frac{\mathcal{L}}{d}\,\delta\left[\frac{v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t)}{\widetilde{\theta}_1} + (\widetilde{\mathbf{W}}_{55})_{1t}\right] \\ &\quad + \sum_{t=1}^{T-1}\sum_{\ell=2}^{t} D_{(\widetilde{\mathbf{W}}_{55})_{\ell t}}\frac{\mathcal{L}}{d}\,\delta(\widetilde{\mathbf{W}}_{55})_{\ell t}, \end{aligned} \tag{80}$$

where (79) contains every later-time autoregressive contribution. The raw chain rule is

$$
\begin{aligned}
D_{w_{24}}\frac{\mathcal{L}}{d} &= v_{31}D_{\widetilde{\theta}_1}\frac{\mathcal{L}}{d} + \frac{v_{31}}{\widetilde{\theta}_1}\sum_{t=1}^{T-1}\left(1+(\widetilde{\mathbf{W}}_{55})_{1t}-(\widetilde{\mathbf{\Theta}}_4)_{1t}\right)D_{(\widetilde{\mathbf{\Theta}}_4)_{1t}}\frac{\mathcal{L}}{d},\\
D_{(\mathbf{w}_{25})_0}\frac{\mathcal{L}}{d} &= v_{31}D_{\widetilde{\theta}_1}\frac{\mathcal{L}}{d} + \frac{v_{31}}{\widetilde{\theta}_1}\sum_{t=1}^{T-1}\left((\widetilde{\mathbf{W}}_{55})_{1t}-(\widetilde{\mathbf{\Theta}}_4)_{1t}\right)D_{(\widetilde{\mathbf{\Theta}}_4)_{1t}}\frac{\mathcal{L}}{d},\\
D_{(\mathbf{w}_{25})_t}\frac{\mathcal{L}}{d} &= \frac{v_{31}}{\widetilde{\theta}_1}D_{(\widetilde{\mathbf{\Theta}}_4)_{1t}}\frac{\mathcal{L}}{d}, \qquad 1\le t<T,\\
\nabla_{w_{44}}\frac{\mathcal{L}}{d} &= v_{33}\left[\sum_{t=1}^{T-1}D_{(\widetilde{\mathbf{\Theta}}_4)_{1t}}\frac{\mathcal{L}}{d}+\sum_{t=1}^{T-1}\sum_{\ell=2}^{t}D_{(\widetilde{\mathbf{W}}_{55})_{\ell t}}\frac{\mathcal{L}}{d}\right],\\
\nabla_{\mathbf{w}_{45}}\frac{\mathcal{L}}{d} &= v_{33}\left[\sum_{t=1}^{T-1}D_{(\widetilde{\mathbf{\Theta}}_4)_{1t}}\frac{\mathcal{L}}{d}\,\mathbf{e}_t^{\mathsf{T}}+\sum_{t=1}^{T-1}\sum_{\ell=2}^{t}D_{(\widetilde{\mathbf{W}}_{55})_{\ell t}}\frac{\mathcal{L}}{d}\,\mathbf{e}_t^{\mathsf{T}}\right],\\
\nabla_{\mathbf{w}_{54}}\frac{\mathcal{L}}{d} &= v_{33}\left[\sum_{t=1}^{T-1}D_{(\widetilde{\mathbf{\Theta}}_4)_{1t}}\frac{\mathcal{L}}{d}\,\mathbf{e}_1+\sum_{t=1}^{T-1}\sum_{\ell=2}^{t}D_{(\widetilde{\mathbf{W}}_{55})_{\ell t}}\frac{\mathcal{L}}{d}\,\mathbf{e}_\ell\right],\\
\nabla_{\mathbf{W}_{55}}\frac{\mathcal{L}}{d} &= v_{33}\left[\sum_{t=1}^{T-1}D_{(\widetilde{\mathbf{\Theta}}_4)_{1t}}\frac{\mathcal{L}}{d}\,\mathbf{e}_1\mathbf{e}_t^{\mathsf{T}}+\sum_{t=1}^{T-1}\sum_{\ell=2}^{t}D_{(\widetilde{\mathbf{W}}_{55})_{\ell t}}\frac{\mathcal{L}}{d}\,\mathbf{e}_\ell\mathbf{e}_t^{\mathsf{T}}\right].
\end{aligned}
\tag{81}
$$

With $\mathbf{K}$ from (201), these identities give

$$
\left[\mathbf{I}-\frac{1}{2}\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}(1\ \ \mathbf{e}_0^{\mathsf{T}})\right]\begin{pmatrix}\dot{w}_{24}\\ \dot{\mathbf{w}}_{25}^{\mathsf{T}}\end{pmatrix}=\frac{v_{31}}{\widetilde{\theta}_1 v_{33}}\mathbf{K}\begin{pmatrix}\dot{w}_{44}\\ \dot{\mathbf{w}}_{45}^{\mathsf{T}}\\ \dot{\mathbf{w}}_{54}\\ \mathrm{vec}(\dot{\mathbf{W}}_{55})\end{pmatrix}. \tag{82}
$$

Also,

$$
\begin{pmatrix}w_{24}\\ \mathbf{w}_{25}^{\mathsf{T}}\end{pmatrix}=\left[\mathbf{I}-\frac{1}{2}\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}(1\ \ \mathbf{e}_0^{\mathsf{T}})\right]\begin{pmatrix}w_{24}\\ \mathbf{w}_{25}^{\mathsf{T}}\end{pmatrix}+\frac{\widetilde{\theta}_1}{2v_{31}}\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}, \tag{83}
$$

and (46) implies

$$
\begin{aligned}
&\mathbf{K}\begin{pmatrix}w_{44}\\ \mathbf{w}_{45}^{\mathsf{T}}\\ \mathbf{w}_{54}\\ \mathrm{vec}(\mathbf{W}_{55})\end{pmatrix}+\frac{v_{31}}{\widetilde{\theta}_1 v_{33}}\mathbf{K}\mathbf{K}^{\mathsf{T}}\left[\mathbf{I}-\frac{1}{2}\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}(1\ \ \mathbf{e}_0^{\mathsf{T}})\right]\begin{pmatrix}w_{24}\\ \mathbf{w}_{25}^{\mathsf{T}}\end{pmatrix}\\
&\quad=\mathbf{K}\begin{pmatrix}0\\ \mathbf{0}_T\\ \mathbf{0}_T\\ \mathrm{vec}\left(\left[\dfrac{(\widetilde{\mathbf{\Theta}}_4)_{\ell t}-\frac{1}{2}\mathbb{1}_{\{\ell=1\}}}{v_{33}}\right]_{\ell,t}\right)\end{pmatrix}.
\end{aligned}
\tag{84}
$$

Here and below, $\widetilde{\mathbf{\Theta}}_4$ is expressed in $(\widetilde{\theta}_1,\widetilde{\theta}_2,\widetilde{\theta}_5)$ using (73). The coefficient bounds give

$$
\sum_{t=1}^{T-1}\left|\frac{(\widetilde{\mathbf{\Theta}}_4)_{1t}}{v_{33}}-\frac{1}{2v_{33}}\right|^2+\sum_{t=1}^{T-1}\sum_{\ell=2}^{t}\left|\frac{(\widetilde{\mathbf{\Theta}}_4)_{\ell t}}{v_{33}}\right|^2\le d^{o(1)}. \tag{85}
$$

For $R_{t,\mathrm{ls}} = R_t|_{\widetilde{\theta}_5=\widetilde{\theta}_5^\star}$, (69) gives

$$\sum_{t=1}^{T} \|R_t - R_{t,\mathrm{ls}}\|_{d,n}^2 = T^2 \left(D_{\widetilde{\theta}_5} \frac{\mathcal{L}}{d}\right)^{\mathsf{T}} \mathbf{H}^{-1} D_{\widetilde{\theta}_5} \frac{\mathcal{L}}{d}$$
$$\leq d^{o(1)} \left\| D_{\widetilde{\mathbf{\Theta}}_4} \frac{\mathcal{L}}{d} \right\|_F^2,$$
$$\left\| D_{\widetilde{\mathbf{\Theta}}_4} \frac{\mathcal{L}}{d} \right\|_F^2 = \frac{\widetilde{\theta}_1^2}{v_{31}^2} \sum_{t=1}^{T-1} \left| D_{(\mathbf{w}_{25})_t} \frac{\mathcal{L}}{d} \right|^2 + \frac{1}{v_{33}^2} \sum_{t=1}^{T-1} \sum_{\ell=2}^{t} \left| D_{(\mathbf{W}_{55})_{\ell t}} \frac{\mathcal{L}}{d} \right|^2.$$

In the first inequality we used

$$D_{\widetilde{\theta}_5} \frac{\mathcal{L}}{d} = (D_{\widetilde{\theta}_5} \operatorname{vec} \widetilde{\mathbf{\Theta}}_4)^{\mathsf{T}} \operatorname{vec} \left(D_{\widetilde{\mathbf{\Theta}}_4} \frac{\mathcal{L}}{d} \right), \qquad \|\mathbf{H}^{-1}\|_2 + \|D_{\widetilde{\theta}_5} \operatorname{vec} \widetilde{\mathbf{\Theta}}_4\|_2^2 \leq d^{o(1)}.$$

Consequently,

$$\begin{aligned} &\left(\sum_{t=1}^{T} \|R_t - R_{t,\mathrm{ls}}\|_{d,n}^2 \right)^{1/2} \\ &\leq d^{o(1)} \left(\sum_{t=1}^{T-1} \left| D_{(\mathbf{w}_{25})_t} \frac{\mathcal{L}}{d} \right|^2 + \sum_{t=1}^{T-1} \sum_{\ell=2}^{t} \left| D_{(\mathbf{W}_{55})_{\ell t}} \frac{\mathcal{L}}{d} \right|^2 \right)^{1/2} \\ &\leq d^{-1/2+o(1)} \sqrt{-\frac{\mathrm{d}}{\mathrm{d}\tau} \frac{\mathcal{L}}{d}}. \end{aligned} \tag{86}$$

For $0 \leq s \leq 1$, (60) gives

$$\mathcal{L}|_{\widetilde{\theta}_5=\widetilde{\theta}_5^\star + s(\widetilde{\theta}_5 - \widetilde{\theta}_5^\star)} \leq \mathcal{L}.$$

The same inverse-Jacobian bound along this segment therefore yields

$$\begin{aligned} &\left\| \{\mathrm{RHS}\,(84)\} - \{\mathrm{RHS}\,(84)\}|_{\widetilde{\theta}_5=\widetilde{\theta}_5^\star} \right\|_2 \\ &\leq \|\widetilde{\theta}_5 - \widetilde{\theta}_5^\star\|_2 \int_0^1 \left\| D_{\widetilde{\theta}_5} \{\mathrm{RHS}\,(84)\}|_{\widetilde{\theta}_5=\widetilde{\theta}_5^\star + s(\widetilde{\theta}_5 - \widetilde{\theta}_5^\star)} \right\|_2 \mathrm{d}s \\ &\leq d^{o(1)} \left(\sum_{t=1}^{T} \|R_t - R_{t,\mathrm{ls}}\|_{d,n}^2 \right)^{1/2}. \end{aligned} \tag{87}$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \frac{v_{31}}{\widetilde{\theta}_1 v_{33}} = -\frac{\dot{w}_{24} + \dot{\mathbf{w}}_{25}\mathbf{e}_0}{(w_{24} + \mathbf{w}_{25}\mathbf{e}_0)^2 v_{33}} - \frac{\dot{v}_{33}}{(w_{24} + \mathbf{w}_{25}\mathbf{e}_0) v_{33}^2},$$
$$\left| \frac{\mathrm{d}}{\mathrm{d}\tau} \frac{v_{31}}{\widetilde{\theta}_1 v_{33}} \right| \leq C \sqrt{-d \frac{\mathrm{d}}{\mathrm{d}\tau} \frac{\mathcal{L}}{d}},$$

we have

$$\begin{aligned} &\int_0^{d^{-1/2}} \left| \frac{\mathrm{d}}{\mathrm{d}\tau} \frac{v_{31}}{\widetilde{\theta}_1 v_{33}} \right| \left(\sum_{t=1}^{T} \|R_t - R_{t,\mathrm{ls}}\|_{d,n}^2 \right)^{1/2} \mathrm{d}\tau \\ &\leq d^{o(1)} \int_0^{d^{-1/2}} \left(-\frac{\mathrm{d}}{\mathrm{d}\tau} \frac{\mathcal{L}}{d} \right) \mathrm{d}\tau \leq d^{o(1)}. \end{aligned} \tag{88}$$

For the conditional minimizer, the differentiated normal equations in (68) and (73) give

$$\sup_{1/100 \leq a \leq 2} \sum_{\xi \in \{v_{31}, w_{13}, v_{33}\}} \sum_{j=0}^{1} \left\| D_\xi^j \left[\{\mathrm{RHS}\,(84)\}|_{\substack{\widetilde{\theta}_1 = a,\ \widetilde{\theta}_2 = v_{31} w_{13} \\ \widetilde{\theta}_5 = \widetilde{\theta}_5^\star}} \right] \right\|_2 \leq d^{o(1)}. \tag{89}$$

Moreover, for $1/100 \le a \le 2$,

$$\left\| \left(\mathbf{I} + \frac{v_{31}^2}{a^2 v_{33}^2} \mathbf{K}\mathbf{K}^\mathsf{T} \right)^{-3/2} \right\|_2 \le 1, \qquad \sum_{\xi \in \{v_{31}, v_{33}\}} \left\| D_\xi \left(\mathbf{I} + \frac{v_{31}^2}{a^2 v_{33}^2} \mathbf{K}\mathbf{K}^\mathsf{T} \right)^{-3/2} \right\|_2 \le CT^4.$$

Using

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \frac{v_{31}}{\widetilde{\theta}_1 v_{33}} = -\frac{v_{31} \dot{\widetilde{\theta}}_1}{\widetilde{\theta}_1^2 v_{33}} + \frac{\dot{v}_{31}}{\widetilde{\theta}_1 v_{33}} - \frac{v_{31} \dot{v}_{33}}{\widetilde{\theta}_1 v_{33}^2},$$

the chain rule gives

$$\begin{aligned} &\left\| \int_0^\tau \left(\frac{\mathrm{d}}{\mathrm{d}s} \frac{v_{31}}{\widetilde{\theta}_1 v_{33}} \right) \left(\mathbf{I} + \frac{v_{31}^2}{\widetilde{\theta}_1^2 v_{33}^2} \mathbf{K}\mathbf{K}^\mathsf{T} \right)^{-3/2} \{\mathrm{RHS}\,(84)\}|_{\widetilde{\theta}_5 = \widetilde{\theta}_5^\star} \, \mathrm{d}s \right. \\ &\quad \left. - \left[\int_1^{\widetilde{\theta}_1} \frac{-v_{31}}{a^2 v_{33}} \left(\mathbf{I} + \frac{v_{31}^2}{a^2 v_{33}^2} \mathbf{K}\mathbf{K}^\mathsf{T} \right)^{-3/2} \{\mathrm{RHS}\,(84)\}|_{\substack{\widetilde{\theta}_1 = a,\ \widetilde{\theta}_2 = v_{31} w_{13} \\ \widetilde{\theta}_5 = \widetilde{\theta}_5^\star}} \, \mathrm{d}a \right]_0^\tau \right\|_2 \\ &\quad \le d^{o(1)} \int_0^\tau (|\dot{v}_{31}| + |\dot{w}_{13}| + |\dot{v}_{33}|)\, \mathrm{d}s \le d^{o(1)}. \end{aligned} \tag{90}$$

Both endpoint integrals are bounded by $d^{o(1)}$ by (89). Substituting $\eta_0 \mapsto \widetilde{\theta}_1$ and RHS (204) $\mapsto$ RHS (84) in (205), and applying (87)–(90), gives

$$\sup_{0 \le \tau \le d^{-1/2}} \left\| \left[\mathbf{I} - \frac{1}{2} \begin{pmatrix} 1 \\ \mathbf{e}_0 \end{pmatrix} (1 \ \ \mathbf{e}_0^\mathsf{T}) \right] \begin{pmatrix} w_{24}(\tau) \\ \mathbf{w}_{25}(\tau)^\mathsf{T} \end{pmatrix} \right\|_2 \le d^{o(1)}.$$

The last four lines of (81) preserve (202). Thus (83) and (85) yield

$$\sup_{0 \le \tau \le d^{-1/2}} \left\{ |w_{24}| + \|\mathbf{w}_{25}\|_2 + \sum_{B \in \{44,45,54,55\}} \|\mathbf{W}_B\|_F \right\} \le d^{o(1)}.$$

(ii) Hold $(-\widetilde{\theta}_2)$ fixed. In the coordinates consisting of $\widetilde{\theta}_1$ and all lower polynomial coefficients, $\mathcal{L}/d$ is a strictly convex quadratic: every leading coefficient is linear in $\widetilde{\theta}_1$. The first residual controls $\widetilde{\theta}_1$, and the finite Gram bound therefore gives a PL constant $d^{-o(1)}$. The inverse recurrence Jacobian has norm $d^{o(1)}$. Its auxiliary variations have the following raw lift with all three raw scalars fixed:

$$\begin{aligned} \delta(\mathbf{w}_{25})_0 &= \frac{\delta \widetilde{\theta}_1}{v_{31}}, \\ \delta(\mathbf{w}_{25})_t &= \frac{\widetilde{\theta}_1 (\delta \widetilde{\mathbf{\Theta}}_4)_{1t} + \big((\widetilde{\mathbf{\Theta}}_4)_{1t} - (\widetilde{\mathbf{W}}_{55})_{1t} \big) \delta \widetilde{\theta}_1}{v_{31}}, \\ \delta(\mathbf{W}_{55})_{\ell t} &= \frac{\delta (\widetilde{\mathbf{W}}_{55})_{\ell t}}{v_{33}} \quad (\ell \ge 2). \end{aligned}$$

Set every other raw variation to zero. Part (i) bounds its squared ordinary Euclidean norm by $d^{o(1)}$ times the prescribed squared coordinate norm. Duality and the raw mobility d consequently give

$$-\frac{\mathrm{d}}{\mathrm{d}\tau} \frac{\mathcal{L}}{d} \ge d^{1-o(1)} \left\{ \frac{\mathcal{L}}{d} - \frac{\min_{\widetilde{\theta}_1, \widetilde{\mathbf{\Theta}}_4} \mathcal{L}}{d} \right\}. \tag{91}$$

Lemma E.1, the finite coefficient transfer, the bounded first derivative of the function, and (78) imply, uniformly on the interval,

$$\frac{2T}{d} \min_{\widetilde{\theta}_1, \widetilde{\mathbf{\Theta}}_4} \mathcal{L} \le \frac{299}{200} \frac{d}{n}.$$

While $2T\mathcal{L}/d > 3d/(2n)$, (91) gives $-\frac{\mathrm{d}}{\mathrm{d}\tau}(\mathcal{L}/d) \ge d^{1-o(1)}(d/n)/(400T)$. Since $\mathcal{L}(0)/d \le 49/(100T)$ and d/n is bounded below by a fixed positive constant, this level must be reached in time $d^{-1+o(1)} < d^{-1/2}$. This proves entry into (77). □

F HANDOFF AND ESTIMATES

F.1 HANDOFF ESTIMATES

Residual formulation. All norms in this section are the appendix spectral norms. The finite spectral law is $\nu_{d,n}$, and the comparison law is $\mu_{d/n}$. The recurrence in Appendix D.2 fixes

$$R_t(0) = -1, \qquad [\lambda^t]R_t = \widetilde{\theta}_1\widetilde{\theta}_2^{t-1}. \tag{92}$$

After $\widetilde{\theta}_1, (-\widetilde{\theta}_2)$ are fixed, changing the t-th trainable coefficient column changes R_{t+1} by an arbitrary polynomial with zero constant term and degree at most t.

Lemma F.1. *Assume that the moment Gram matrices through degree T of $\nu_{d,n}$ are positive definite and $\widetilde{\theta}_1\widetilde{\theta}_2 \neq 0$. A point is stationary in every trainable coefficient column if and only if*

$$\langle R_t, q\rangle_{d,n} = 0 \quad (q(0) = 0,\ \deg q \le t-1), \qquad 1 \le t \le T. \tag{93}$$

Consequently, for fixed $\widetilde{\theta}_1, (-\widetilde{\theta}_2)$, each residual is the unique minimizer of

$$\min_{\substack{R(0)=-1,\ \deg R\le t\\ [\lambda^t]R=\widetilde{\theta}_1\widetilde{\theta}_2^{t-1}}} \|R\|_{d,n}^2. \tag{94}$$

Moreover, differentiating the minimized value in $\widetilde{\theta}_1, (-\widetilde{\theta}_2)$ gives

$$\nabla_{\widetilde{\theta}_1,(-\widetilde{\theta}_2)}\left(\frac{\mathcal{L}}{d}\right) = \nabla_{\widetilde{\theta}_1,(-\widetilde{\theta}_2)}\left[\frac{1}{2T}\sum_{t=1}^{T}\min_{\substack{R(0)=-1,\ \deg R\le t\\ [\lambda^t]R=(-1)^{t-1}\widetilde{\theta}_1(-\widetilde{\theta}_2)^{t-1}}}\|R\|_{d,n}^2\right]. \tag{95}$$

Proof. **(i)** Fix $2 \le m \le T$ and $q(0) = 0$, $\deg q \le m-1$. Set

$$\delta\widetilde{\theta}_1 = \delta\widetilde{\theta}_2 = 0, \qquad \delta R_1 = 0.$$

By (48), the matrices $\mathbf{M}_s$ in (73) satisfy

$$\det \mathbf{M}_s = \prod_{\ell=1}^{s}\widetilde{\theta}_1\widetilde{\theta}_2^{\ell-1} \neq 0,$$
$$\operatorname{span}\{p_1,\dots,p_s\} = \operatorname{span}\{\lambda,\dots,\lambda^s\}, \qquad 1 \le s < T.$$

(ii) Since $p_1 = \widetilde{\theta}_1\lambda$, (46) gives

$$\delta\Big[v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_s) + \widetilde{\theta}_1(\widetilde{\mathbf{W}}_{55})_{1s}\Big]\lambda = p_1(\delta\widetilde{\mathbf{\Theta}}_4)_{1s}.$$

Thus (47) yields, for $1 \le s < T$,

$$\delta R_{s+1} = (1+\widetilde{\theta}_2\lambda)\delta R_s + \sum_{\ell=1}^{s}(\widetilde{\mathbf{\Theta}}_4)_{\ell s}\delta R_\ell + \sum_{\ell=1}^{s}p_\ell(\delta\widetilde{\mathbf{\Theta}}_4)_{\ell s}. \tag{96}$$

For $s = 1,\dots,T-1$, choose successively

$$(\delta\widetilde{\mathbf{\Theta}}_4)_{1:s,s} = \mathbf{M}_s^{-1}\operatorname{coeff}_{1:s}\Big[\mathbb{1}\{s+1=m\}q - (1+\widetilde{\theta}_2\lambda)\delta R_s - \sum_{\ell=1}^{s}(\widetilde{\mathbf{\Theta}}_4)_{\ell s}\delta R_\ell\Big].$$

Substitution into (96) gives

$$\delta R_s = \mathbb{1}\{s=m\}q, \qquad 1 \le s \le T.$$

Since $\widetilde{\theta}_1 \neq 0$ implies $v_{31} \neq 0$, and Lemma D.2 gives $v_{33} > 0$, take

$$\delta(\mathbf{w}_{25}\mathbf{e}_s) = \frac{\widetilde{\theta}_1}{v_{31}}(\delta\widetilde{\mathbf{\Theta}}_4)_{1s}, \quad 1 \le s < T,$$
$$\delta(\mathbf{W}_{55})_{\ell s} = \frac{1}{v_{33}}(\delta\widetilde{\mathbf{\Theta}}_4)_{\ell s}, \quad 2 \le \ell \le s < T,$$

and set every other raw variation to zero. Then

$$D_{\bar{\boldsymbol{\theta}}}\left(\frac{\mathcal{L}}{d}\right)[\delta\bar{\boldsymbol{\theta}}] = D_{\widetilde{\mathbf{\Theta}}_4}\left(\frac{\mathcal{L}}{d}\right)[\delta\widetilde{\mathbf{\Theta}}_4]$$
$$= \frac{1}{T}\sum_{s=1}^{T}\langle R_s, \delta R_s\rangle_{d,n} = \frac{1}{T}\langle R_m, q\rangle_{d,n}.$$

Conversely, for every $\delta\widetilde{\mathbf{\Theta}}_4$,

$$\delta R_s(0) = 0, \qquad \deg \delta R_s \le s-1,$$
$$(93) \implies D_{\widetilde{\mathbf{\Theta}}_4}\left(\frac{\mathcal{L}}{d}\right)[\delta\widetilde{\mathbf{\Theta}}_4] = \frac{1}{T}\sum_{s=1}^{T}\langle R_s, \delta R_s\rangle_{d,n} = 0.$$

Therefore

$$D_{\widetilde{\mathbf{\Theta}}_4}\left(\frac{\mathcal{L}}{d}\right) = 0 \iff (93).$$

(iii) The displayed orthogonality is exactly the normal equation for (94), so the Hilbert-space projection theorem gives existence and uniqueness. The fixed leading coefficient is a smooth function of $\widetilde{\theta}_1, (-\widetilde{\theta}_2)$, and the positive Gram matrix makes the minimizing lower coefficients smooth functions of them. Differentiating the minimized quadratic and using the normal equations to remove the derivatives of those lower coefficients gives (95). □

Corollary F.2. *If $\widetilde{\theta}_1\widetilde{\theta}_2 \neq 0$ and (93) holds, then*

$$D^2_{\widetilde{\mathbf{\Theta}}_4}\left(\frac{\mathcal{L}}{d}\right)[\delta\widetilde{\mathbf{\Theta}}_4, \delta\widetilde{\mathbf{\Theta}}_4] = \frac{1}{T}\sum_{t=1}^{T}\left\|D_{\widetilde{\mathbf{\Theta}}_4}R_t[\delta\widetilde{\mathbf{\Theta}}_4]\right\|^2_{d,n} > 0 \quad (\delta\widetilde{\mathbf{\Theta}}_4 \neq 0). \tag{97}$$

The Hessian of the minimized function is

$$D^2_{(\widetilde{\theta}_1,-\widetilde{\theta}_2)}\left[\min_{\widetilde{\mathbf{\Theta}}_4}\frac{\mathcal{L}}{d}\right]$$
$$= D^2_{(\widetilde{\theta}_1,-\widetilde{\theta}_2)}\frac{\mathcal{L}}{d} - D^2_{(\widetilde{\theta}_1,-\widetilde{\theta}_2),\widetilde{\mathbf{\Theta}}_4}\frac{\mathcal{L}}{d}\left(D^2_{\widetilde{\mathbf{\Theta}}_4}\frac{\mathcal{L}}{d}\right)^{-1}D^2_{\widetilde{\mathbf{\Theta}}_4,(\widetilde{\theta}_1,-\widetilde{\theta}_2)}\frac{\mathcal{L}}{d}. \tag{98}$$

Consequently,

$$D^2_{\bar{\boldsymbol{\theta}}'}\frac{\mathcal{L}}{d} \succ 0 \iff D^2_{(\widetilde{\theta}_1,-\widetilde{\theta}_2)}\frac{\min_{\widetilde{\mathbf{\Theta}}_4}\mathcal{L}}{d} \succ 0.$$

Proof. The leading coefficient $\widetilde{\theta}_1\widetilde{\theta}_2^{t-1}$ does not depend on $\widetilde{\mathbf{\Theta}}_4$. For $j = 1, 2$,

$$\left(D^j_{\widetilde{\mathbf{\Theta}}_4}R_t\right)(0) = 0, \qquad \deg\left(D^j_{\widetilde{\mathbf{\Theta}}_4}R_t\right) \le t-1, \qquad \left\langle R_t, D^2_{\widetilde{\mathbf{\Theta}}_4}R_t[\delta\widetilde{\mathbf{\Theta}}_4, \delta\widetilde{\mathbf{\Theta}}_4]\right\rangle_{d,n} = 0.$$

Hence

$$D^2_{\widetilde{\mathbf{\Theta}}_4}\frac{\mathcal{L}}{d}[\delta\widetilde{\mathbf{\Theta}}_4, \delta\widetilde{\mathbf{\Theta}}_4] = \frac{1}{T}\sum_{t=1}^{T}\Big\{\|D_{\widetilde{\mathbf{\Theta}}_4}R_t[\delta\widetilde{\mathbf{\Theta}}_4]\|^2_{d,n}$$
$$+ \langle R_t, D^2_{\widetilde{\mathbf{\Theta}}_4}R_t[\delta\widetilde{\mathbf{\Theta}}_4, \delta\widetilde{\mathbf{\Theta}}_4]\rangle_{d,n}\Big\}.$$

Moreover, (73) gives

$$\det(D_{\widetilde{\mathbf{\Theta}}_4}\widetilde{\theta}_5) = \prod_{t=1}^{T-1} \det \mathbf{M}_t = \prod_{t=1}^{T-1}\prod_{\ell=1}^{t} \widetilde{\theta}_1 \widetilde{\theta}_2^{\ell-1} \neq 0.$$

Thus (97) holds and the coefficient Hessian is invertible. Differentiating its normal equations gives

$$D_{\widetilde{\mathbf{\Theta}}_4}\mathcal{L}\left(\widetilde{\theta}_1, \widetilde{\theta}_2, \arg\min_{\widetilde{\mathbf{\Theta}}_4} \mathcal{L}\right) = 0,$$
$$D^2_{\widetilde{\mathbf{\Theta}}_4}\frac{\mathcal{L}}{d} D_{(\widetilde{\theta}_1,-\widetilde{\theta}_2)}\left(\arg\min_{\widetilde{\mathbf{\Theta}}_4} \mathcal{L}\right) + D^2_{\widetilde{\mathbf{\Theta}}_4,(\widetilde{\theta}_1,-\widetilde{\theta}_2)}\frac{\mathcal{L}}{d} = 0.$$

Substitution proves (98). Finally,

$$\begin{aligned} &D^2\frac{\mathcal{L}}{d}[\delta\widetilde{\boldsymbol{\theta}}', \delta\widetilde{\boldsymbol{\theta}}'] \\ &= D^2_{(\widetilde{\theta}_1,-\widetilde{\theta}_2)}\frac{\min_{\widetilde{\mathbf{\Theta}}_4}\mathcal{L}}{d}\left[\begin{pmatrix}\delta\widetilde{\theta}_1\\ -\delta\widetilde{\theta}_2\end{pmatrix}, \begin{pmatrix}\delta\widetilde{\theta}_1\\ -\delta\widetilde{\theta}_2\end{pmatrix}\right] \\ &\quad + \left\| \left(D^2_{\widetilde{\mathbf{\Theta}}_4}\frac{\mathcal{L}}{d}\right)^{1/2} \left((\delta\widetilde{\mathbf{\Theta}}_4)_{\ell t}\right)_{1\le\ell\le t<T} + \left(D^2_{\widetilde{\mathbf{\Theta}}_4}\frac{\mathcal{L}}{d}\right)^{-1/2} D^2_{\widetilde{\mathbf{\Theta}}_4,(\widetilde{\theta}_1,-\widetilde{\theta}_2)}\frac{\mathcal{L}}{d}\begin{pmatrix}\delta\widetilde{\theta}_1\\ -\delta\widetilde{\theta}_2\end{pmatrix}\right\|_2^2. \end{aligned}$$

□

Lemma F.3. *Let τ be any Stage-1 time at which $\mathcal{L}_{\mathrm{GD}}(\tau) \le d^{-1/4}$, and keep exactly the same raw parameters as the initial parameters of Stage 2, resetting only the time origin. Then $\mathbf{W}_{33} = \mathbf{0}_{d\times d}$, scalar-identity invariance is preserved, and $v_{33}(\tau) \ge v_{33}(0) > 0$. At this endpoint,*

$$\begin{aligned} &[v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0) - \eta_0]^2 + [-v_{31}w_{13} - \eta_0]^2 \\ &\quad + \sum_{t=1}^{T-1}\left|v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t) + v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0)(\widetilde{\mathbf{W}}_{55})_{1t} - \eta_0\right|^2 \\ &\quad + \sum_{t=1}^{T-1}\sum_{\ell=2}^{t}(\widetilde{\mathbf{W}}_{55})_{\ell t}^2 \le d^{-1/4+o(1)}. \end{aligned} \tag{99}$$

The autoregressively generated polynomials from the unchanged raw endpoint therefore satisfy

$$\frac{1}{2Td}\sum_{t=1}^{T}\mathbb{E}\ \mathrm{tr}\Big(\left[p_t(\mathbf{S}) - \{\mathbf{I} - (\mathbf{I} - \eta_0\mathbf{S})^t\}\right]^2\Big) \le d^{-1/4+o(1)}. \tag{100}$$

Consequently its Stage-2 outcome loss obeys

$$\sqrt{\frac{\mathcal{L}(0)}{d}} \le \left\{\frac{1}{2Td}\sum_{t=1}^{T}\mathbb{E}\ \mathrm{tr}\big[(\mathbf{I} - \eta_0\mathbf{S})^{2t}\big]\right\}^{1/2} + d^{-1/8+o(1)}. \tag{101}$$

Proof. The scalar-identity statement and the lower bound on v_{33} are the invariant identities of Lemmas C.3 and C.1. Lemma G.14 and $\mathcal{L}_{\mathrm{GD}}(\tau) \le d^{-1/4}$ give

$$\begin{aligned} &[v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0) - \eta_0]^2 + [-v_{31}w_{13} - \eta_0]^2 \\ &\quad + \sum_{t=1}^{T-1}|v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t) + \eta_0(\widetilde{\mathbf{W}}_{55})_{1t} - \eta_0|^2 + \sum_{t=1}^{T-1}\sum_{\ell=2}^{t}(\widetilde{\mathbf{W}}_{55})_{\ell t}^2 \le d^{-5/4+o(1)}. \end{aligned} \tag{102}$$

This controls the teacher-forced coefficient $\eta_0(\widetilde{\mathbf{W}}_{55})_{1t}$. The autoregressive recurrence instead contains $v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0)(\widetilde{\mathbf{W}}_{55})_{1t}$. Lemma G.15 gives

$$\sum_{t=1}^{T-1}(\widetilde{\mathbf{W}}_{55})_{1t}^2 \le d^{1+o(1)}.$$

Therefore

$$\sum_{t=1}^{T-1}\Big|v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)+v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)(\widetilde{\mathbf{W}}_{55})_{1t}-\eta_0\Big|^2$$
$$\le 2\sum_{t=1}^{T-1}|v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)+\eta_0(\widetilde{\mathbf{W}}_{55})_{1t}-\eta_0|^2$$
$$+2|v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)-\eta_0|^2\sum_{t=1}^{T-1}(\widetilde{\mathbf{W}}_{55})_{1t}^2\le d^{-1/4+o(1)}.$$

Together with (102), this proves (99). In particular, both leading products are positive for all sufficiently large d.

We compare p_t with $1-(1-\eta_0\lambda)^t$. Subtract at each step the exact GD identity

$$\mathbf{I}-(\mathbf{I}-\eta_0\mathbf{S})^{t+1}=(\mathbf{I}-\eta_0\mathbf{S})\{\mathbf{I}-(\mathbf{I}-\eta_0\mathbf{S})^t\}+\eta_0\mathbf{S}$$

from the recurrence in Lemma D.1. The difference satisfies the exact identity

$$\begin{aligned}
&p_{t+1}(\lambda)-\{1-(1-\eta_0\lambda)^{t+1}\}\\
&=(1+v_{31}w_{13}\lambda)\big[p_t(\lambda)-\{1-(1-\eta_0\lambda)^t\}\big]\\
&\quad+(\eta_0+v_{31}w_{13})\lambda\{1-(1-\eta_0\lambda)^t\}\\
&\quad+\big[v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)+v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)(\widetilde{\mathbf{W}}_{55})_{1t}-\eta_0\big]\lambda\\
&\quad+\sum_{\ell=2}^{t}(\widetilde{\mathbf{W}}_{55})_{\ell t}\big[p_\ell(\lambda)-\{1-(1-\eta_0\lambda)^\ell\}\big]+\sum_{\ell=2}^{t}(\widetilde{\mathbf{W}}_{55})_{\ell t}\{1-(1-\eta_0\lambda)^\ell\}.
\end{aligned}\tag{103}$$

The initial difference is $[v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)-\eta_0]\lambda$. By (99), for all sufficiently large d,

$$|v_{31}w_{13}|\le 2,\qquad \sum_{t=1}^{T-1}\sum_{\ell=2}^{t}(\widetilde{\mathbf{W}}_{55})_{\ell t}^2\le d^{-1/4+o(1)}.$$

Moreover, the sum of the absolute monomial coefficients of $1-(1-\eta_0\lambda)^\ell$ is at most $2^\ell+1$. Taking Euclidean norms of the monomial coefficient vectors in (103), applying Cauchy-Schwarz to each sum over ℓ, and inducting on t therefore gives

$$\sum_{t=1}^{T}\sum_{j=0}^{t}\big|[\lambda^j]\big[p_t-\{1-(1-\eta_0\lambda)^t\}\big]\big|^2\le\exp\{CT\log(e+T)\}\,[\text{left side of (99)}]\,.$$

To see the induction explicitly, multiplication by $1+v_{31}w_{13}\lambda$ increases a coefficient norm by at most three; the sum multiplying earlier differences has norm at most $\sqrt{T}\,d^{-1/8+o(1)}<1$; and all remaining terms in the t-th row have coefficient norm at most $C2^T$ times the square root of the left side of (99). Thus the maximum accumulated coefficient norm is multiplied by at most $C(t+1)$ at each of at most T steps, which is the displayed exponential factor. Finally, the finite-Wishart moment matrix through degree T has operator norm at most $\exp\{CT\log(e+T)\}$ by Lemma G.1. Both exponential factors are $d^{o(1)}$ under (17); together with (99) this proves (100).

Finally, view the T matrix polynomials as one vector in their Hilbert direct sum with squared norm $(2Td)^{-1}\sum_{t=1}^{T}\mathbb{E}\,\mathrm{tr}\big((\cdot)^2\big)$. Since

$$p_t(\mathbf{S})-\mathbf{I}=\big[p_t(\mathbf{S})-\{\mathbf{I}-(\mathbf{I}-\eta_0\mathbf{S})^t\}\big]-(\mathbf{I}-\eta_0\mathbf{S})^t,$$

Minkowski's inequality and (100) give (101). Every estimate concerns the unchanged raw endpoint; no raw-parameter comparison or re-gauging is used. □

Lemma F.4. *For all sufficiently large d in the fixed-aspect-ratio regime (17), the finite-Wishart transfer estimate in Lemma G.5 gives the three strict boundary inequalities below for the sublevel reached in Lemma E.2.*

(i)

$$\frac{1}{2T}\left\{\min_{R(0)=-1,\deg R\le 1}\|R\|_{d,n}^2+\sum_{t=2}^{T}\min_{\substack{R(0)=-1,\deg R\le t\\ [\lambda^t]R=0}}\|R\|_{d,n}^2\right\}>\frac{3d}{4nT}. \tag{104}$$

(ii)

$$\frac{1}{2T}\sum_{t=1}^{T}\min_{\substack{R(0)=-1,\deg R\le t\\ [\lambda^t]R=0}}\|R\|_{d,n}^2>\frac{3d}{4nT}. \tag{105}$$

(iii)

$$\frac{1}{2T}\left\{\sum_{t=1}^{T-1}\min_{\substack{R(0)=-1,\deg R\le t\\ [\lambda^t]R=0}}\|R\|_{d,n}^2+\min_{\substack{R(0)=-1,\deg R\le T\\ [\lambda^T]R\text{ free}}}\|R\|_{d,n}^2\right\}>\frac{3d}{4nT}. \tag{106}$$

The sublevel $\mathcal{L}/d\le 3d/(4nT)$ is therefore bounded away from $\widetilde{\theta}_1=0,(-\widetilde{\theta}_2)=0,|(-\widetilde{\theta}_2)|=\infty,|\widetilde{\theta}_1|=\infty$.

Proof. **(i)** For the first boundary, the first MP minimum is $(d/n)/(1+d/n)$, while the t-th minimum in the remaining sum is $(d/n)^{t-1}/\sum_{r=0}^{t-1}(d/n)^r$. Since $2/(1+d/n)-3/2\ge 1/4$, these formulas give

$$\frac{1}{2T}\left\{\frac{d/n}{1+d/n}+\sum_{k=1}^{T-1}\frac{(d/n)^k}{\sum_{r=0}^{k}(d/n)^r}\right\}-\frac{3d}{4nT}>\frac{(d/n)^{3/2}}{8T}. \tag{107}$$

For the $\widetilde{\theta}_1\to 0$ boundary, the same formula gives

$$\frac{1}{2T}\sum_{k=0}^{T-1}\frac{(d/n)^k}{\sum_{r=0}^{k}(d/n)^r}-\frac{1}{2T}\left\{\frac{d/n}{1+d/n}+\sum_{k=1}^{T-1}\frac{(d/n)^k}{\sum_{r=0}^{k}(d/n)^r}\right\}=\frac{1}{2T(1+d/n)}>\frac{d/n}{2T}. \tag{108}$$

For $|(-\widetilde{\theta}_2)|\to\infty$, the limiting boundary is

$$\frac{1}{2T}\left\{\sum_{k=0}^{T-2}\frac{(d/n)^k}{\sum_{r=0}^{k}(d/n)^r}+\frac{(d/n)^T}{\sum_{r=0}^{T}(d/n)^r}\right\}.$$

Subtracting the boundary value for $(-\widetilde{\theta}_2)\to 0$ gives

$$\begin{aligned}&\frac{1}{2T}\left\{\sum_{k=0}^{T-2}\frac{(d/n)^k}{\sum_{r=0}^{k}(d/n)^r}+\frac{(d/n)^T}{\sum_{r=0}^{T}(d/n)^r}\right\}\\&\quad-\frac{1}{2T}\left\{\frac{d/n}{1+d/n}+\sum_{k=1}^{T-1}\frac{(d/n)^k}{\sum_{r=0}^{k}(d/n)^r}\right\}\\&=\frac{1}{2T}\left\{1-\frac{d/n}{1+d/n}-\frac{(d/n)^{T-1}}{\sum_{r=0}^{T-1}(d/n)^r}+\frac{(d/n)^T}{\sum_{r=0}^{T}(d/n)^r}\right\}>\frac{1}{4T}.\end{aligned} \tag{109}$$

Here $T\ge 8$ and $d/n\le 9/20$. Together with (107), this places the boundary at infinity above $3d/(4nT)$.

(ii) Apply (157) with $f=g=R$ and take the same constrained minima under both norms. Each MP minimum is at most $\|-1\|_{\mathrm{MP}}^2=1$, since $R=-1$ is feasible. Equation (159) therefore bounds the change in each displayed averaged boundary value by $(d/n)^{3/2}/(8192T)$. Equation (107) places the entry region below the smallest MP boundary value by more than $(d/n)^{3/2}/(8T)$; the other two MP gaps in (108)-(109) are larger. In particular, the first transferred finite boundary value exceeds

the entry region by at least $(d/n)^{3/2}/(32T)$ for all sufficiently large d. Thus all three transferred finite boundary values remain strictly above that sublevel.

(iii) The recurrence fixes $[\lambda^t]R_t = (-1)^{t-1}\widetilde{\theta}_1(-\widetilde{\theta}_2)^{t-1}$, and the projection theorem gives

$$\frac{\mathcal{L}}{d} \geq \frac{1}{2T}\sum_{t=1}^{T} \min_{\substack{R(0)=-1,\deg R\leq t \\ [\lambda^t]R=(-1)^{t-1}\widetilde{\theta}_1(-\widetilde{\theta}_2)^{t-1}}} \|R\|_{d,n}^2. \tag{110}$$

Every term is a coercive quadratic in its leading coefficient, so $\widetilde{\theta}_1, \widetilde{\theta}_1(-\widetilde{\theta}_2), \ldots, \widetilde{\theta}_1(-\widetilde{\theta}_2)^{T-1}$ are bounded on any bounded-loss sequence. If $|(-\widetilde{\theta}_2)| \to \infty$, all products except possibly the last tend to zero; freeing that last coefficient gives the lower limit in (106), which excludes this case first. With $(-\widetilde{\theta}_2)$ now bounded, $\widetilde{\theta}_1 \to 0$ makes all products vanish and gives (105). Finally $(-\widetilde{\theta}_2) \to 0$ makes all products except the first vanish; minimizing over the remaining first product gives (104). These are lower bounds on limiting losses, not equalities for arbitrary parameter paths. The first quadratic already excludes $|\widetilde{\theta}_1| \to \infty$. □

Lemma F.5. *Under the conditions of Lemma F.4, for all sufficiently large d, the region*

$$\left\{\widetilde{\boldsymbol{\theta}}' : \widetilde{\theta}_1 \neq 0, \quad \frac{\mathcal{L}(\widetilde{\boldsymbol{\theta}}')}{d} \leq \frac{3d}{4nT}\right\}$$

is compact and satisfies

$$\frac{13}{20} < \widetilde{\theta}_1 \leq \frac{7}{5}, \qquad \frac{1}{14} < -\widetilde{\theta}_2 < \frac{38}{13}, \qquad \frac{d}{n}(-\widetilde{\theta}_2)^2 < \frac{722}{845} < 1.$$

It contains exactly one auxiliary critical point $\widetilde{\boldsymbol{\theta}}'^{\star}$, with

$$\mathcal{L}(\widetilde{\boldsymbol{\theta}}'^{\star}) = \mathcal{L}^{\star}, \qquad D^2_{\widetilde{\boldsymbol{\theta}}'}\mathcal{L}(\widetilde{\boldsymbol{\theta}}'^{\star}) \succ 0.$$

Proof. **(i)** By (157) and the first two constrained MP minima, throughout this region,

$$(1-\widetilde{\theta}_1)^2 + \frac{d}{n}\widetilde{\theta}_1^2 + \frac{d/n}{1+d/n}\{1-\widetilde{\theta}_1(-\widetilde{\theta}_2)\}^2 + \left(\frac{d}{n}\right)^2 \widetilde{\theta}_1^2(-\widetilde{\theta}_2)^2$$
$$\leq \sum_{t=1}^{T}\|R_t\|_{\mathrm{MP}}^2 \leq \frac{1}{1-\varepsilon_T}\frac{2T\mathcal{L}}{d} \leq \frac{8}{5}\frac{d}{n}.$$

For $\widetilde{\theta}_1 \leq 13/20$,

$$(1-\widetilde{\theta}_1)^2 + \frac{d}{n}\widetilde{\theta}_1^2 - \frac{8}{5}\frac{d}{n} \geq \frac{49-471d/n}{400} \geq \frac{19}{4000} > 0.$$

Also, $(1-\widetilde{\theta}_1)^2 \leq 8d/(5n) \leq 4/25$. Thus

$$\frac{13}{20} < \widetilde{\theta}_1 \leq \frac{7}{5}.$$

Subtracting the minimum $(d/n)/(1+d/n)$ of the first quadratic gives

$$\{1-\widetilde{\theta}_1(-\widetilde{\theta}_2)\}^2 \leq \frac{3}{5} + \frac{8}{5}\frac{d}{n} \leq \frac{19}{25} < \frac{81}{100}.$$

Therefore

$$\frac{1}{10} < \widetilde{\theta}_1(-\widetilde{\theta}_2) < \frac{19}{10}, \qquad \frac{1}{14} < -\widetilde{\theta}_2 < \frac{38}{13}, \qquad \frac{d}{n}(-\widetilde{\theta}_2)^2 < \frac{722}{845}.$$

The finite Gram bound and the triangular inverse in (72) now imply

$$\sum_{t=1}^{T}\|R_t\|_{\mathrm{coeff}}^2 \leq d^{o(1)}, \qquad \sum_{1\leq \ell\leq t<T} |(\widetilde{\boldsymbol{\Theta}}_4)_{\ell t}|^2 \leq d^{o(1)}.$$

Since $\mathcal{L}$ is continuous and $\widetilde{\theta}_1 > 13/20$ on this region,

$$\left\{\widetilde{\boldsymbol{\theta}}' : \widetilde{\theta}_1 \neq 0, \quad \mathcal{L}(\widetilde{\boldsymbol{\theta}}')/d \leq 3d/(4nT)\right\} \quad \text{is closed and bounded, hence compact.}$$

It is nonempty by Lemma E.2. At a minimizer in this region, a nonzero auxiliary gradient would give, for small $s > 0$,

$$\mathcal{L}\left(\widetilde{\boldsymbol{\theta}}' - sD_{\widetilde{\boldsymbol{\theta}}'}\mathcal{L}\right) = \mathcal{L}(\widetilde{\boldsymbol{\theta}}') - s\|D_{\widetilde{\boldsymbol{\theta}}'}\mathcal{L}\|_2^2 + o(s) < \mathcal{L}(\widetilde{\boldsymbol{\theta}}'),$$

a contradiction. Hence this region contains a critical point.

(ii) By Lemma F.1,

$$2T\,\frac{\min_{\widetilde{\boldsymbol{\Theta}}_4}\mathcal{L}}{d} = \sum_{k=0}^{T-1} q_{k,\nu_{d,n}} - 2\widetilde{\theta}_1 B_{\nu_{d,n}}(-\widetilde{\theta}_2) + \widetilde{\theta}_1^2 C_{\nu_{d,n}}(-\widetilde{\theta}_2). \tag{111}$$

Thus every critical point in the region satisfies

$$\widetilde{\theta}_1 = \frac{B_{\nu_{d,n}}(-\widetilde{\theta}_2)}{C_{\nu_{d,n}}(-\widetilde{\theta}_2)}, \qquad \left(\frac{B_{\nu_{d,n}}^2}{C_{\nu_{d,n}}}\right)'(-\widetilde{\theta}_2) = 0, \qquad \widetilde{\theta}_5 = \widetilde{\theta}_5^\star.$$

Lemma G.8 and the invertibility of $D_{\widetilde{\boldsymbol{\Theta}}_4}\widetilde{\theta}_5$ imply

$$\widetilde{\boldsymbol{\theta}}' = \widetilde{\boldsymbol{\theta}}'^\star, \qquad \widetilde{\theta}_1^\star = \frac{B_{\nu_{d,n}}(-\widetilde{\theta}_2^\star)}{C_{\nu_{d,n}}(-\widetilde{\theta}_2^\star)} > 0, \qquad \left(\frac{B_{\nu_{d,n}}^2}{C_{\nu_{d,n}}}\right)''(-\widetilde{\theta}_2^\star) < 0.$$

At this point, differentiating (111) and completing the square gives

$$\begin{aligned} &D^2_{(\widetilde{\theta}_1,-\widetilde{\theta}_2)}\frac{\min_{\widetilde{\boldsymbol{\Theta}}_4}\mathcal{L}}{d}\left[\begin{pmatrix}\delta\widetilde{\theta}_1\\ -\delta\widetilde{\theta}_2\end{pmatrix},\begin{pmatrix}\delta\widetilde{\theta}_1\\ -\delta\widetilde{\theta}_2\end{pmatrix}\right] \\ &\quad= \frac{C_{\nu_{d,n}}(-\widetilde{\theta}_2^\star)}{T}\left\{\delta\widetilde{\theta}_1 + \left(\frac{B_{\nu_{d,n}}}{C_{\nu_{d,n}}}\right)'(-\widetilde{\theta}_2^\star)\delta\widetilde{\theta}_2\right\}^2 \\ &\qquad - \frac{1}{2T}\left(\frac{B_{\nu_{d,n}}^2}{C_{\nu_{d,n}}}\right)''(-\widetilde{\theta}_2^\star)(\delta\widetilde{\theta}_2)^2 > 0 \quad \text{if } (\delta\widetilde{\theta}_1,\delta\widetilde{\theta}_2) \neq (0,0). \end{aligned}$$

Corollary F.2 therefore gives $D^2_{\widetilde{\boldsymbol{\theta}}'}\mathcal{L}(\widetilde{\boldsymbol{\theta}}'^\star) \succ 0$.

(iii) Take a minimizing sequence with

$$\mathcal{L}(\widetilde{\boldsymbol{\theta}}'_k)/d \leq \frac{3d}{4nT}, \qquad \mathcal{L}(\widetilde{\boldsymbol{\theta}}'_k) \longrightarrow \inf_{\widetilde{\boldsymbol{\theta}}' : \widetilde{\theta}_1 \neq 0}\mathcal{L}.$$

Part (i) and continuity give a subsequence satisfying

$$\widetilde{\boldsymbol{\theta}}'_{k_j} \longrightarrow \widetilde{\boldsymbol{\theta}}', \qquad \mathcal{L}(\widetilde{\boldsymbol{\theta}}') = \inf_{\widetilde{\boldsymbol{\theta}}' : \widetilde{\theta}_1 \neq 0}\mathcal{L}, \qquad D_{\widetilde{\boldsymbol{\theta}}'}\mathcal{L}(\widetilde{\boldsymbol{\theta}}') = 0.$$

By part (ii),

$$\widetilde{\boldsymbol{\theta}}' = \widetilde{\boldsymbol{\theta}}'^\star.$$

Lemma G.16 gives

$$\forall\, \widetilde{\boldsymbol{\theta}}' \text{ with } \widetilde{\theta}_1 \neq 0, \quad \exists\, \bar{\boldsymbol{\theta}} \text{ satisfying (11), (51)–(53)} : \quad \widetilde{\boldsymbol{\theta}}'(\bar{\boldsymbol{\theta}}) = \widetilde{\boldsymbol{\theta}}'.$$

At raw points with $\widetilde{\theta}_1 = 0$,

$$p_1 = 0, \qquad \frac{\mathcal{L}(\bar{\boldsymbol{\theta}})}{d} \geq \frac{1}{2T} > \frac{3d}{4nT}.$$

Consequently,

$$\mathcal{L}(\widetilde{\boldsymbol{\theta}}'^\star) = \min_{\widetilde{\boldsymbol{\theta}}' : \widetilde{\theta}_1 \neq 0}\mathcal{L} = \min_{\substack{\bar{\boldsymbol{\theta}} \text{ satisfying (11)} \\ \text{and the inherited conservation laws}}} \mathcal{L}(\bar{\boldsymbol{\theta}}) = \mathcal{L}^\star.$$

□

F.2 Raw-parameter estimates

Lemma F.6. *Under (17), along the Stage-2 flow, every variation of the active auxiliary parameters has a raw lift satisfying*

$$(\delta v_{31})^2 + (\delta w_{13})^2 + (\delta v_{33})^2 + \frac{1}{d}\{(\delta w_{24})^2 + \|\delta \mathbf{w}_{25}\|^2 + \|\delta w_{44}\|_F^2 + \|\delta \mathbf{w}_{45}\|_F^2\} + \frac{1}{d}\{\|\delta \mathbf{w}_{54}\|_F^2 + \|\delta \mathbf{W}_{55}\|_F^2\} \le C\big(1 + (d/n)^{-1/2} + v_{31}^2 + v_{33}^2\big)\left\{|\delta\widetilde{\theta}_1|^2 + |\delta\widetilde{\theta}_2|^2 + \frac{1}{d}\|\delta\widetilde{\mathbf{\Theta}}_4\|_F^2\right\}. \tag{112}$$

Proof. Choose

$$\delta v_{31} = \frac{v_{31}}{\widetilde{\theta}_1}\delta\widetilde{\theta}_1, \qquad \delta w_{13} = \frac{\delta\widetilde{\theta}_2 - (\widetilde{\theta}_2/\widetilde{\theta}_1)\delta\widetilde{\theta}_1}{v_{31}}, \tag{113}$$
$$\delta v_{33} = \delta w_{24} = \delta \mathbf{w}_{25} = \delta w_{44} = \delta \mathbf{w}_{45} = \delta \mathbf{w}_{54} = 0,$$

and, for $1 \le t < T$,

$$\delta(\mathbf{W}_{55})_{\ell t} = \frac{(\delta\widetilde{\mathbf{\Theta}}_4)_{\ell t}}{v_{33}}, \qquad 1 \le \ell \le t. \tag{114}$$

All other entries of $\delta\mathbf{W}_{55}$ are zero. Substitution into (45) and (46) gives

$$\begin{aligned} D\widetilde{\theta}_1[\delta\bar{\boldsymbol{\theta}}] &= \frac{\widetilde{\theta}_1}{v_{31}}\delta v_{31} = \delta\widetilde{\theta}_1, \\ D\widetilde{\theta}_2[\delta\bar{\boldsymbol{\theta}}] &= w_{13}\delta v_{31} + v_{31}\delta w_{13} = \delta\widetilde{\theta}_2, \\ D(\widetilde{\mathbf{\Theta}}_4)_{1t}[\delta\bar{\boldsymbol{\theta}}] &= v_{33}\delta(\mathbf{W}_{55})_{1t} = (\delta\widetilde{\mathbf{\Theta}}_4)_{1t}, \\ D(\widetilde{\mathbf{\Theta}}_4)_{\ell t}[\delta\bar{\boldsymbol{\theta}}] &= v_{33}\delta(\mathbf{W}_{55})_{\ell t} = (\delta\widetilde{\mathbf{\Theta}}_4)_{\ell t}, \qquad 2 \le \ell \le t < T. \end{aligned}$$

Lemmas E.2 and F.5, together with (72), give

$$c \le \widetilde{\theta}_1 \le C, \qquad c\sqrt{d/n} \le -\widetilde{\theta}_2 \le C, \qquad \|\widetilde{\mathbf{\Theta}}_4\|_F^2 \le \Lambda_T.$$

The inherited invariants imply

$$v_{31}^4 + \frac{\sigma^2}{d}v_{31}^2 \ge \widetilde{\theta}_2^2 \ge c(d/n), \qquad |v_{31}| \ge c(d/n)^{1/4},$$
$$v_{33}^2 = (v_{33}^{\mathrm{init}})^2 + \frac{1}{d}\sum_{B\in\{44,45,54,55\}}\|\mathbf{W}_B\|_F^2, \qquad v_{33} \ge v_{33}^{\mathrm{init}} \ge c.$$

The first lower bound uses $\sigma^2/d = o(\sqrt{d/n})$. Thus

$$|\delta v_{31}|^2 + |\delta w_{13}|^2 \le C\big(v_{31}^2 + (d/n)^{-1/2}\big)\big(|\delta\widetilde{\theta}_1|^2 + |\delta\widetilde{\theta}_2|^2\big),$$
$$\frac{1}{d}\|\delta\mathbf{W}_{55}\|_F^2 \le \frac{C}{d}\|\delta\widetilde{\mathbf{\Theta}}_4\|_F^2.$$

These inequalities prove (112). □

Lemma F.7. *In the fixed-aspect-ratio regime of Theorem 3.1, there is a constant $c > 0$ such that along the Stage-2 flow,*

$$-\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathcal{L}}{d} \ge \frac{c(d/n)}{T\big(1 + (d/n)^{-1/2} + v_{31}^2 + v_{33}^2\big)}\left\{\frac{\mathcal{L} - \mathcal{L}^*}{d}\right\}. \tag{115}$$

Proof. For fixed $\widetilde{\theta}_1, \widetilde{\theta}_2$, let $R_{t,\mathrm{ls}}$ be the constrained least-squares residual. The normal equations give

$$\langle R_{t,\mathrm{ls}}, R_t - R_{t,\mathrm{ls}}\rangle_{d,n} = 0, \qquad \|R_t\|_{d,n}^2 = \|R_{t,\mathrm{ls}}\|_{d,n}^2 + \|R_t - R_{t,\mathrm{ls}}\|_{d,n}^2.$$

Hence

$$\frac{\mathcal{L}-\mathcal{L}^*}{d}=\frac{1}{2T}\sum_{t=1}^{T}\|R_t-R_{t,\mathrm{ls}}\|_{d,n}^2+\frac{\min_{\widetilde{\mathbf{\Theta}}_4}\mathcal{L}-\mathcal{L}^*}{d}. \tag{116}$$

Equations (69) and (73) imply

$$\left\|D_{\widetilde{\mathbf{\Theta}}_4}\frac{\mathcal{L}}{d}\right\|_F^2\geq\frac{1}{T^2\Lambda_T^2}\sum_{t=1}^{T}\|R_t-R_{t,\mathrm{ls}}\|_{d,n}^2.$$

For the raw blocks $\mathbf{w}_{25}$ and $\mathbf{W}_{55}$,

$$D_{(\mathbf{w}_{25})_t}\frac{\mathcal{L}}{d}=\frac{v_{31}}{\widetilde{\theta}_1}D_{(\widetilde{\mathbf{\Theta}}_4)_{1t}}\frac{\mathcal{L}}{d},\quad 1\leq t<T,$$
$$D_{(\mathbf{W}_{55})_{\ell t}}\frac{\mathcal{L}}{d}=v_{33}D_{(\widetilde{\mathbf{\Theta}}_4)_{\ell t}}\frac{\mathcal{L}}{d},\quad 2\leq\ell\leq t<T.$$

Using $|v_{31}|^2\geq c\sqrt{d/n}$ and $v_{33}\geq c$ in the energy identity therefore gives

$$\begin{aligned}-\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathcal{L}}{d}&\geq cd\sqrt{d/n}\left\|D_{\widetilde{\mathbf{\Theta}}_4}\frac{\mathcal{L}}{d}\right\|_F^2\\&\geq\frac{d^{1-o(1)}}{T}\,\frac{1}{2T}\sum_{t=1}^{T}\|R_t-R_{t,\mathrm{ls}}\|_{d,n}^2.\end{aligned} \tag{117}$$

Finally, the dual of (112), followed by (74), gives

$$\begin{aligned}-\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathcal{L}}{d}&\geq\frac{c}{1+(d/n)^{-1/2}+v_{31}^2+v_{33}^2}\left\{\sum_{i=1}^{2}\left|D_{\widetilde{\theta}_i}\frac{\mathcal{L}}{d}\right|_{\widetilde{\mathbf{\Theta}}_4}\right|^2+d\left\|D_{\widetilde{\mathbf{\Theta}}_4}\frac{\mathcal{L}}{d}\right\|_F^2\right\}\\&\geq\frac{c(d/n)}{T\big(1+(d/n)^{-1/2}+v_{31}^2+v_{33}^2\big)}\,\frac{\mathcal{L}-\mathcal{L}^*}{d}.\end{aligned}$$

This proves (115). □

Convergence part of Theorem 3.1. Lemmas G.8 and G.5 give the finite function separation, unique minimizer, and transfer budget (159); the finite Heavy-Ball transfer is proved in Lemma G.12 below. Begin Stage 2 from a Stage-1 iterate with $\mathcal{L}_{\mathrm{GD}}\leq d^{-1/4}$. Lemmas F.3 and E.2 show that the unchanged Stage-2 flow reaches the positive compact sublevel after an initial interval of length $d^{-1+o(1)}$; no reset is made. The reduced vector field is polynomial. On every finite interval $[0,\tau^\star)$ contained in its maximal existence interval,

$$\begin{aligned}\int_0^{\tau^\star}&\Big[(\dot v_{31})^2+(\dot w_{13})^2+(\dot v_{33})^2\\&+\frac{1}{d}\{\dot w_{24}^2+\|\dot{\mathbf{w}}_{25}\|^2+\|\dot w_{44}\|_F^2+\|\dot{\mathbf{w}}_{45}\|_F^2+\|\dot{\mathbf{w}}_{54}\|_F^2+\|\dot{\mathbf{W}}_{55}\|_F^2\}\Big]\,\mathrm{d}\tau\\&=\frac{\mathcal{L}(0)-\mathcal{L}(\tau^\star)}{d}\leq\frac{\mathcal{L}(0)}{d}.\end{aligned}$$

Hence

$$\begin{aligned}\int_0^{\tau^\star}&\Big[(\dot v_{31})^2+(\dot w_{13})^2+(\dot v_{33})^2\\&+\frac{1}{d}\{\dot w_{24}^2+\|\dot{\mathbf{w}}_{25}\|^2+\|\dot w_{44}\|_F^2+\|\dot{\mathbf{w}}_{45}\|_F^2+\|\dot{\mathbf{w}}_{54}\|_F^2+\|\dot{\mathbf{W}}_{55}\|_F^2\}\Big]^{1/2}\,\mathrm{d}\tau\\&\leq\sqrt{\frac{\tau^\star\mathcal{L}(0)}{d}}.\end{aligned}$$

If the maximal time were finite, the raw trajectory would be Cauchy and would converge to a finite parameter value, from which the smooth ODE extends. Thus the flow exists globally.

After entry, loss monotonicity keeps the trajectory in the region $\mathcal{L}/d \le 3d/(4nT)$. Lemma F.5 bounds its auxiliary variables there; Lemma E.2 supplies the bounds on the initial interval. In particular Lemma F.7 applies from the original Stage-2 time zero.

(i) The initial loss bound in Lemma E.2 and the fixed aspect-ratio regime give

$$\frac{\mathcal{L}(0)}{d} \le \frac{49}{100T} \le C\frac{d/n}{T}. \tag{118}$$

Consequently,

$$\frac{\mathcal{L}(0) - \mathcal{L}^*}{d} \le C\frac{d/n}{T}. \tag{119}$$

Lemma F.7 first gives the intrinsic-clock estimate

$$\frac{\mathcal{L}(\tau) - \mathcal{L}^*}{d} \le \frac{\mathcal{L}(0) - \mathcal{L}^*}{d} \exp\left\{-c\frac{d/n}{T}\int_0^\tau \frac{\mathrm{d}r}{1 + (d/n)^{-1/2} + v_{31}(r)^2 + v_{33}(r)^2}\right\}. \tag{120}$$

Direct differentiation of $\sqrt{1 + v_{31}^2 + v_{33}^2}$, followed by Cauchy-Schwarz and the energy identity, gives

$$\left|\frac{\mathrm{d}}{\mathrm{d}\tau}\sqrt{1 + v_{31}(\tau)^2 + v_{33}(\tau)^2}\right| \le \sqrt{-\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathcal{L}(\tau) - \mathcal{L}^*}{d}}.$$

Since the integral of the right-hand square is bounded by the initial normalized excess,

$$\begin{aligned}
&(d/n)^{-1/2} + 1 + v_{31}(\tau)^2 + v_{33}(\tau)^2 \\
&\quad\le (d/n)^{-1/2} + 2\{1 + v_{31}(0)^2 + v_{33}(0)^2\} + 2\frac{\mathcal{L}(0) - \mathcal{L}^*}{d}\tau \\
&\quad\le d^{o(1)} + C\frac{d/n}{T}\tau.
\end{aligned} \tag{121}$$

At the direct Stage-1 endpoint, the product bounds in Lemma F.3 replace the exact zero-error identities. Lemma G.15 gives $v_{31}(0)^2 + v_{33}(0)^2 = d^{o(1)}$ at the Stage-2 time origin, while Lemma C.3 keeps $v_{33}(0) > 0$. Hence $\eta_0 \in [1/3, 9/10]$, the fixed scalar initialization, and (119) give the last line of (121). Then

$$-\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathcal{L}(\tau) - \mathcal{L}^*}{d} \ge \frac{c(d/n)/T}{d^{o(1)} + C(d/n)\tau/T}\frac{\mathcal{L}(\tau) - \mathcal{L}^*}{d}.$$

Direct integration gives, after changing the absolute exponent,

$$\frac{\mathcal{L}(\tau) - \mathcal{L}^*}{d} \le C\frac{d/n}{T}\left(1 + \frac{(d/n)\tau}{T\,d^{o(1)}}\right)^{-\alpha}. \tag{122}$$

(ii) Let $\widehat{R}_{T,t}$ denote the finite-Wishart model-optimal residual at horizon T. For the fixed $(\widetilde{\theta}_1, (-\widetilde{\theta}_2))$ constrained least-squares residuals used in the proof of Lemma F.7, orthogonality gives

$$\begin{aligned}
\sum_{t=1}^{T}\|R_t - \widehat{R}_{T,t}\|_{d,n}^2 = \sum_{t=1}^{T}\|R_t - R_{t,\mathrm{ls}}\|_{d,n}^2 \\
+ \sum_{t=1}^{T}\|R_{t,\mathrm{ls}} - \widehat{R}_{T,t}\|_{d,n}^2,
\end{aligned}$$

and the first sum is exactly

$$2T\frac{\mathcal{L} - \min_{\widetilde{\Theta}_4}\mathcal{L}}{d}.$$

The affine fixed-leading-coefficient projection in (166) gives exactly

$$\sum_{t=1}^{T}\|R_{t,\mathrm{ls}} - \widehat{R}_{T,t}\|_{d,n}^2 = \sum_{k=0}^{T-1} c_{k,\nu_{d,n}}\left\{\widetilde{\theta}_1(-\widetilde{\theta}_2)^k - \widetilde{\theta}_1^\star(-\widetilde{\theta}_2^\star)^k\right\}^2.$$

Lemmas E.2 and F.5 place the current parameters and the minimizer in

$$\frac{1}{100} \le \widetilde{\theta}_1 \le 2, \qquad \frac{1}{14} \le (-\widetilde{\theta}_2) \le \frac{38}{13}.$$

On this rectangle, $(d/n)(-\widetilde{\theta}_2)^2 \le 722/845 < 1$. The MP coefficients in (168) and the transfer estimate (169) give

$$C_\nu(-\widetilde{\theta}_2) \ge c, \qquad \sum_{j=0}^{2} \{|B_\nu^{(j)}(-\widetilde{\theta}_2)| + |C_\nu^{(j)}(-\widetilde{\theta}_2)|\} \le C,$$
$$\left|\left(\frac{B_\nu}{C_\nu}\right)'(-\widetilde{\theta}_2)\right| \le C.$$

For $Q = B^2/C$ and x_∞ from Lemma G.2, the same estimates yield

$$\sum_{j=0}^{2} \sup_{1/14 \le -\widetilde{\theta}_2 \le 38/13} \left|\left(\frac{B_\nu^2}{C_\nu} - Q\right)^{(j)}(-\widetilde{\theta}_2)\right|$$
$$\le CT^2 \left(\frac{722}{845}\right)^T + \exp\{CT\log(e+T)\}\left(\frac{d}{n}\right)^{-6T}\left(\frac{1}{d}+\frac{1}{n}\right) = o(1).$$

The ratio d/n ranges over a fixed compact subset of $(0, 1/10]$. Lemma G.2 and the implicit-function theorem make x_∞ continuous in d/n. Since

$$\lim_{-\widetilde{\theta}_2 \to x_\infty} \frac{Q(x_\infty) - Q(-\widetilde{\theta}_2)}{|\widetilde{\theta}_2 + x_\infty|^2} = -\frac{1}{2}Q''(x_\infty) > 0,$$

uniqueness and compactness give, uniformly in d/n,

$$-Q''(x_\infty) \ge c, \qquad \inf_{\substack{1/14 \le -\widetilde{\theta}_2 \le 38/13 \\ -\widetilde{\theta}_2 \ne x_\infty}} \frac{Q(x_\infty) - Q(-\widetilde{\theta}_2)}{|\widetilde{\theta}_2 + x_\infty|^2} \ge c.$$

Lemma G.8 then implies

$$c|\widetilde{\theta}_2^\star + x_\infty|^2 \le Q(x_\infty) - Q(-\widetilde{\theta}_2^\star)$$
$$\le 2 \sup_{1/14 \le -\widetilde{\theta}_2 \le 38/13} \left|\left(\frac{B_\nu^2}{C_\nu} - Q\right)(-\widetilde{\theta}_2)\right| = o(1).$$

Consequently, for some fixed $c, C > 0$ and all sufficiently large d,

$$-\left(\frac{B_\nu^2}{C_\nu}\right)''(-\widetilde{\theta}_2) \ge C^{-1}, \quad |\widetilde{\theta}_2 - \widetilde{\theta}_2^\star| \le c,$$
$$\frac{B_\nu(-\widetilde{\theta}_2^\star)^2}{C_\nu(-\widetilde{\theta}_2^\star)} - \frac{B_\nu(-\widetilde{\theta}_2)^2}{C_\nu(-\widetilde{\theta}_2)} \ge C^{-1}, \quad |\widetilde{\theta}_2 - \widetilde{\theta}_2^\star| \ge c.$$

For $|\widetilde{\theta}_2 - \widetilde{\theta}_2^\star| \le c$, Taylor's formula gives

$$\frac{B_\nu(-\widetilde{\theta}_2^\star)^2}{C_\nu(-\widetilde{\theta}_2^\star)} - \frac{B_\nu(-\widetilde{\theta}_2)^2}{C_\nu(-\widetilde{\theta}_2)}$$
$$= -|\widetilde{\theta}_2 - \widetilde{\theta}_2^\star|^2 \int_0^1 (1-s)\left(\frac{B_\nu^2}{C_\nu}\right)''\left(-\widetilde{\theta}_2^\star + s(\widetilde{\theta}_2^\star - \widetilde{\theta}_2)\right) \mathrm{d}s$$
$$\ge \frac{1}{2C}|\widetilde{\theta}_2 - \widetilde{\theta}_2^\star|^2.$$

For $|\widetilde{\theta}_2 - \widetilde{\theta}_2^\star| \ge c$, the second bound and $|\widetilde{\theta}_2 - \widetilde{\theta}_2^\star| < 3$ give the same inequality with $1/(9C)$. Thus

$$\frac{B_\nu(-\widetilde{\theta}_2^\star)^2}{C_\nu(-\widetilde{\theta}_2^\star)} - \frac{B_\nu(-\widetilde{\theta}_2)^2}{C_\nu(-\widetilde{\theta}_2)} \ge c|\widetilde{\theta}_2 - \widetilde{\theta}_2^\star|^2.$$

Using $\widetilde{\theta}_1^\star = B_\nu(-\widetilde{\theta}_2^\star)/C_\nu(-\widetilde{\theta}_2^\star)$ and (60), we obtain

$$\begin{aligned}
&|\widetilde{\theta}_1 - \widetilde{\theta}_1^\star|^2 + |\widetilde{\theta}_2 - \widetilde{\theta}_2^\star|^2 \\
&\quad \le C\left\{\left|\widetilde{\theta}_1 - \frac{B_\nu(-\widetilde{\theta}_2)}{C_\nu(-\widetilde{\theta}_2)}\right|^2 + |\widetilde{\theta}_2 - \widetilde{\theta}_2^\star|^2\right\} \\
&\quad \le CT\frac{\min_{\widetilde{\mathbf{\Theta}}_4} \mathcal{L}(\widetilde{\theta}_1, (-\widetilde{\theta}_2)) - \mathcal{L}^*}{d} \le CT\frac{\mathcal{L} - \mathcal{L}^*}{d}.
\end{aligned}$$

The rectangle is convex and contains the segment joining the current parameters to the minimizer. Uniformly on this rectangle, the geometric MP coefficients and their first derivatives, together with the coefficient transfer, give

$$\sum_{k=0}^{T-1} c_{k,\nu}(-\widetilde{\theta}_2)^{2k} \le C, \qquad \sum_{k=1}^{T-1} k^2 c_{k,\nu}\widetilde{\theta}_1^2(-\widetilde{\theta}_2)^{2k-2} \le C.$$

Indeed, throughout the rectangle, $(d/n)(-\widetilde{\theta}_2)^2 \le 722/845 < 1$, so the geometric sums and their derivatives are uniformly bounded; the finite errors are $d^{-1+o(1)} = o(1)$. The mean-value theorem and the preceding product-distance estimate give

$$\sum_{t=1}^{T} \|R_{t,\mathrm{ls}} - \widehat{R}_{T,t}\|_{d,n}^2 \le CT\frac{\min_{\widetilde{\mathbf{\Theta}}_4} \mathcal{L} - \mathcal{L}^*}{d}.$$

Conversely, the normalized function has zero gradient at its minimizer and Hessian bounded by C/T on the same compact rectangle. Since $c_{0,\nu}, c_{1,\nu} \ge c$ and $\widetilde{\theta}_1 \ge c$, its first two leading coefficients give

$$\begin{aligned}
&\frac{\min_{\widetilde{\mathbf{\Theta}}_4} \mathcal{L} - \mathcal{L}^*}{d} \\
&\quad \le \frac{C}{T}\left\{|\widetilde{\theta}_1 - \widetilde{\theta}_1^\star|^2 + |(-\widetilde{\theta}_2) - (-\widetilde{\theta}_2^\star)|^2\right\} \\
&\quad \le \frac{C}{T}\sum_{t=1}^{T} \|R_{t,\mathrm{ls}} - \widehat{R}_{T,t}\|_{d,n}^2.
\end{aligned}$$

Combining both estimates with the orthogonal decomposition yields

$$\frac{c}{T}\sum_{t=1}^{T} \|R_t(\tau) - \widehat{R}_{T,t}\|_{d,n}^2 \le \frac{\mathcal{L}(\tau) - \mathcal{L}^*}{d} \le \frac{C}{T}\sum_{t=1}^{T} \|R_t(\tau) - \widehat{R}_{T,t}\|_{d,n}^2. \tag{123}$$

Here c, C may depend on the fixed comparison bounds in $n = \Theta(d)$, but not on d, n, T. Combining this inequality with (122) yields

$$\frac{\mathcal{L}(\tau) - \mathcal{L}^*}{d} + \sum_{t=1}^{T} \|R_t(\tau) - \widehat{R}_{T,t}\|_{d,n}^2 \le CT^C\left(\frac{d}{n}\right)^{-C}\frac{d/n}{T}\left(1 + \frac{(d/n)\tau}{T\,d^{o(1)}}\right)^{-\alpha}. \tag{124}$$

Equation (120) is the corresponding exponential statement in the gauge-rescaled clock. Solving (124) for the required Stage-2 accuracy gives the time bound in Theorem 3.1; no raw convergent subsequence is used. □

F.3 Proofs of Lemmas 4.9–4.10

Lemma G.9 identifies $p_t^\star(\mathbf{S})\mathbf{w}^\star$ with the jointly optimal time-varying Heavy-Ball trajectory. The model optimum instead fixes the leading coefficient through the two shared auxiliary parameters. Lemma G.12 compares these constrained polynomials with the degree-wise optima, and Corollary G.13 gives the required late-horizon error.

Completion of the proof of Theorem 3.1. For every polynomial f, conditioning on $\mathbf{S}$ gives $\mathbb{E}\|f(\mathbf{S})\mathbf{w}^\star\|^2 = d\|f\|_{d,n}^2$. The exact decomposition is

$$\begin{aligned}
\mathbf{w}_{t+1}(\tau) - \mathbf{w}_t^{\mathrm{HB}} = &\{R_{t+1}(\tau) - \widehat{R}_{T,t+1}\}(\mathbf{S})\mathbf{w}^\star \\
&+ \widehat{R}_{T,t+1}(\mathbf{S})\mathbf{w}^\star - (\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star).
\end{aligned} \tag{125}$$

Equations (124) and (174) give, uniformly for $T/2 \le t < T$,

$$\frac{d\|R_{t+1}(\tau) - \widehat{R}_{T,t+1}\|_{d,n}^2}{\mathbb{E}\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star\|^2} \le \frac{C}{(1-d/n)(d/n)^{2T+t}} \left(1 + \frac{(d/n)\tau}{T\, d^{o(1)}}\right)^{-\alpha}. \tag{126}$$

Here we used $T^{C-1}(d/n)^{1-C} \le C(d/n)^{-2T}$ for large d. Choose τ so that (126) is at most $\epsilon/4$ simultaneously for all these indices. It suffices to take

$$\tau = O\left(d^{o(1)} \frac{T}{d/n} \left[\frac{1}{(1-d/n)(d/n)^{3T}\epsilon}\right]^{1/\alpha}\right).$$

Corollary G.13 bounds the second term in (125), in the same squared relative norm, by $\exp[-2(\log d)^{1/4}] \le \epsilon/4$ for sufficiently large d. The triangle inequality therefore yields

$$\frac{\left(\mathbb{E}\|\mathbf{w}_{t+1}(\tau) - \mathbf{w}_t^{\mathrm{HB}}\|^2\right)^{1/2}}{\left(\mathbb{E}\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star\|^2\right)^{1/2}} \le \frac{\sqrt{\epsilon}}{2} + \frac{\sqrt{\epsilon}}{2} = \sqrt{\epsilon}.$$

Squaring proves the claimed comparison. Under (17), both $T/(d/n)$ and $((1-d/n)(d/n)^{3T})^{-1}$ are $d^{o(1)}$, so the chosen Stage-2 time remains $O(d^{o(1)}\epsilon^{-1/\alpha})$. □

G Technical estimates

Assume $0 < d/n \le 9/20$, and let $\mu_{d/n}$ be the Marchenko-Pastur law with support

$$\left[(1-\sqrt{d/n})^2, (1+\sqrt{d/n})^2\right].$$

For every rollout $\mathbf{w}_t = p_t(\mathbf{S})\mathbf{w}^\star$, write $R_t = p_t - 1$. Conditioning on $\mathbf{S}$ gives the exact finite-dimensional identity

$$\frac{\mathcal{L}}{d} = \frac{1}{2T}\sum_{t=1}^{T} \frac{1}{d}\mathbb{E}\,\mathrm{tr}\big(R_t(\mathbf{S})^2\big). \tag{127}$$

Lemma G.1. *For every integer $k \ge 1$ and $1 \le d \le n$,*

$$0 \le \frac{1}{d}\mathbb{E}\,\mathrm{tr}(\mathbf{S}^k) - \mathbb{E}_{\lambda\sim\mu_{d/n}}[\lambda^k] \le \frac{(2k-1)!!}{n} \le \frac{\exp\{Ck\log(e+k)\}}{n}. \tag{128}$$

The difference is zero for $k = 0$.

Proof. The finite real-Wishart moment formula (Bryc, 2008, Corollary 2.17) is a sum over $(2k-1)!!$ pairings. The noncrossing terms sum to the MP moment (Bishop et al., 2018, Corollary 3.3); every remaining term is nonnegative and at most $1/n$ when $d \le n$ (Redelmeier, 2011, Sections 3 and 5). Therefore

$$0 \le \frac{1}{d}\mathbb{E}\,\mathrm{tr}\,\mathbf{S}^k - \mathbb{E}_{\lambda\sim\mu_{d/n}}[\lambda^k] \le \frac{(2k-1)!!}{n} \le \frac{(2k)^k}{n}.$$

These finite identities impose no fixed-order or fixed-aspect-ratio limit. □

G.1 Function sign and rational certificate

The following lemma records the exact coefficient argument used in the function proof.

Lemma G.2. *For $0 < d/n \le 9/20$, the infinite function $B(x)^2/C(x)$ in (131) has exactly one critical point $x_\infty \in (0, (d/n)^{-1/2})$ on its convergence interval. It satisfies*

$$\frac{1+d/n}{1+d/n+(d/n)^2} < x_\infty < 1, \qquad \frac{\mathrm{d}^2}{\mathrm{d}x^2}\frac{B(x)^2}{C(x)}\bigg|_{x=x_\infty} < 0. \tag{129}$$

Moreover,

$$x_\infty \le 1 - \frac{(d/n)^2}{128}. \tag{130}$$

Proof. The proof first establishes strict unimodality of the infinite series

$$B(x)=\sum_{k=0}^{\infty}\frac{\left(\frac{d}{n}\right)^k}{\sum_{r=0}^{k}\left(\frac{d}{n}\right)^r}x^k,\qquad C(x)=\sum_{k=0}^{\infty}\frac{\left(\frac{d}{n}\right)^k\sum_{r=0}^{k+1}\left(\frac{d}{n}\right)^r}{\sum_{r=0}^{k}\left(\frac{d}{n}\right)^r}x^{2k},\qquad Q(x)=\frac{B(x)^2}{C(x)} \tag{131}$$

on $0<x<\left(\frac{d}{n}\right)^{-1/2}$, and then transfers the unique nondegenerate critical point to the finite sums. Pairing the (i,j) and (j,i) terms in $2B'C-BC'$ gives the absolutely convergent identity

$$\begin{aligned}&\frac{x}{2}\{2B'(x)C(x)-B(x)C'(x)\}\\&\quad=\sum_{0\le j<i}(i-j)\frac{\left(\frac{d}{n}\right)^i}{\sum_{r=0}^{i}\left(\frac{d}{n}\right)^r}\frac{\left(\frac{d}{n}\right)^j}{\sum_{r=0}^{j}\left(\frac{d}{n}\right)^r}x^{i+2j}\\&\quad\times\left\{\sum_{r=0}^{j+1}\left(\frac{d}{n}\right)^r-\sum_{r=0}^{i+1}\left(\frac{d}{n}\right)^r x^{i-j}\right\}.\end{aligned} \tag{132}$$

Since

$$\left(\sum_{r=0}^{m+1}\left(\frac{d}{n}\right)^r\right)^2-\left(\sum_{r=0}^{m}\left(\frac{d}{n}\right)^r\right)\left(\sum_{r=0}^{m+2}\left(\frac{d}{n}\right)^r\right)=\left(\frac{d}{n}\right)^{m+1}>0,$$

the ratios $\sum_{r=0}^{m}\left(\frac{d}{n}\right)^r/\sum_{r=0}^{m+1}\left(\frac{d}{n}\right)^r$ strictly increase. Hence every summand in (132) is nonnegative, with at least one strict, when

$$0<x\le\frac{1+\frac{d}{n}}{1+\frac{d}{n}+(\frac{d}{n})^2}.$$

Every summand is negative when $1\le x<\left(\frac{d}{n}\right)^{-1/2}$. Therefore

$$Q'(x)>0\quad\left(0<x\le\frac{1+\frac{d}{n}}{1+\frac{d}{n}+(\frac{d}{n})^2}\right),\qquad Q'(x)<0\quad(1\le x<\left(\frac{d}{n}\right)^{-1/2}). \tag{133}$$

It remains to prove that Q' crosses zero only once in

$$\left(\frac{1+d/n}{1+d/n+(d/n)^2},1\right).$$

Multiply numerator and denominator by $(1+x)^2$:

$$(1+x)^2B(x)^2=\sum_{j\ge0}A_jx^j,\qquad(1+x)^2C(x)=\sum_{j\ge0}D_jx^j,\qquad D_j>0. \tag{134}$$

All D_j are positive. If

$$\chi_j=\frac{1+\left(\frac{d}{n}\right)^j}{1-\left(\frac{d}{n}\right)^j},\qquad F_k=\frac{\sum_{j=1}^{k+1}\chi_j}{1-\left(\frac{d}{n}\right)^{k+2}},\qquad E_k=F_{k-2}+2\left(\frac{d}{n}\right)F_{k-1}+\left(\frac{d}{n}\right)^2F_k,$$

and

$$v_m=\frac{1-\left(\frac{d}{n}\right)^{m+2}}{1-\left(\frac{d}{n}\right)^{m+1}},$$

direct convolution gives, for $m\ge1$,

$$\frac{A_{2m}}{D_{2m}}=(1-\left(\frac{d}{n}\right))^2\left(\frac{d}{n}\right)^{m-1}\frac{E_{2m}}{\left(\frac{d}{n}\right)v_m+v_{m-1}}, \tag{135}$$

$$\frac{A_{2m+1}}{D_{2m+1}}=(1-\left(\frac{d}{n}\right))^2\left(\frac{d}{n}\right)^{m-1}\frac{E_{2m+1}}{2v_m}. \tag{136}$$

For completeness, the convolution used here is

$$[x^k]B(x)^2 = \frac{(1-\left(\frac{d}{n}\right))^2\left(\frac{d}{n}\right)^k}{1-\left(\frac{d}{n}\right)^{k+2}} \sum_{j=1}^{k+1} \frac{1+\left(\frac{d}{n}\right)^j}{1-\left(\frac{d}{n}\right)^j}. \tag{137}$$

The first differences are

$$\frac{A_1}{D_1} - \frac{A_0}{D_0} = \frac{\left(\frac{d}{n}\right)}{(1+\left(\frac{d}{n}\right))^2} > 0, \tag{138}$$

$$\frac{A_2}{D_2} - \frac{A_1}{D_1} = \frac{\left(\frac{d}{n}\right)\left(-2\left(\frac{d}{n}\right)^5 + \left(\frac{d}{n}\right)^4 + 9\left(\frac{d}{n}\right)^3 + 13\left(\frac{d}{n}\right)^2 + 8\left(\frac{d}{n}\right) + 2\right)}{(1+\left(\frac{d}{n}\right))^2(1+\left(\frac{d}{n}\right)+\left(\frac{d}{n}\right)^2)(\left(\frac{d}{n}\right)^3 + 2\left(\frac{d}{n}\right)^2 + 3\left(\frac{d}{n}\right) + 1)} > 0, \tag{139}$$

$$\frac{A_3}{D_3} - \frac{A_2}{D_2} = \frac{\left(\frac{d}{n}\right)^2\left(-2\left(\frac{d}{n}\right)^5 + 3\left(\frac{d}{n}\right)^3 + 5\left(\frac{d}{n}\right)^2 + 4\left(\frac{d}{n}\right) + 3\right)}{(1+\left(\frac{d}{n}\right)^2)(1+\left(\frac{d}{n}\right)+\left(\frac{d}{n}\right)^2)^2(\left(\frac{d}{n}\right)^3 + 2\left(\frac{d}{n}\right)^2 + 3\left(\frac{d}{n}\right) + 1)} > 0. \tag{140}$$

The last two numerators are positive because $2\left(\frac{d}{n}\right)^5 < 2\left(\frac{d}{n}\right)^3$.

Equations (135)-(136) show that

$$\frac{A_{2m+1}}{D_{2m+1}} > \frac{A_{2m+2}}{D_{2m+2}} \iff (\left(\frac{d}{n}\right) v_{m+1} + v_m)E_{2m+1} - 2\left(\frac{d}{n}\right) v_m E_{2m+2} > 0, \quad m \geq 1, \tag{141}$$

$$\frac{A_{2m}}{D_{2m}} > \frac{A_{2m+1}}{D_{2m+1}} \iff 2v_m E_{2m} - (\left(\frac{d}{n}\right) v_m + v_{m-1})E_{2m+1} > 0, \quad m \geq 2. \tag{142}$$

The direct rational certificate below proves both inequalities and hence shows

$$\frac{A_0}{D_0} < \frac{A_1}{D_1} < \frac{A_2}{D_2} < \frac{A_3}{D_3} > \frac{A_4}{D_4} > \frac{A_5}{D_5} > \cdots. \tag{143}$$

We next use a power-series form of Descartes' rule. If $f(x) = \sum_{j\geq 0} b_j x^j$ has positive radius of convergence and the nonzero coefficient sequence has at most r sign changes, then f has at most r positive zeros, counted with multiplicity. To prove this, induct on r. The case $r = 0$ is immediate. Let k be the last index of the first constant-sign coefficient block and put

$$g(x) = xf'(x) - kf(x) = \sum_{j\geq 0}(j-k)b_j x^j.$$

The first block changes sign, the coefficient at k vanishes, and the next block keeps its sign, so g has $r-1$ sign changes. If the positive zeros of f have total multiplicity Z, Rolle's theorem applied to $x^{-k}f(x)$ shows that g has at least $Z-1$ positive zeros counted with multiplicity. Induction gives $Z - 1 \leq r - 1$.

For any real y, the coefficients of

$$(1+x)^2B(x)^2 - y(1+x)^2C(x)$$

are $D_j(\frac{A_j}{D_j} - y)$. By (143), they have at most two sign changes. Hence every horizontal line intersects Q at at most two positive points, counted with multiplicity. At a critical point x, the function $Q(\cdot) - Q(x)$ has a zero of multiplicity at least two; the same Descartes bound makes that multiplicity exactly two. Thus every critical point is nondegenerate.

By (133), at least one critical point lies in $\left(\frac{1+d/n}{1+d/n+(d/n)^2}, 1\right)$. The coefficient definition (134) gives

$$Q(0) = \frac{A_0}{D_0} > 0, \qquad \lim_{x\uparrow(d/n)^{-1/2}} Q(x) = 0.$$

If there were more than one critical point, nondegeneracy and the endpoint derivative signs would give at least three. Let x_1 and x_3 be the first and last local maxima, and let x_2 be a point of minimum value on the interval between them. Then $Q(x_1) > Q(0) > 0$, $Q(x_3) > 0$, and $Q(x_2) < \min\{Q(x_1), Q(x_3)\}$. Hence one can choose y with

$$\max\{Q(x_2), 0\} < y < \min\{Q(x_1), Q(x_3)\}.$$

The intermediate-value theorem gives one intersection of $Q(x) = y$ between x_1 and x_2, one between x_2 and x_3, and one after x_3, because the right endpoint limit is zero. This contradicts the Descartes bound of two positive intersections. Therefore Q cannot have three or more critical points. Since every critical point is nondegenerate, the sign of Q' changes at each one; the positive sign on the left and negative sign on the right would force any number larger than one to be at least three. Hence Q has exactly one critical point, it is a strict nondegenerate maximum, and it lies in $\big((1+d/n)/(1+d/n+(d/n)^2), 1\big)$.

We now quantify its separation from 1 in the range used by Theorem 3.1. At $x = 1$, every summand on the right side of (132) is negative. Keeping only the term $(i, j) = (1, 0)$ therefore gives

$$\sum_{0\le j<i} (i-j)\frac{(\frac{d}{n})^i}{\sum_{r=0}^{i}(\frac{d}{n})^r}\frac{(\frac{d}{n})^j}{\sum_{r=0}^{j}(\frac{d}{n})^r}x^{i+2j}\left\{\sum_{r=0}^{j+1}(\frac{d}{n})^r - \sum_{r=0}^{i+1}(\frac{d}{n})^r x^{i-j}\right\}\Bigg|_{x=1}$$
$$\le \frac{\frac{d}{n}}{\sum_{r=0}^{1}(\frac{d}{n})^r}\left\{\sum_{r=0}^{1}(\frac{d}{n})^r - \sum_{r=0}^{2}(\frac{d}{n})^r\right\} = -\frac{(\frac{d}{n})^3}{1+\frac{d}{n}}. \tag{144}$$

For $0 \le x \le 1$, termwise differentiation is justified by absolute convergence. The bounds

$$\frac{(d/n)^k}{\sum_{r=0}^{k}(d/n)^r} \le (d/n)^k, \qquad \sum_{r=0}^{k}(d/n)^r \le (1-d/n)^{-1}$$

then show that the absolute value of that derivative is at most

$$\frac{3}{1-\left(\frac{d}{n}\right)}\sum_{0\le j<i}(i^2-j^2)\left(\frac{d}{n}\right)^{i+j}$$
$$= \frac{3\left(\frac{d}{n}\right)\left(1+2\left(\frac{d}{n}\right)+3\left(\frac{d}{n}\right)^2\right)}{(1-\left(\frac{d}{n}\right))^5(1+\left(\frac{d}{n}\right))^2}. \tag{145}$$

Differentiating the two geometric sums $\sum_{j=0}^{i-1}(d/n)^j$ and $\sum_{j=0}^{i-1} j^2(d/n)^j$ gives directly

$$\sum_{0\le j<i}(i^2-j^2)\left(\frac{d}{n}\right)^{i+j} = \frac{\frac{d}{n}\left(1+2\frac{d}{n}+3(\frac{d}{n})^2\right)}{(1-\frac{d}{n})^4(1+\frac{d}{n})^2}.$$

The first displayed expression is simply the finite formulas for $\sum_{j=0}^{i-1}\left(\frac{d}{n}\right)^j$ and $\sum_{j=0}^{i-1} j^2\left(\frac{d}{n}\right)^j$, followed by summation in i. The mean-value theorem, (144), and (145) now give

$$\frac{x}{2}\{2B'(x)C(x) - B(x)C'(x)\}\Big|_{x=1-(d/n)^2/128}$$
$$\le -\frac{(d/n)^3}{1+d/n}\left\{1 - \frac{3(1+2(d/n)+3(d/n)^2)}{128(1-d/n)^5(1+d/n)}\right\} < 0. \tag{146}$$

Indeed, the fraction subtracted in braces is increasing in $\left(\frac{d}{n}\right)$. At the endpoint,

$$\frac{3(1+2\left(\frac{d}{n}\right)+3\left(\frac{d}{n}\right)^2)}{128(1-\left(\frac{d}{n}\right))^5(1+\left(\frac{d}{n}\right))}\Bigg|_{\left(\frac{d}{n}\right)=9/20} = \frac{3761250}{4670479} < 1.$$

The left side of (132) has the sign of Q'. Since Q' changes sign only once, at its unique maximizer x_∞,

$$x_\infty < 1 - \frac{\left(\frac{d}{n}\right)^2}{128}. \tag{147}$$

The finite-horizon function is handled by the exact coefficient-ratio lemma below; no C^2 perturbation of the finite function is used here. (147) proves the last assertion. □

Lemma G.3. *The rational inequalities* (141)-(142) *hold for every* $0 < \left(\frac{d}{n}\right) \le 9/20$ *in their stated index ranges.*

Proof. For (141), set $r = \left(\frac{d}{n}\right)^m$ and $S = \sum_{j=1}^{2m-1} \chi_j$. Substituting the definitions of E_k makes the left side affine in S:

$$\Delta_m^{\mathrm{o}} = \alpha_{\mathrm{o}}(\left(\frac{d}{n}\right), r)S + \beta_{\mathrm{o}}(\left(\frac{d}{n}\right), r).$$

After $r = \left(\frac{d}{n}\right) s$, exact common-denominator reduction gives

$$\alpha_{\mathrm{o}}(\left(\frac{d}{n}\right), \left(\frac{d}{n}\right) s) = \frac{(1 - \left(\frac{d}{n}\right))\mathcal{A}_{\mathrm{o}}(\left(\frac{d}{n}\right), s)}{(1 - \left(\frac{d}{n}\right)^4 s^2)(1 - \left(\frac{d}{n}\right)^6 s^2)(1 - \left(\frac{d}{n}\right)^3 s^2)(1 - \left(\frac{d}{n}\right)^5 s^2)}, \tag{148}$$

$$\alpha_{\mathrm{o}}\chi_1 + \beta_{\mathrm{o}} = \frac{(1 - \left(\frac{d}{n}\right))\mathcal{L}_{\mathrm{o}}(\left(\frac{d}{n}\right), s)}{\left[\begin{array}{l}(1 - (\frac{d}{n})^2 s^2)(1 - (\frac{d}{n})^2 s)^2(1 + (\frac{d}{n})^2 s) \\ \quad \times (1 - (\frac{d}{n})^6 s^2)(1 - (\frac{d}{n})^3 s^2)(1 - (\frac{d}{n})^5 s^2)\end{array}\right]}. \tag{149}$$

For (142), set $r = \left(\frac{d}{n}\right)^m$ and use the same S. After $r = \left(\frac{d}{n}\right)^2 s$,

$$\alpha_{\mathrm{e}}(\left(\frac{d}{n}\right), \left(\frac{d}{n}\right)^2 s) = \frac{(1 - \left(\frac{d}{n}\right))\mathcal{A}_{\mathrm{e}}(\left(\frac{d}{n}\right), s)}{(1 - \left(\frac{d}{n}\right)^4 s^2)(1 - \left(\frac{d}{n}\right)^6 s^2)(1 - \left(\frac{d}{n}\right)^5 s^2)(1 - \left(\frac{d}{n}\right)^7 s^2)}, \tag{150}$$

$$\alpha_{\mathrm{e}}(\chi_1 + \chi_2) + \beta_{\mathrm{e}} = \frac{\mathcal{L}_{\mathrm{e}}(\left(\frac{d}{n}\right), s)}{\left[\begin{array}{l}(1 + \frac{d}{n})(1 - (\frac{d}{n})^2 s)^2(1 + (\frac{d}{n})^2 s) \\ \quad \times (1 - (\frac{d}{n})^3 s)^2(1 + (\frac{d}{n})^3 s)(1 - (\frac{d}{n})^5 s^2)(1 - (\frac{d}{n})^7 s^2)\end{array}\right]}. \tag{151}$$

Every denominator is positive for $0 < \left(\frac{d}{n}\right) \le 9/20$ and $0 < s \le 1$. The four numerators are

$$\begin{aligned}
\mathcal{A}_{\mathrm{o}} &= 4(\tfrac{d}{n})^{16}s^6 + (\tfrac{d}{n})^2 + 2(\tfrac{d}{n}) + 1 + (-3(\tfrac{d}{n})^{13} - 3(\tfrac{d}{n})^{11} + 2(\tfrac{d}{n})^{10})s^4 \\
&\quad + (2(\tfrac{d}{n})^{10} - 2(\tfrac{d}{n})^9 - 2(\tfrac{d}{n})^8 + 2(\tfrac{d}{n})^7)s^3 + ((\tfrac{d}{n})^9 + 3(\tfrac{d}{n})^8 - 3(\tfrac{d}{n})^7 - 3(\tfrac{d}{n})^6 - 2(\tfrac{d}{n})^5)s^2 \\
&\quad + ((\tfrac{d}{n})^6 - 2(\tfrac{d}{n})^4 + (\tfrac{d}{n})^2)s, \\
\mathcal{L}_{\mathrm{o}} &= -12(\tfrac{d}{n})^{20}s^9 + (12(\tfrac{d}{n})^{18} + 8(\tfrac{d}{n})^{17})s^8 \\
&\quad + (4(\tfrac{d}{n})^{18} + 4(\tfrac{d}{n})^{17} + 2(\tfrac{d}{n})^{16} + 12(\tfrac{d}{n})^{15} + 2(\tfrac{d}{n})^{14})s^7 \\
&\quad + (-4(\tfrac{d}{n})^{16} - 12(\tfrac{d}{n})^{15} - 10(\tfrac{d}{n})^{14} - 10(\tfrac{d}{n})^{13} + 2(\tfrac{d}{n})^{12} - 6(\tfrac{d}{n})^{11})s^6 \\
&\quad + (2(\tfrac{d}{n})^{15} - 2(\tfrac{d}{n})^{14} - 6(\tfrac{d}{n})^{13} - 12(\tfrac{d}{n})^{12} - 8(\tfrac{d}{n})^{11} - 8(\tfrac{d}{n})^{10} + 2(\tfrac{d}{n})^9)s^5 \\
&\quad + (-2(\tfrac{d}{n})^{13} + 6(\tfrac{d}{n})^{12} + 8(\tfrac{d}{n})^{11} + 10(\tfrac{d}{n})^{10} + 8(\tfrac{d}{n})^9 + 2(\tfrac{d}{n})^8)s^4 \\
&\quad + (-2(\tfrac{d}{n})^{11} + 4(\tfrac{d}{n})^{10} + 10(\tfrac{d}{n})^9 + 10(\tfrac{d}{n})^8 + 12(\tfrac{d}{n})^7 + 6(\tfrac{d}{n})^6)s^3 \\
&\quad + (2(\tfrac{d}{n})^9 - 12(\tfrac{d}{n})^7 - 4(\tfrac{d}{n})^6 - 4(\tfrac{d}{n})^5 - 4(\tfrac{d}{n})^4 - 2(\tfrac{d}{n})^3)s^2 \\
&\quad + (2(\tfrac{d}{n})^6 - 6(\tfrac{d}{n})^5 - 12(\tfrac{d}{n})^4 - 4(\tfrac{d}{n})^3)s + 4(\tfrac{d}{n})^2 + 6(\tfrac{d}{n}) + 2, \\
\mathcal{A}_{\mathrm{e}} &= 4(\tfrac{d}{n})^{18}s^6 + (\tfrac{d}{n})^2 + 2(\tfrac{d}{n}) + 1 + (-2(\tfrac{d}{n})^{14} - 3(\tfrac{d}{n})^{13} + 4(\tfrac{d}{n})^{12} - 3(\tfrac{d}{n})^{11})s^4 \\
&\quad + (-2(\tfrac{d}{n})^{11} + 2(\tfrac{d}{n})^{10} + 2(\tfrac{d}{n})^9 - 2(\tfrac{d}{n})^8)s^3 + (-(\tfrac{d}{n})^9 - (\tfrac{d}{n})^8 + (\tfrac{d}{n})^7 + (\tfrac{d}{n})^6 - 4(\tfrac{d}{n})^5)s^2 \\
&\quad + (-(\tfrac{d}{n})^6 + 2(\tfrac{d}{n})^4 - (\tfrac{d}{n})^2)s, \\
\mathcal{L}_{\mathrm{e}} &= 8(\tfrac{d}{n})^{25}s^8 + (-8(\tfrac{d}{n})^{23} - 16(\tfrac{d}{n})^{22} + 8(\tfrac{d}{n})^{20})s^7 \\
&\quad + (3(\tfrac{d}{n})^{20} + 6(\tfrac{d}{n})^{19} - 6(\tfrac{d}{n})^{18} - 2(\tfrac{d}{n})^{17} - (\tfrac{d}{n})^{16})s^6 \\
&\quad + (5(\tfrac{d}{n})^{18} + 9(\tfrac{d}{n})^{17} + 8(\tfrac{d}{n})^{16} + 4(\tfrac{d}{n})^{15} - 5(\tfrac{d}{n})^{14} - 5(\tfrac{d}{n})^{13})s^5 \\
&\quad + (-(\tfrac{d}{n})^{16} - 2(\tfrac{d}{n})^{15} - 8(\tfrac{d}{n})^{13} - 3(\tfrac{d}{n})^{12} - 2(\tfrac{d}{n})^{11})s^4 \\
&\quad + ((\tfrac{d}{n})^{14} + (\tfrac{d}{n})^{13} - 3(\tfrac{d}{n})^{12} - 5(\tfrac{d}{n})^{11} + 3(\tfrac{d}{n})^{10} + 7(\tfrac{d}{n})^9 + 7(\tfrac{d}{n})^8 + 5(\tfrac{d}{n})^7)s^3 \\
&\quad + (-2(\tfrac{d}{n})^{10} - 2(\tfrac{d}{n})^9 + (\tfrac{d}{n})^8 - 2(\tfrac{d}{n})^7 + 4(\tfrac{d}{n})^6 + (\tfrac{d}{n})^4)s^2 \\
&\quad + ((\tfrac{d}{n})^8 - (\tfrac{d}{n})^7 + 4(\tfrac{d}{n})^5 - 5(\tfrac{d}{n})^4 \\
&\qquad - 11(\tfrac{d}{n})^3 - 4(\tfrac{d}{n})^2)s
\end{aligned}$$

$$+1+4(\tfrac{d}{n})+4(\tfrac{d}{n})^2-(\tfrac{d}{n})^4.$$

For $0<d/n\le 9/20$ and $0\le s\le 1$, direct grouping of the displayed polynomials by the sign of each monomial gives

$$\mathcal{A}_{\mathrm{o}}\left(\frac{d}{n},s\right)\ge\frac{5}{6},\qquad \mathcal{L}_{\mathrm{o}}\left(\frac{d}{n},s\right)\ge\frac{1}{2},\qquad \mathcal{A}_{\mathrm{e}}\left(\frac{d}{n},s\right)\ge\frac{2}{3},$$
$$\mathcal{L}_{\mathrm{e}}\left(\frac{d}{n},s\right)\ge 1+\frac{d}{n}\begin{bmatrix}4-11(\frac{d}{n})^2-6(\frac{d}{n})^3-3(\frac{d}{n})^6-2(\frac{d}{n})^8\\-2(\frac{d}{n})^9-7(\frac{d}{n})^{10}-6(\frac{d}{n})^{11}-13(\frac{d}{n})^{12}-5(\frac{d}{n})^{13}\\-2(\frac{d}{n})^{14}-2(\frac{d}{n})^{15}-6(\frac{d}{n})^{16}-16(\frac{d}{n})^{21}-8(\frac{d}{n})^{22}\end{bmatrix}\ge 1. \tag{152}$$

Indeed, the absolute sums of the negative monomials in the first three displayed polynomials are respectively less than $1/6$, $3/2$, and $1/3$ after replacing d/n and s by their endpoint bounds; their constant positive parts are at least 1, 2, and 1. For the last polynomial, use $s^k\le 1$ on the negative terms, together with

$$4(\tfrac{d}{n})^2(1-s)\ge 0,\qquad 2(\tfrac{d}{n})^{17}+6(\tfrac{d}{n})^{18}\le 6(\tfrac{d}{n})^{17}.$$

The polynomial inside the square brackets in (152) is decreasing on $[0,9/20]$. Its negative coefficients of degree at least six have absolute sum 72, so

$$\begin{aligned}\mathcal{L}_{\mathrm{e}}\left(\frac{d}{n},s\right)&\ge 1+\frac{d}{n}\left[4-11\left(\frac{9}{20}\right)^2-6\left(\frac{9}{20}\right)^3-72\left(\frac{9}{20}\right)^6\right]\\&>1+\frac{d}{2n}\ge 1.\end{aligned}$$

Thus all four numerators are strictly positive on the stated domain. Equations (148)-(151), together with $S\ge\chi_1$ for $m\ge 1$ and $S\ge\chi_1+\chi_2$ for $m\ge 2$, prove (141)-(142). □

For every horizon covered by Lemma D.5, let $(\widetilde{\theta}_1^{\mathrm{MP}},(-\widetilde{\theta}_2^{\mathrm{MP}}))$ be its unique positive point. Combining Lemmas G.7 and D.5, the MP-optimal residuals are

$$R_{T,t}=(1-\widetilde{\theta}_1^{\mathrm{MP}}(-\widetilde{\theta}_2^{\mathrm{MP}})^{t-1})R^*_{t-1}+(-1)^{t-1}\widetilde{\theta}_1^{\mathrm{MP}}(-\widetilde{\theta}_2^{\mathrm{MP}})^{t-1}\pi_t,\qquad 1\le t\le T. \tag{153}$$

Lemma G.4. *Let $\pi_0=1$, $\pi_1=\lambda-1$, and*

$$\pi_{t+1}=\{\lambda-(1+\left(\frac{d}{n}\right))\}\pi_t-\left(\frac{d}{n}\right)\pi_{t-1}. \tag{154}$$

Then $\{\pi_t\}$ is the monic orthogonal family for $\mu_{d/n}$, and

$$\|R_t^*\|^2=\frac{(\frac{d}{n})^t}{\sum_{r=0}^t(\frac{d}{n})^r},\qquad \pi_t(0)=(-1)^t,\qquad \|\pi_t\|^2=\left(\frac{d}{n}\right)^t. \tag{155}$$

Moreover,

$$R_t^*(\lambda)=-\frac{(\frac{d}{n})^t}{\sum_{r=0}^t(\frac{d}{n})^r}\sum_{j=0}^t(-1)^j\left(\frac{d}{n}\right)^{-j}\pi_j(\lambda). \tag{156}$$

Proof. Section 4 of Kusalik et al. (2005) with parameter n/d and its degree-t polynomial evaluated at $(n/d)\lambda$ and multiplied by $(d/n)^t$, gives (154), orthogonality, and squared norms $(d/n)^t$. The recurrence gives $\pi_t(0)=(-1)^t$. For every $R(0)=-1$ of degree at most t, Cauchy-Schwarz gives

$$1=|R(0)|^2\le\|R\|_{\mathrm{MP}}^2\sum_{j=0}^t\frac{\pi_j(0)^2}{\|\pi_j\|_{\mathrm{MP}}^2},$$

with equality only at (156). This proves the remaining identity in (155) and uniqueness. □

Lemma G.5. *Under (17), for all sufficiently large d and all polynomials f, g of degree at most T,*

$$\begin{aligned}&|\langle f,g\rangle_{d,n} - \langle f,g\rangle_{\mathrm{MP}}| \\ &\quad\le \exp\{CT\log(e+T)\}\left(\frac{d}{n}\right)^{-T}\left(\frac{1}{d}+\frac{1}{n}\right)\|f\|_{\mathrm{MP}}\|g\|_{\mathrm{MP}}.\end{aligned} \tag{157}$$

In particular,

$$\frac{1}{2}\|f\|_{\mathrm{MP}}^2 \le \|f\|_{d,n}^2 \le \frac{3}{2}\|f\|_{\mathrm{MP}}^2. \tag{158}$$

For every fixed absolute C, for all sufficiently large d,

$$\exp\{CT\log(e+T)\}\left(\frac{d}{n}\right)^{-6T}\left(\frac{1}{d}+\frac{1}{n}\right) = d^{-1+o(1)} \le \frac{(d/n)^{3/2}}{8192T}. \tag{159}$$

Proof. The recurrence (154) gives $\|\operatorname{coeff}_j(\pi_j)\|_1 \le 4^j$. By (155) and Cauchy-Schwarz,

$$\begin{aligned}\|f\|_{\mathrm{coeff}} &\le \sum_{j=0}^{T} 4^j\left(\frac{d}{n}\right)^{-j}|\langle f,\pi_j\rangle_{\mathrm{MP}}| \\ &\le \sqrt{T+1}\,4^T\left(\frac{d}{n}\right)^{-T/2}\|f\|_{\mathrm{MP}}.\end{aligned} \tag{160}$$

Expanding fg and applying (128) yields

$$\begin{aligned}&|\langle f,g\rangle_{d,n} - \langle f,g\rangle_{\mathrm{MP}}| \\ &\quad\le \frac{\exp\{CT\log(e+T)\}}{n}\left(\sum_{i=0}^{T}|[\lambda^i]f|\right)\left(\sum_{j=0}^{T}|[\lambda^j]g|\right) \\ &\quad\le \frac{(T+1)\exp\{CT\log(e+T)\}}{n}\|f\|_{\mathrm{coeff}}\|g\|_{\mathrm{coeff}}.\end{aligned}$$

This proves (157). The remaining claims follow from

$$T\log(e+T)+6T\log(n/d) = o(\log d), \qquad \log\frac{(d/n)^{3/2}}{8192T} = O(\log\log d). \qquad \square$$

Lemma G.6. *Assume (17) and write $\nu = \nu_{d,n}$, $\mu = \mu_{d/n}$. For $1 \le t \le T$, $\ell \in \mathbb{R}$, and $\sigma \in \{\nu,\mu\}$, let $R_{t,\ell}^{\sigma}$ be the unique minimizer of*

$$\min\left\{\|R\|_\sigma^2 : R(0) = -1,\ \deg R \le t,\ [\lambda^t]R = \ell\right\}.$$

For each fixed $A \ge 0$ and all sufficiently large d, uniformly for $|\ell| \le A$,

$$\begin{aligned}&\|R_{t,\ell}^{\nu} - R_{t,\ell}^{\mu}\|_{\mathrm{coeff}} + \left|\|R_{t,\ell}^{\nu}\|_{d,n}^2 - \|R_{t,\ell}^{\mu}\|_{\mathrm{MP}}^2\right| \\ &\quad\le C_A\exp\{CT\log(e+T)\}\left(\frac{d}{n}\right)^{-6T}\left(\frac{1}{d}+\frac{1}{n}\right).\end{aligned} \tag{161}$$

Write

$$\mathcal{Q}_{k,\sigma}(a) := \min_{\substack{R(0)=-1,\ \deg R\le k+1\\ [\lambda^{k+1}]R=(-1)^k a}}\|R\|_\sigma^2 = q_{k,\sigma} - 2b_{k,\sigma}a + c_{k,\sigma}a^2, \tag{162}$$

Then, uniformly for $0 \le k < T$,

$$|q_{k,\nu}-q_{k,\mu}| + |b_{k,\nu}-b_{k,\mu}| + |c_{k,\nu}-c_{k,\mu}| \le \exp\{CT\log(e+T)\}\left(\frac{d}{n}\right)^{-6T}\left(\frac{1}{d}+\frac{1}{n}\right). \tag{163}$$

The constants C and C_A are independent of $d, n, T, d/n$, with C_A depending only on A.

Proof. Equation (158) ensures uniqueness and, by minimality and the feasible polynomial $-1+\ell\lambda^t$,

$$\|R^{\nu}_{t,\ell}\|_{\mathrm{MP}} \le \sqrt{3}\|R^{\mu}_{t,\ell}\|_{\mathrm{MP}} \le \sqrt{3}\|-1+\ell\lambda^t\|_{\mathrm{MP}} \le \sqrt{3}(1+A3^T).$$

The difference has zero constant and leading coefficients. The two orthogonality conditions therefore give

$$\begin{aligned}\|R^{\nu}_{t,\ell}-R^{\mu}_{t,\ell}\|^2_{d,n} &= \langle R^{\mu}_{t,\ell}, R^{\nu}_{t,\ell}-R^{\mu}_{t,\ell}\rangle_{\mathrm{MP}} \\ &\quad - \langle R^{\mu}_{t,\ell}, R^{\nu}_{t,\ell}-R^{\mu}_{t,\ell}\rangle_{d,n}.\end{aligned}$$

By (157)-(158),

$$\|R^{\nu}_{t,\ell}-R^{\mu}_{t,\ell}\|_{\mathrm{MP}} \le C_A \exp\{CT\log(e+T)\}\left(\frac{d}{n}\right)^{-T}\left(\frac{1}{d}+\frac{1}{n}\right).$$

Equation (160) gives the coefficient bound in (161). For the values, minimality gives

$$\begin{aligned}\|R^{\nu}_{t,\ell}\|^2_{d,n} - \|R^{\nu}_{t,\ell}\|^2_{\mathrm{MP}} &\le \|R^{\nu}_{t,\ell}\|^2_{d,n} - \|R^{\mu}_{t,\ell}\|^2_{\mathrm{MP}} \\ &\le \|R^{\mu}_{t,\ell}\|^2_{d,n} - \|R^{\mu}_{t,\ell}\|^2_{\mathrm{MP}}.\end{aligned}$$

Apply (157) to both sides. Uniqueness and orthogonality also give

$$R^{\sigma}_{t,\ell} = R^{\sigma}_{t,0} + \ell\big(R^{\sigma}_{t,1} - R^{\sigma}_{t,0}\big),$$

so $\mathcal{Q}_{k,\sigma}$ is quadratic, with

$$\begin{aligned}q_{k,\sigma} &= \mathcal{Q}_{k,\sigma}(0), \qquad b_{k,\sigma} = \frac{\mathcal{Q}_{k,\sigma}(-1)-\mathcal{Q}_{k,\sigma}(1)}{4}, \\ c_{k,\sigma} &= \frac{\mathcal{Q}_{k,\sigma}(1)+\mathcal{Q}_{k,\sigma}(-1)-2\mathcal{Q}_{k,\sigma}(0)}{2},\end{aligned}$$

so the value bound with $A=1$ proves (163). □

Lemma G.7. *For every $\ell \in \mathbb{R}$,*

$$\operatorname*{arg\,min}_{\substack{\deg R\le t,\ R(0)=-1\\ [\lambda^t]R=\ell}} \|R\|^2_{\mathrm{MP}} = \{1+(-1)^t\ell\}R^*_{t-1} + \ell\pi_t, \tag{164}$$

$$\min_{\substack{\deg R\le t,\ R(0)=-1\\ [\lambda^t]R=\ell}} \|R\|^2_{\mathrm{MP}} = \frac{\left(\frac{d}{n}\right)^{t-1}}{\sum_{r=0}^{t-1}\left(\frac{d}{n}\right)^r}\{1+(-1)^t\ell\}^2 + \left(\frac{d}{n}\right)^t \ell^2. \tag{165}$$

Proof. Every admissible R is uniquely $q+\ell\pi_t$, where $\deg q \le t-1$ and $q(0) = -1-(-1)^t\ell$. Lemma G.4 gives

$$\|q\|^2 \ge \{1+(-1)^t\ell\}^2 \frac{\left(\frac{d}{n}\right)^{t-1}}{\sum_{r=0}^{t-1}\left(\frac{d}{n}\right)^r},$$

with equality only at $q = \{1+(-1)^t\ell\}R^*_{t-1}$. Since $\langle q, \pi_t\rangle = 0$,

$$\begin{aligned}\|R\|^2 &= \|q\|^2 + 2\ell\langle q,\pi_t\rangle + \ell^2\|\pi_t\|^2 \\ &\ge \{1+(-1)^t\ell\}^2 \frac{\left(\frac{d}{n}\right)^{t-1}}{\sum_{r=0}^{t-1}\left(\frac{d}{n}\right)^r} + \ell^2\left(\frac{d}{n}\right)^t.\end{aligned}$$

This proves (164)-(165), including uniqueness. □

Lemma G.8. *For d sufficiently large, assume the fixed-aspect-ratio regime (17). Put $\nu = \nu_{d,n}$ and $\mu = \mu_{d/n}$. For $0 \le k < T$, define*

$$\begin{aligned}\min_{\substack{R(0)=-1,\ \deg R\le k+1\\ [\lambda^{k+1}]R=(-1)^k a}} \|R\|^2_{d,n} &= q_{k,\nu} - 2b_{k,\nu}a + c_{k,\nu}a^2, \\ \min_{\substack{R(0)=-1,\ \deg R\le k+1\\ [\lambda^{k+1}]R=(-1)^k a}} \|R\|^2_{\mathrm{MP}} &= q_{k,\mu} - 2b_{k,\mu}a + c_{k,\mu}a^2.\end{aligned} \tag{166}$$

Then $b_{k,\nu}, c_{k,\nu} > 0$, and

$$B_\nu(x) = \sum_{k=0}^{T-1} b_{k,\nu} x^k, \quad C_\nu(x) = \sum_{k=0}^{T-1} c_{k,\nu} x^{2k},$$
$$B_\mu(x) = \sum_{k=0}^{T-1} b_{k,\mu} x^k, \quad C_\mu(x) = \sum_{k=0}^{T-1} c_{k,\mu} x^{2k}. \tag{167}$$

The function $B_\nu(x)^2/C_\nu(x)$ has exactly one critical point on $x > 0$; it is a strict nondegenerate maximum. Moreover,

$$(-\widetilde{\theta}_2^\star) \le 1 - \frac{(d/n)^2}{128}.$$

The same assertions hold for $B_\mu(x)^2/C_\mu(x)$, with critical point $(-\widetilde{\theta}_2)_\mu$.

Proof. **(i)** The norm equivalence (158) makes the constrained minima unique. The coefficient $c_{k,\sigma}$ is the value of the strictly positive quadratic form obtained by varying the leading coefficient while re-minimizing the free coordinates, so $c_{k,\nu}, c_{k,\mu} > 0$.

For the MP law, (165) gives, with $S_k = \sum_{r=0}^{k} (d/n)^r$,

$$b_{k,\mu} = \frac{(d/n)^k}{S_k}, \qquad c_{k,\mu} = \frac{(d/n)^k S_{k+1}}{S_k}. \tag{168}$$

Lemma G.6, applied to (166), gives

$$|b_{k,\nu} - b_{k,\mu}| + |c_{k,\nu} - c_{k,\mu}| \le \exp\{CT\log(e+T)\} \left(\frac{d}{n}\right)^{-6T} \left(\frac{1}{d} + \frac{1}{n}\right), \qquad 0 \le k < T. \tag{169}$$

Using

$$\log\left[\frac{\exp\{CT\log(e+T)\}(d/n)^{-7T}}{d}\right] = -\log d + CT\log(e+T) + 7T\log\frac{n}{d} + O(1)$$
$$= -\log d + o(\log d),$$

the transfer in (169) is at most $\frac{1}{4}(1 - d/n)(d/n)^T$. Thus $b_{k,\nu}, c_{k,\nu} > 0$.

It remains to prove the finite function sign. Write, only in this proof, $b_k = b_{k,\mu}$, $c_k = c_{k,\mu}$, and put $q_j = 0$ outside $0 \le j \le 2T-2$. Then

$$[x^j]B_\mu(x)^2 = q_j, \qquad [x^j](1+x)^2 B_\mu(x)^2 = q_j + 2q_{j-1} + q_{j-2}.$$

Set

$$A_{j,T} = q_j + 2q_{j-1} + q_{j-2}, \qquad D_{2m,T} = c_m + c_{m-1}, \quad D_{2m+1,T} = 2c_m,$$

where $c_{-1} = c_T = 0$, and $r_{j,T} = A_{j,T}/D_{j,T}$. For $j < T$, the coefficients are identical to those of the infinite series in (134); hence the exact certificate (143) gives

$$r_{0,T} < r_{1,T} < r_{2,T} < r_{3,T} > r_{4,T} > \cdots > r_{T-1,T}.$$

For $b_k = (d/n)^k/S_k$,

$$\frac{b_{k+1}}{b_k} = \frac{d}{n}\frac{S_k}{S_{k+1}} < \frac{d}{n}.$$

For $T-1 \le j \le 2T-2$, the finite convolution gives

$$q_{j+1} = \sum_{i=j-T+2}^{T-1} b_i b_{j+1-i} \le \frac{d}{n} \sum_{i=j-T+2}^{T-1} b_i b_{j-i} \le \frac{d}{n}\, q_j.$$

Consequently $q_{j+1} \le (d/n) q_j$ for $T-1 \le j \le 2T-2$, while the infinite convolution identity gives

$$\frac{q_j}{q_{j-1}} \le \frac{d}{n}\frac{j+1}{j}, \qquad 1 \le j < T.$$

At the first two truncation transitions, $T \geq 8$ and $0 < d/n \leq 1/10$ from (17) give

$$\frac{A_{T,T}}{A_{T-1,T}} \leq \frac{1 + \frac{8}{7}(2 + d/n)(d/n)}{2 + \frac{6}{7(d/n)}}$$
$$= \frac{(d/n)\{7 + 16(d/n) + 8(d/n)^2\}}{6 + 14(d/n)} \leq \frac{3}{2}\frac{d}{n}, \tag{170}$$
$$\frac{A_{T+1,T}}{A_{T,T}} \leq \frac{1 + (2 + d/n)(d/n)}{2 + \frac{7}{8(d/n)}}$$
$$= \frac{8(d/n)(1 + d/n)^2}{7 + 16(d/n)} \leq \frac{3}{2}\frac{d}{n}. \tag{171}$$

For the last inequalities, substitute $d/n \leq 1/10$ in each numerator after factoring out d/n, and drop the positive linear term in its denominator. To treat both parities of T, observe that for $4 \leq m < T$,

$$\frac{c_m}{c_{m-1}} = \frac{d}{n}\left(1 - \frac{(d/n)^m}{S_m^2}\right) \geq \frac{d}{n}\{1 - (d/n)^4\}.$$

Each of the two denominator ratios is either $(1 + c_m/c_{m-1})/2$ or $2(c_m/c_{m-1})/(1 + c_m/c_{m-1})$, with $4 \leq m < T$. Thus

$$\min\left\{\frac{D_{T,T}}{D_{T-1,T}}, \frac{D_{T+1,T}}{D_{T,T}}\right\} \geq \min\left\{\frac{1}{2}, \frac{2(d/n)\{1 - (d/n)^4\}}{1 + (d/n)\{1 - (d/n)^4\}}\right\} \geq \frac{9}{5}\frac{d}{n}.$$

Here $1/2 \geq (9/5)(d/n)$, and the second entry of the minimum is at least $2(d/n)(1 - 10^{-4})/(11/10) > (9/5)(d/n)$. Both truncation transitions therefore strictly decrease $r_{j,T}$, regardless of the parity of T.

For $T + 1 \leq j \leq 2T - 2$, put

$$a_j = \frac{q_j}{q_{j-1}}, \qquad b_j = \frac{q_{j-1}}{q_{j-2}}.$$

The convolution estimate above gives $0 < a_j, b_j \leq d/n$, and

$$\frac{A_{j+1,T}}{A_{j,T}} \leq \frac{(2 + d/n)a_j + 1}{a_j + 2 + b_j^{-1}} \leq \frac{(2 + d/n)(d/n) + 1}{(d/n) + 2 + (d/n)^{-1}} = \frac{d}{n}.$$

Here the second inequality follows because the displayed fraction is increasing in each of a_j and b_j on the indicated interval. For the denominator coefficients, if $j = 2m + 1$, then

$$\frac{D_{j+1,T}}{D_{j,T}} = \frac{1 + c_{m+1}/c_m}{2} > \frac{1}{2} > \frac{d}{n},$$

whereas if $j = 2m$, then

$$\frac{D_{j+1,T}}{D_{j,T}} = \frac{2c_m}{c_m + c_{m-1}} > \frac{d}{n},$$

because $m \geq 4$,

$$\frac{c_m}{c_{m-1}} = \frac{d}{n}\left(1 - \frac{(d/n)^m}{S_m^2}\right) \geq \frac{d}{n}\{1 - (d/n)^4\},$$

and $(2 - d/n)(1 - (d/n)^4) > 1$ on $0 < d/n \leq 9/20$. Therefore

$$r_{0,T} < r_{1,T} < r_{2,T} < r_{3,T} > r_{4,T} > \cdots > r_{2T,T}.$$

The exact certificate also gives an absolute C such that

$$\min_{0 \leq j < 2T} |r_{j+1,T} - r_{j,T}| \geq C^{-1}(d/n)^T. \tag{172}$$

For $j < T$, the direct rational bounds in Lemma G.3 give, for $0 \leq s \leq 1$,

$$\min\left\{\mathcal{A}_{\mathrm{o}}\left(\frac{d}{n}, s\right), \mathcal{L}_{\mathrm{o}}\left(\frac{d}{n}, s\right), \mathcal{A}_{\mathrm{e}}\left(\frac{d}{n}, s\right), \mathcal{L}_{\mathrm{e}}\left(\frac{d}{n}, s\right)\right\} \geq \frac{1}{2}.$$

The remaining denominators are bounded above by absolute constants; the explicit factors in (135)-(142) and (140) are bounded below by a fixed multiple of $(d/n)^T$. For the two truncation transitions, the bounds in (170)-(171) and the common denominator-ratio lower bound give

$$\max\left\{\frac{r_{T,T}}{r_{T-1,T}}, \frac{r_{T+1,T}}{r_{T,T}}\right\} \le \frac{(3/2)(d/n)}{(9/5)(d/n)} = \frac{5}{6} < \frac{9}{10}.$$

For $T+1 \le j \le 2T-2$, the estimates above give

$$\frac{r_{j+1,T}}{r_{j,T}} = \frac{A_{j+1,T}/A_{j,T}}{D_{j+1,T}/D_{j,T}} \le \frac{9}{10}.$$

For odd j, this is $2(d/n) \le 9/10$. For even j, it follows from

$$\frac{1+(d/n)\{1-(d/n)^4\}}{2\{1-(d/n)^4\}} \le \frac{9}{10},$$

because $4/5 - d/n - (9/5)(d/n)^4 + (d/n)^5$ is decreasing and positive at $d/n = 9/20$. For the last pair,

$$\frac{A_{2T,T}}{A_{2T-1,T}} = \frac{b_{T-1}}{2(b_{T-1}+b_{T-2})} \le \frac{d/n}{2}, \qquad \frac{D_{2T,T}}{D_{2T-1,T}} = \frac{1}{2},$$

so $r_{2T,T}/r_{2T-1,T} \le d/n < 9/10$ as well. Thus every tail gap is at least

$$\frac{1}{10} r_{j,T} \ge \frac{1}{10} r_{2T,T} = \frac{1}{10}\frac{(d/n)^{T-1}}{S_{T-1}S_T} \ge \frac{1}{10}\left(1-\frac{d}{n}\right)^2 (d/n)^T,$$

which proves (172) for all remaining indices.

The coefficient transfer in (169) changes every convolution coefficient by at most

$$CT\exp\{CT\log(e+T)\}(d/n)^{-6T}\left(\frac{1}{d}+\frac{1}{n}\right),$$

and changes every $D_{j,T}$ by the same expression without T. More explicitly, with

$$\varepsilon_T := \exp\{CT\log(e+T)\}(d/n)^{-6T}\left(\frac{1}{d}+\frac{1}{n}\right), \qquad |A_{j,T}| \le C,$$

$$\widehat{A}_{j,T} = A_{j,T} + \Delta A_{j,T}, \qquad |\Delta A_{j,T}| \le CT\varepsilon_T, \qquad \widehat{D}_{j,T} = D_{j,T} + \Delta D_{j,T}, \qquad |\Delta D_{j,T}| \le C\varepsilon_T,$$

we have $\varepsilon_T (d/n)^{-T} = o(1)$, hence $|\Delta D_{j,T}| \le D_{j,T}/2$, and therefore

$$|\widehat{r}_{j,T} - r_{j,T}| \le \frac{2|\Delta A_{j,T}|}{D_{j,T}} + \frac{2|A_{j,T}||\Delta D_{j,T}|}{D_{j,T}^2}$$
$$\le CT\exp\{CT\log(e+T)\}(d/n)^{-8T}\left(\frac{1}{d}+\frac{1}{n}\right) = o((d/n)^T).$$

$$\log\left[\frac{CT\exp\{CT\log(e+T)\}(d/n)^{-8T}(1/d+1/n)}{(d/n)^T}\right]$$
$$\le -\log d + CT\log(e+T) + 9T\log\frac{n}{d} + \log T + O(1) = -\log d + o(\log d).$$

Since the preceding error is $o((d/n)^T)$, for all sufficiently large d,

$$\sup_{j\le 2T}|\widehat{r}_{j,T} - r_{j,T}| \le \tfrac{1}{4}C^{-1}(d/n)^T, \qquad \operatorname{sgn}(\widehat{r}_{j+1,T} - \widehat{r}_{j,T}) = \operatorname{sgn}(r_{j+1,T} - r_{j,T}),$$

and

$$\min_{j<2T}|\widehat{r}_{j+1,T} - \widehat{r}_{j,T}| \ge \tfrac{1}{2}C^{-1}(d/n)^T.$$

For the finite Wishart ratios

$$\widehat{r}_{j,T} = \frac{[x^j](1+x)^2 B_\nu(x)^2}{[x^j](1+x)^2 C_\nu(x)}. \tag{173}$$

For any $y \in \mathbb{R}$, the coefficients of

$$(1+x)^2\{B_\nu(x)^2 - yC_\nu(x)\}$$

are positive multiples of $\widehat{r}_{j,T} - y$, and hence have at most two sign changes. Descartes' rule gives at most two positive intersections of $B_\nu(x)^2/C_\nu(x)$ with any horizontal line, counted with multiplicity. If $x_0 > 0$ is critical for $Q_\nu(x) = B_\nu(x)^2/C_\nu(x)$, then

$$Q_\nu(x) - Q_\nu(x_0) = Q'_\nu(x_0)(x-x_0) + \tfrac{1}{2}Q''_\nu(x_0)(x-x_0)^2 + \cdots$$

has multiplicity at least two. If $Q''_\nu(x_0) = 0$, the multiplicity is at least three, contradicting the preceding Descartes bound. Hence every critical point is nondegenerate. Moreover, $\widehat{r}_{1,T} > \widehat{r}_{0,T}$ and $\widehat{r}_{2T-1,T} > \widehat{r}_{2T,T}$, so the function derivative is positive near zero and negative at infinity. By the definition of the coefficient ratios,

$$Q_\nu(0) = \widehat{r}_{0,T}, \qquad \lim_{x\to\infty} Q_\nu(x) = \widehat{r}_{2T,T}.$$

The first identity also shows that the derivative is positive immediately to the right of zero, while the second ratio gap shows that it is negative for all sufficiently large x. Moreover,

$$r_{0,T} = \frac{1}{1+d/n} \geq \frac{20}{29}, \qquad r_{2T,T} = \frac{(d/n)^{T-1}}{S_{T-1}S_T} \leq \left(\frac{9}{20}\right)^7 < \frac{1}{256}.$$

The ratio-transfer bound above is $o((d/n)^T)$, hence is smaller than $1/8$ for all sufficiently large d. Therefore

$$\widehat{r}_{2T,T} < \frac{1}{256} + \frac{1}{8} < \frac{20}{29} - \frac{1}{8} < \widehat{r}_{0,T}.$$

If there were more than one critical point, nondegeneracy and the endpoint derivative signs would give at least three. Let x_1 and x_3 be the first and last local maxima, and let x_2 be a point of minimum value on the interval between them. Then

$$Q_\nu(x_1) > Q_\nu(0), \qquad Q_\nu(x_3) > \lim_{x\to\infty} Q_\nu(x), \qquad Q_\nu(x_2) < \min\{Q_\nu(x_1), Q_\nu(x_3)\}.$$

Together with the strict endpoint inequality, this permits a number y such that

$$\max\Big\{Q_\nu(x_2), \lim_{x\to\infty} Q_\nu(x)\Big\} < y < \min\{Q_\nu(x_1), Q_\nu(x_3)\}.$$

The intermediate-value theorem gives one intersection of $Q_\nu(x) = y$ between x_1 and x_2, one between x_2 and x_3, and one after x_3. This contradicts the Descartes bound of two positive intersections. Thus there is exactly one positive critical point.

The coefficient estimate in (169), applied to the numerator $x\{2B'C - BC'\}/2$, together with (146), gives

$$\frac{x}{2}\{2B'_\nu(x)C_\nu(x) - B_\nu(x)C'_\nu(x)\}\Big|_{x=1-(d/n)^2/128} < 0$$

for all sufficiently large d in the displayed regime. For μ, the same conclusion follows directly from the finite tail $\sum_{k\geq T} k^2(d/n)^k$ and (146). Since the function is increasing near zero and has only one critical point, its critical point satisfies $(-\widetilde{\theta}_2^\star) \leq 1 - (d/n)^2/128$. □

G.2 Heavy-Ball estimates

Lemma G.9. *The jointly optimal trajectory defined after (7) exists and is unique. Its residual at every time t is the unique minimizer of*

$$\min_{\substack{R(0)=-1 \\ \deg R \leq t}} \|R\|_{d,n}^2.$$

Under (17), for all sufficiently large d, uniformly for $1 \leq t \leq T$,

$$\frac{1}{d}\mathbb{E}\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star\|^2 \geq \frac{1}{2}\|R_t^*\|_{\mathrm{MP}}^2 \geq \frac{1-d/n}{2}\left(\frac{d}{n}\right)^t, \tag{174}$$

and

$$\frac{\mathbb{E}\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star - R_t^*(\mathbf{S})\mathbf{w}^\star\|^2}{d\|R_t^*\|_{\mathrm{MP}}^2} \le \exp\{CT\log(e+T)\}\left(\frac{d}{n}\right)^{-2T}\left(\frac{1}{d}+\frac{1}{n}\right)^2. \tag{175}$$

Moreover, uniformly for $0 \le t < T$,

$$\begin{aligned}|\alpha_t^{\mathrm{HB}} - 1| + |\beta_t^{\mathrm{HB}} - d/n| &\le C\left(\frac{d}{n}\right)^{t+1} \\ &\quad + \exp\{CT\log(e+T)\}\left(\frac{d}{n}\right)^{-2T}\left(\frac{1}{d}+\frac{1}{n}\right).\end{aligned} \tag{176}$$

Thus both coefficient errors tend to zero uniformly for $T/2 \le t < T$ *in the joint scaling* (17).

Proof. Every trajectory in (7) has a residual R_t with $R_t(0) = -1$ and $\deg R_t \le t$. Use the same definition of $p_t^\star$ also at $t = T$. The finite moment Gram matrices are positive definite, so the minimizers exist uniquely, and their normal equations are

$$\langle p_t^\star - 1, \lambda p\rangle_{d,n} = 0 \quad (\deg p < t), \qquad p_0^\star = 0, \qquad p_t^\star(0) - 1 = -1. \tag{177}$$

For $t \ge 1$,

$$\deg(p_t^\star - 1) < t \implies 0 = \langle p_t^\star - 1, \lambda(p_t^\star - 1)\rangle_{d,n} = \int_0^\infty \lambda(p_t^\star - 1)^2 \,\mathrm{d}\nu_{d,n} > 0,$$

a contradiction. Hence

$$\deg(p_t^\star - 1) = t, \qquad [\lambda^t](p_t^\star - 1) \ne 0, \qquad \int_0^\infty \lambda(p_t^\star - 1)(p_j^\star - 1)\,\mathrm{d}\nu_{d,n} = 0 \quad (0 \le j < t).$$

Define the HB coefficients by

$$\begin{aligned}\alpha_0^{\mathrm{HB}} &= \frac{\langle 1, \lambda\rangle_{d,n}}{\langle \lambda, \lambda\rangle_{d,n}} = \frac{1}{1 + d/n + 1/n}, && \beta_0^{\mathrm{HB}} = 0, \\ \alpha_t^{\mathrm{HB}} &= -\frac{[\lambda^{t+1}](p_{t+1}^\star - 1)}{[\lambda^t](p_t^\star - 1)}, && 1 \le t < T, \\ \beta_t^{\mathrm{HB}} &= \frac{[\lambda^t](p_{t+1}^\star - 1) + \alpha_t^{\mathrm{HB}}[\lambda^{t-1}](p_t^\star - 1)}{[\lambda^t](p_t^\star - 1)} - 1, && 1 \le t < T.\end{aligned} \tag{178}$$

For $1 \le t < T$, the leading terms cancel, and

$$\begin{aligned}&\deg\{p_{t+1}^\star - 1 + \alpha_t^{\mathrm{HB}}\lambda(p_t^\star - 1)\} \le t, \\ &\int_0^\infty \lambda\{p_{t+1}^\star - 1 + \alpha_t^{\mathrm{HB}}\lambda(p_t^\star - 1)\}(p_j^\star - 1)\,\mathrm{d}\nu_{d,n} \\ &\qquad = \alpha_t^{\mathrm{HB}}\langle p_t^\star - 1, \lambda^2(p_j^\star - 1)\rangle_{d,n} = 0, \qquad 0 \le j \le t-2.\end{aligned}$$

Thus

$$p_{t+1}^\star - 1 + \alpha_t^{\mathrm{HB}}\lambda(p_t^\star - 1) \in \mathrm{span}\{p_t^\star - 1, p_{t-1}^\star - 1\}.$$

Evaluating at zero and comparing the degree-t coefficients gives

$$\begin{aligned}p_1^\star(\lambda) &= \alpha_0^{\mathrm{HB}}\lambda, \\ p_{t+1}^\star - 1 &= (1 + \beta_t^{\mathrm{HB}} - \alpha_t^{\mathrm{HB}}\lambda)(p_t^\star - 1) - \beta_t^{\mathrm{HB}}(p_{t-1}^\star - 1), \qquad 1 \le t < T.\end{aligned} \tag{179}$$

This is the recurrence in Pedregosa & Scieur (2020, Theorem 3.1). Substituting $\lambda = \mathbf{S}$ gives

$$\begin{aligned}p_1^\star(\mathbf{S})\mathbf{w}^\star &= \alpha_0^{\mathrm{HB}}\mathbf{X}\mathbf{y}^\mathsf{T}, \\ p_{t+1}^\star(\mathbf{S})\mathbf{w}^\star &= p_t^\star(\mathbf{S})\mathbf{w}^\star - \alpha_t^{\mathrm{HB}}\{\mathbf{S}p_t^\star(\mathbf{S})\mathbf{w}^\star - \mathbf{X}\mathbf{y}^\mathsf{T}\} \\ &\quad + \beta_t^{\mathrm{HB}}\{p_t^\star(\mathbf{S})\mathbf{w}^\star - p_{t-1}^\star(\mathbf{S})\mathbf{w}^\star\}.\end{aligned}$$

Therefore $p_t^\star(\mathbf{S})\mathbf{w}^\star$ is an admissible HB trajectory. For every admissible residual sequence, (177) gives the exact identity

$$\begin{aligned}\frac{d}{2T}\sum_{t=1}^{T}\left\{\|R_t\|_{d,n}^2-\|p_t^\star-1\|_{d,n}^2\right\}\\ =\frac{d}{2T}\sum_{t=1}^{T}\|R_t-(p_t^\star-1)\|_{d,n}^2\geq 0.\end{aligned}\tag{180}$$

Equality holds if and only if $R_t = p_t^\star - 1$ for every t. Thus the jointly optimal trajectory is unique, and (178) uniquely determines its coefficients.

The norm equivalence (158), followed by minimization, proves (174). For the finite degree-t minimizer R, both R and R_t^* are orthogonal to their difference in their respective norms. Hence

$$\|R-R_t^*\|_{d,n}^2=\langle R_t^*,R-R_t^*\rangle_{\mathrm{MP}}-\langle R_t^*,R-R_t^*\rangle_{d,n}.$$

Applying (157) and (158) gives

$$\|R-R_t^*\|_{\mathrm{MP}}\leq 2\exp\{CT\log(e+T)\}\left(\frac{d}{n}\right)^{-T}\left(\frac{1}{d}+\frac{1}{n}\right)\|R_t^*\|_{\mathrm{MP}}.$$

The identity $\mathbb{E}\|f(\mathbf{S})\mathbf{w}^\star\|^2 = d\|f\|_{d,n}^2$ proves (175). The coefficient bound (160) also gives

$$\|R-R_t^*\|_{\mathrm{coeff}}\leq\exp\{CT\log(e+T)\}\left(\frac{d}{n}\right)^{-2T}\left(\frac{1}{d}+\frac{1}{n}\right)=d^{-1+o(1)}.$$

For the MP family, comparison of the two highest-degree coefficients in (156) shows that the recurrence coefficients differ from $1, d/n$ by a total of at most $3(d/n)^{t+1}$, including the first step. The leading coefficient of R_t^* has magnitude at least $1 - d/n$; its other coefficients are bounded by $\exp(CT)$ by (156) and the recurrence for π_t. The same coefficient comparison, using the preceding perturbation bound, therefore proves (176). On $T/2 \leq t < T$, its first term vanishes as $T \to \infty$, and its second term is $d^{-1+o(1)}$. □

Lemma G.10. *In the fixed-aspect-ratio regime* (17)*, for all sufficiently large* d*,*

$$(-\widetilde{\theta}_2^{\mathrm{MP}})\leq 1-\frac{\left(\frac{d}{n}\right)^2}{128}.\tag{181}$$

Proof. This is the MP case of Lemma G.8. □

Lemma G.11. *For the optimal residuals* (153)*,*

$$\frac{\|R_{T,t}-R_{t-1}^*\|^2}{\|R_{t-1}^*\|^2}=\left(\widetilde{\theta}_1^{\mathrm{MP}}\right)^2(-\widetilde{\theta}_2^{\mathrm{MP}})^{2t-2}\sum_{r=0}^{t}\left(\frac{d}{n}\right)^r\leq\frac{\left(1-\left(\frac{d}{n}\right)^2/128\right)^{2t-2}}{1-\left(\frac{d}{n}\right)}.\tag{182}$$

Proof. Subtracting R_{t-1}^* in (153) and using $R_{t-1}^* \perp \pi_t$ gives the equality in (182). For $0 \leq (-\widetilde{\theta}_2) \leq 1$, completion of squares and telescoping give

$$\begin{aligned}C_T(-\widetilde{\theta}_2)-B_T(-\widetilde{\theta}_2)&\geq\frac{d}{n}-\frac{1}{4(d/n)}\left(\frac{1}{1+d/n}-\frac{1}{\sum_{r=0}^{T}(d/n)^r}\right)\\&\geq\frac{d}{n}\left(1-\frac{1}{4(1+d/n)}\right)>0.\end{aligned}$$

Thus $0 < \widetilde{\theta}_1^{\mathrm{MP}} < 1$. The upper bound in (182) now follows from (181) and $\sum_{r=0}^{t}(d/n)^r \leq (1 - d/n)^{-1}$. □

Lemma G.12. *Assume the fixed-aspect-ratio regime (17). Let $\widehat{R}_{T,t}$ be the finite-Wishart model-optimal residual, and let $(\widetilde{\theta}_1^\star, (-\widetilde{\theta}_2^\star))$ be the unique positive finite function minimizer. For $1 \le t \le T$, let*

$$\widetilde{R}_t = \operatorname*{arg\,min}_{\substack{R(0)=-1,\ \deg R \le t \\ [\lambda^t]R = (-1)^{t-1}\widetilde{\theta}_1^\star(-\widetilde{\theta}_2^\star)^{t-1}}} \|R\|_{\mathrm{MP}}^2.$$

There is an absolute C such that, for all sufficiently large d, uniformly for $T/2 \le t < T$,

$$\begin{aligned} &\frac{\mathbb{E}\|\widehat{R}_{T,t+1}(\mathbf{S})\mathbf{w}^\star - (\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star)\|^2}{\mathbb{E}\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star\|^2} \\ &\le \frac{C}{1-d/n}\left(1 - \frac{(d/n)^2}{128}\right)^{2t} \\ &\quad + \frac{\exp\{CT\log(e+T)\}}{(1-d/n)(d/n)^{12T+t}}\left(\frac{1}{d}+\frac{1}{n}\right)^2. \end{aligned} \tag{183}$$

Proof. Lemma G.8 and (169) give

$$0 < \widetilde{\theta}_1^\star \le \frac{2\sum_{k=0}^{T-1}(d/n)^k}{(1+d/n)/2} \le 8, \qquad 0 < (-\widetilde{\theta}_2^\star) \le 1 - \frac{(d/n)^2}{128}.$$

The fixed-leading projection (164) gives

$$\widetilde{R}_t = (1 - \widetilde{\theta}_1^\star(-\widetilde{\theta}_2^\star)^{t-1})R_{t-1}^* + (-1)^{t-1}\widetilde{\theta}_1^\star(-\widetilde{\theta}_2^\star)^{t-1}\pi_t. \tag{184}$$

Orthogonality and (155) imply

$$\frac{\|\widetilde{R}_s - R_{s-1}^*\|_{\mathrm{MP}}^2}{\|R_{s-1}^*\|_{\mathrm{MP}}^2} = (\widetilde{\theta}_1^\star)^2(-\widetilde{\theta}_2^\star)^{2s-2}\sum_{r=0}^{s}(d/n)^r \le \frac{64}{1-d/n}\left(1 - \frac{(d/n)^2}{128}\right)^{2s-2}.$$

The finite and MP projections have the same leading coefficient, bounded by 8. Lemma G.6 therefore gives, uniformly for $1 \le s \le T$,

$$\begin{aligned} &\|\widehat{R}_{T,s} - \widetilde{R}_s\|_{\mathrm{coeff}} + \left|\|\widehat{R}_{T,s}\|_{d,n}^2 - \|\widetilde{R}_s\|_{\mathrm{MP}}^2\right| \\ &\quad \le \exp\{CT\log(e+T)\}(d/n)^{-6T}\left(\frac{1}{d}+\frac{1}{n}\right) = d^{-1+o(1)} = o((d/n)^T). \end{aligned} \tag{185}$$

By (158) and the MP support,

$$\|f\|_{d,n} \le \sqrt{\frac{3}{2}(T+1)}(1+\sqrt{d/n})^{2T}\|f\|_{\mathrm{coeff}}, \qquad \deg f \le T.$$

Use the exact decomposition

$$\begin{aligned} \widehat{R}_{T,t+1}(\mathbf{S})\mathbf{w}^\star - (\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star) = &\{\widehat{R}_{T,t+1} - \widetilde{R}_{t+1}\}(\mathbf{S})\mathbf{w}^\star \\ &+ \{\widetilde{R}_{t+1} - R_t^*\}(\mathbf{S})\mathbf{w}^\star \\ &+ R_t^*(\mathbf{S})\mathbf{w}^\star - (\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star). \end{aligned}$$

The first term is controlled by (185) and the coefficient-to-norm bound above. The second is controlled by the tail identity with $s = t+1$, using (158). The third is controlled by (175). Divide by the denominator lower bound (174) and bound the squared norm of the sum by three times the sum of the three squared norms. This proves (183); every polynomial degree is at most $t+1 \le T$. □

Corollary G.13. *Under the assumptions of Lemma G.12, for all sufficiently large d, uniformly for $T/2 \le t < T$,*

$$\frac{\mathbb{E}\|\widehat{R}_{T,t+1}(\mathbf{S})\mathbf{w}^\star - (\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star)\|^2}{\mathbb{E}\|\mathbf{w}_t^{\mathrm{HB}} - \mathbf{w}^\star\|^2} \le \exp\left[-2(\log d)^{1/4}\right]. \tag{186}$$

Proof. By (18), for $t \ge T/2$,

$$\left(1 - \frac{(d/n)^2}{128}\right)^{2t} \le \exp\left[-\frac{(d/n)^2 T}{128}\right] \le \exp\left[-8(\log d)^{1/4}\right].$$

The second term in (183) is $d^{-2+o(1)}$, uniformly over these indices. Both terms are $o(\exp[-2(\log d)^{1/4}])$, proving (186). □

G.3 Teacher-polynomial estimates

All estimates below are uniform over the full interval $1/3 \le \eta_0 \le 9/10$. Since $n = \Theta(d)$,

$$\eta_0^T \ge 3^{-T} = e^{-O(T)} = d^{-o(1)}.$$

Moreover, on the support $[(1-\sqrt{d/n})^2, (1+\sqrt{d/n})^2]$ of the MP law,

$$|1-\eta_0\lambda| \le \max\left\{1-\frac{1}{3}(1-\sqrt{d/n})^2,\ \frac{9}{10}(1+\sqrt{d/n})^2-1\right\} < \frac{17}{20}, \tag{187}$$

since $d/n \le 1/10$.

For integers $a, b \ge 0$ with $a+b \le 2T$, expand $\mathbf{S}^b(\mathbf{I}-\eta_0\mathbf{S})^a$ and apply the finite-Wishart moment estimate term by term. Since $0 < \eta_0 \le 1$, the sum of the absolute binomial coefficients is at most 2^a, and hence

$$\left|\frac{1}{d}\mathbb{E}\,\mathrm{tr}[\mathbf{S}^b(\mathbf{I}-\eta_0\mathbf{S})^a] - \mathbb{E}_{\lambda\sim\mu_{d/n}}[\lambda^b(1-\eta_0\lambda)^a]\right| \le \exp\{CT\log(e+T)\}\left(\frac{1}{d}+\frac{1}{n}\right) = d^{-1+o(1)}. \tag{188}$$

The estimate is uniform in η_0, rather than a perturbation estimate around one distinguished step. The projection spaces are also independent of the particular allowed step, since $\lambda, 1-(1-\eta_0\lambda)^2, \dots, 1-(1-\eta_0\lambda)^t$ span all zero-constant polynomials of degree at most t. Their triangular change of basis has diagonal powers of η_0; the lower bound on η_0^T shows that these and their inverses cost only $d^{o(1)}$. Combining this observation with (188) and the MP Gram lower bound $\exp\{-CT\log(e+T)\}(d/n)^{CT}$, the relative finite-Wishart perturbation is $d^{-1+o(1)} = o(1)$. The Neumann series therefore gives all moment, Gram, projection, and triangular inverse estimates used below, uniformly over $\eta_0 \in [1/3, 9/10]$ and with the same stated scales.

G.4 Stage-1 entry estimates

Proof of Lemma C.4. The polynomials $1-(1-\eta_0\lambda)^\ell$, $1 \le \ell \le t$, have successive degrees and span the range of P_t. All terms of the teacher residual except its degree-$(t+1)$ term belong to that range. Pythagoras gives (33). The finite-Wishart second moment gives $M = 1 + d/n + 1/n$. The MP constrained Gram-Schmidt formula in (168) gives

$$\min_{\substack{q(0)=0\\ \deg q\le t}} \left\|\lambda\{1-(1-\eta_0\lambda)^t\} - q\right\|_{\mathrm{MP}}^2 = \eta_0^{2t}(d/n)^t \frac{\sum_{r=0}^{t+1}(d/n)^r}{\sum_{r=0}^{t}(d/n)^r}.$$

The $t=1$ term is bounded below by a fixed multiple of d/n, while the sum is at most $\sum_{t\ge1}(d/n)^t/(1-d/n) \le 2d/n$. The uniform moment and inverse-Gram transfer in (188) changes this sum by $d^{-1+o(1)} = o(d/n)$. This proves (34).

The same finite Gram and triangular inverse estimates give, simultaneously for all columns and all r in the range of P_t,

$$\sum_{\ell=1}^{t} |\langle r, 1-(1-\eta_0\lambda)^\ell\rangle_{d,n}|^2 \ge \exp\{-CT\log(e+T)\}(d/n)^{CT}\|r\|_{d,n}^2. \tag{189}$$

Enlarging the fixed C also gives the coefficient bound in (36). All of these estimates are deterministic and uniform in the trajectory.

The independent raw coordinate $(\mathbf{w}_{25})_0$ has gradient $v_{31}M\{v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)-\eta_0\}/T$. The independent raw coordinate $(\mathbf{W}_{55})_{\ell t}$ has gradient

$$D_{(\mathbf{W}_{55})_{\ell t}}\frac{\mathcal{L}_{\mathrm{GD}}}{d} = \frac{v_{33}}{T}\left\langle P_t A_t, 1-(1-\eta_0\lambda)^\ell\right\rangle_{d,n}.$$

Their contributions to (29) are disjoint nonnegative squares. Consequently

$$-\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathcal{L}_{\mathrm{GD}}}{d} \ge \frac{dv_{31}^2M^2}{T^2}\{v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)-\eta_0\}^2 + \frac{dv_{33}(0)^2}{T^2}\exp\{-CT\log(e+T)\}(d/n)^{CT}\mathsf{z}.$$

Combining this with (33) proves (35). No bound on the separate products $v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)$, $t \ge 1$, is used. □

Proof of Lemma C.5. At initialization the residual is $-\eta_0\lambda(1-\eta_0\lambda)^t$. The MP support estimate and (188) give

$$\frac{\mathcal{L}_{\mathrm{GD}}(0)}{d} = \frac{\eta_0^2}{2T}\sum_{t=0}^{T-1}\|\lambda(1-\eta_0\lambda)^t\|_{d,n}^2 \le \frac{C}{T}\sum_{t\ge 0}(17/20)^{2t} + d^{-1+o(1)} \le \frac{C}{T}.$$

The energy identity bounds the squared scalar velocities and d^{-1} times the squared ordinary-block velocities by an integrable function of total integral at most $\mathcal{L}_{\mathrm{GD}}(0)/d$. Cauchy-Schwarz on every finite interval bounds every raw displacement. A finite maximal existence time would therefore have a finite raw limit; local existence for the polynomial vector field extends that limit, a contradiction. The lower bound for v_{33} follows from (31).

The fixed projection identity gives

$$D_{w_{13}}\frac{\mathcal{L}_{\mathrm{GD}}}{d} = v_{31}\left[\frac{\mathsf{G}}{T}(v_{31}w_{13}+\eta_0) + \sum_{t=1}^{T-1}\sum_{\ell=1}^{t}\delta(\widetilde{\mathbf{W}}_{55})_{\ell t}D_{(\widetilde{\mathbf{W}}_{55})_{\ell t}}\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right],$$
$$\frac{\mathrm{d}}{\mathrm{d}\tau}(\mathbf{W}_{55})_{\ell t} = -dv_{33}D_{(\widetilde{\mathbf{W}}_{55})_{\ell t}}\frac{\mathcal{L}_{\mathrm{GD}}}{d}.$$

This proves (37). The shared quantities $w_{44}, \mathbf{w}_{45}, \mathbf{w}_{54}, \mathbf{W}_{55}$ remain trainable; the displayed equation for the independent $\mathbf{W}_{55}$ entries is unchanged by their presence.

Integrating by parts, the projected contribution over $[s,\tau]$ is exactly

$$\frac{1}{d}\left[\frac{v_{31}}{v_{33}}\sum_{t=1}^{T-1}\sum_{\ell=1}^{t}\delta(\widetilde{\mathbf{W}}_{55})_{\ell t}(\mathbf{W}_{55})_{\ell t}\right]_s^{\tau} - \frac{1}{d}\int_s^{\tau}\left(\frac{v_{31}}{v_{33}}\right)'\sum_{t=1}^{T-1}\sum_{\ell=1}^{t}\delta(\widetilde{\mathbf{W}}_{55})_{\ell t}(\mathbf{W}_{55})_{\ell t}\,\mathrm{d}r. \tag{190}$$

On $[0,H_d]$, the energy identity gives

$$\sup(|v_{31}|+|w_{13}|+|v_{33}-v_{33}(0)|) \le C\sqrt{H_d/T},$$
$$\sup\|\mathbf{W}_{55}\|_F \le C\sqrt{dH_d/T},$$
$$\int_0^{H_d}(|\dot v_{31}|+|\dot v_{33}|)\,\mathrm{d}r \le C\sqrt{H_d/T}.$$

Since $v_{33}\ge v_{33}(0)$, these inequalities imply

$$\int_0^{H_d}|(v_{31}/v_{33})'|\,\mathrm{d}r \le C\{\sqrt{H_d/T}+H_d/T\}.$$

Substitution into (190), using the coefficient bound in (36), proves (38). Only finite-interval energy and the fixed polynomial projection are used; no later PL inequality is assumed. □

Proof of Lemma C.6. Use the notation of Lemma C.4. The spatial balance at the present initialization is

$$w_{13}^2 + \frac{w_{24}^2+\|\mathbf{w}_{25}\|^2}{d} = v_{31}^2 + \frac{\sigma^2}{d}. \tag{191}$$

All constants in this proof may depend on the fixed scalar initialization, including σ, but not on d,n,T. Fix in advance

$$H_d = \exp\{\sqrt{\log d}\} = d^{o(1)}.$$

The fixed Gram factors satisfy $T\log(e+T)+T\log(n/d) = O(T\log(e+T)) = o(\sqrt{\log d})$. All estimates below hold on this fixed interval. In particular the scalar sizes are $d^{o(1)}$, and the error in (38) is $d^{-1/2+o(1)}$.

(i) Choose

$$\tau_b = d^{-1}\exp\{\tfrac{1}{4}\sqrt{\log d}\} = d^{-1+o(1)}.$$

Initially $w_{24}+\mathbf{w}_{25}\mathbf{e}_t = \sigma$ and $v_{31}=w_{13}=0$. Stop before τ_b at the first contact with

$$|v_{31}|+|w_{13}| = d^{-1/3}, \qquad \left\|(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)_{t=0}^{T-1} - \sigma\mathbf{1}_T\right\|_2 = \frac{1}{2}, \qquad v_{33} = 2v_{33}(0).$$

The initial loss bound gives $\sum_{t=0}^{T-1}\|A_t\|_{d,n}^2 = 2T\mathcal{L}_{\mathrm{GD}}/d \le C$. The norms of λ, $1-(1-\eta_0\lambda)^t$, and $\lambda\{1-(1-\eta_0\lambda)^t\}$ are uniformly bounded, by the same moment transfer used above. Direct differentiation of the residual gives on the stopped interval

$$|\dot v_{31}| \le C/\sqrt{T}, \qquad |\dot w_{13}| \le C|v_{31}|/\sqrt{T}, \qquad \left\| \left(\frac{\mathrm{d}}{\mathrm{d}\tau}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t) \right)_{t=0}^{T-1} \right\|_2 \le Cd|v_{31}|.$$

For the last bound, the positional realization has Gram matrix $\mathbf{I}_T + \mathbf{1}_T\mathbf{1}_T^{\mathsf{T}}$, of norm $T+1$, and the coefficient-gradient vector has norm at most C/T. Integrating gives

$$|v_{31}| \le C\tau_b/\sqrt{T} = d^{-1+o(1)}, \qquad |w_{13}| \le C\tau_b^2/T = d^{-2+o(1)},$$
$$\left\|(w_{24} + \mathbf{w}_{25}\mathbf{e}_t)_{t=0}^{T-1} - \sigma\mathbf{1}_T\right\|_2 \le Cd\tau_b^2/\sqrt{T} = d^{-1+o(1)}.$$

Also $|v_{33} - v_{33}(0)| \le \sqrt{\tau_b\mathcal{L}_{\mathrm{GD}}(0)/d} = d^{-1/2+o(1)}$. These strict bounds exclude all the stopping events.

For completeness, the projected transient on this layer can be bounded without any quotient-velocity closure. Its exact derivative is

$$\begin{aligned}\dot{\mathsf{z}} = &-2T\left| D_{v_{33}}\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right|^2 - 2dT \sum_{B\in\{44,45,54,55\}} \left\|\nabla_{\mathbf{W}_B}\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right\|_F^2 \\ &+ 2\sum_{t=1}^{T-1}\left\langle P_tA_t, (v_{31}w_{13})'P_t(\lambda p_t) + \big(v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)\big)'\lambda\right\rangle_{d,n}.\end{aligned}$$

Here, for $B \in \{44,45,54,55\}$,

$$D_{\mathbf{W}_B}A_t = P_tD_{\mathbf{W}_B}A_t, \qquad D_{v_{33}}A_t = P_tD_{v_{33}}A_t.$$

Applying (189) to the independent $\mathbf{W}_{55}$ squares and dividing by $2\sqrt{\mathsf{z}}$ gives

$$\begin{aligned}D^+\sqrt{\mathsf{z}} \le &-\frac{d}{T}\exp\{-CT\log(e+T)\}(d/n)^{CT}\sqrt{\mathsf{z}} \\ &+ C\left(\sqrt{T}\,|(v_{31}w_{13})'| + \|\big(v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)\big)'_{1\le t<T}\|_2\right).\end{aligned} \tag{192}$$

At $\mathsf{z} = 0$ the same inequality holds for the upper right Dini derivative. The preceding short-layer bounds make their sum at most C, since $\|(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)_{t=0}^{T-1}\|_2|\dot v_{31}| = O(1)$ and $|v_{31}|\left\|\big(\frac{\mathrm{d}}{\mathrm{d}\tau}(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)\big)_{t=0}^{T-1}\right\|_2 = d^{-1+o(1)}$. For any fixed C, the choice of τ_b ensures

$$\frac{d\tau_b}{T}\exp\{-CT\log(e+T)\}(d/n)^{CT} \ge 20\log d$$

for sufficiently large d. Variation of constants therefore yields

$$\begin{aligned}\sqrt{\mathsf{z}(\tau_b)} &\le d^{-20} + \frac{T}{d}\exp\{CT\log(e+T)\}(n/d)^{CT}, \\ \int_0^{\tau_b}\sqrt{\mathsf{z}}\,\mathrm{d}r &\le \frac{T}{d}\exp\{CT\log(e+T)\}(n/d)^{CT}(1+\tau_b).\end{aligned}$$

The exact v_{31}-gradient, separated by the fixed projections, is

$$\begin{aligned}\dot v_{31} = &-\frac{M(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)}{T}\{v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)-\eta_0\} - \frac{w_{13}\mathsf{G}}{T}(v_{31}w_{13}+\eta_0) \\ &-\frac{1}{T}\sum_{t=1}^{T-1}\langle P_tA_t, (w_{24}+\mathbf{w}_{25}\mathbf{e}_t)\lambda + w_{13}P_t(\lambda p_t)\rangle_{d,n}.\end{aligned}$$

The first term is at most $-1/(CT)$ throughout this layer. The absolute integral of the last line is at most

$$\frac{\sqrt{T}}{d}\exp\{CT\log(e+T)\}(n/d)^{CT}(1+\tau_b) = o(\tau_b/T),$$

and the integral of the second term is also $o(\tau_b/T)$. Therefore

$$v_{31}(\tau_b) \le -\tau_b/(CT) \le -1/(Cd), \qquad v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)\big|_{\tau_b} \ge 1/(Cd),$$
$$|v_{31}w_{13}|_{\tau_b} \le d^{-3+o(1)}, \qquad \mathsf{z}(\tau_b) \le d^{-2+o(1)}. \tag{193}$$

The exact loss decomposition now gives

$$\frac{\eta_0^2(M+\mathsf{G})}{2T} - \frac{\mathcal{L}_{\mathrm{GD}}(\tau_b)}{d} = \frac{1}{2T}\Big[2\eta_0 M v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0) - M\{v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)\}^2 - 2\eta_0\mathsf{G}v_{31}w_{13} - \mathsf{G}(v_{31}w_{13})^2 - \mathsf{z}\Big] \ge \frac{1}{CdT}. \tag{194}$$

At every point with $v_{31}=0$, both products vanish, so $\mathcal{L}_{\mathrm{GD}}/d \ge \eta_0^2(M+\mathsf{G})/(2T)$. This is a lower bound on that face, not a claim that the loss is constant there. Loss monotonicity and (194) therefore imply $v_{31}(\tau) < 0$ for every $\tau \ge \tau_b$.

(ii) Stop at the first time after τ_b with $-v_{31}w_{13} = 3\eta_0/4$, truncating at $H_d/2$ until the time comparison below excludes that truncation. Before this event, (38) and $v_{31} < 0$ give

$$w_{13} \ge -d^{-1/2+o(1)}, \qquad -d^{-1/2+o(1)} \le -v_{31}w_{13} \le 3\eta_0/4,$$

where the second inequality also uses the scalar energy bound on $[0, H_d]$. Thus $\mathsf{G}(-v_{31}w_{13} - \eta_0)^2/(2T) \le \eta_0^2(d/n)/T$.

The loss deficit in (194) is positive and nondecreasing. The loss decomposition and (191) give

$$\frac{\eta_0^2(M+\mathsf{G})}{2} - T\frac{\mathcal{L}_{\mathrm{GD}}}{d} \le C\left(|v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)| + \mathsf{G}|v_{31}w_{13}|\right),$$
$$|v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)| \le \sqrt{2}\,|v_{31}|\sqrt{\sigma^2+dv_{31}^2}, \qquad |v_{31}w_{13}| \le |v_{31}|\sqrt{v_{31}^2+\sigma^2/d}.$$

Consequently

$$v_{31}^2 \ge C^{-1}\min\left\{\left(\frac{\eta_0^2(M+\mathsf{G})}{2} - T\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right)^2, \frac{1}{\sqrt{d}}\left(\frac{\eta_0^2(M+\mathsf{G})}{2} - T\frac{\mathcal{L}_{\mathrm{GD}}}{d}\right)\right\}. \tag{195}$$

To check this inversion, absorb the fixed σ into C and bound the first line above by $C|v_{31}|\sqrt{1+dv_{31}^2}$, square, and distinguish $dv_{31}^2 \le 1$ from $dv_{31}^2 \ge 1$.

As long as $\mathcal{L}_{\mathrm{GD}}/d \ge \eta_0^2 M/(8T)$ and the stopping event has not occurred,

$$\frac{\mathcal{L}_{\mathrm{GD}}}{d} - \frac{\mathsf{G}}{2T}(v_{31}w_{13}+\eta_0)^2 \ge \frac{1}{CT}.$$

Since the deficit is at most C/T, (195) therefore yields

$$-\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathcal{L}_{\mathrm{GD}}}{d} \ge \exp\{-CT\log(e+T)\}(d/n)^{CT} \times \min\left\{d\left(\frac{\eta_0^2(M+\mathsf{G})}{2T} - \frac{\mathcal{L}_{\mathrm{GD}}}{d}\right)^2, \frac{\sqrt{d}}{T}\left(\frac{\eta_0^2(M+\mathsf{G})}{2T} - \frac{\mathcal{L}_{\mathrm{GD}}}{d}\right)\right\}. \tag{196}$$

Starting from the deficit $1/(CdT)$, integrate the quadratic lower bound until the deficit reaches $1/(T\sqrt{d})$, then the linear lower bound until it reaches $\eta_0^2(M+\mathsf{G})/(2T) - \eta_0^2 M/(8T)$. The resulting time is at most

$$T\exp\{CT\log(e+T)\}(n/d)^{CT}(1+d^{-1/2}\log d) = \exp\{o(\sqrt{\log d})\}.$$

Thus within that time either $-v_{31}w_{13} = 3\eta_0/4$, or

$$\frac{\mathcal{L}_{\mathrm{GD}}}{d} \le \frac{\eta_0^2 M}{8T}, \qquad \frac{\eta_0}{2} \le v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0) \le \frac{3\eta_0}{2}, \qquad -v_{31} \ge C^{-1}d^{-1/4}. \tag{197}$$

The last two conclusions follow from the time-zero term of the loss and $|v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)|^2 \le 2d\,v_{31}^2(v_{31}^2+\sigma^2/d)$. They persist thereafter by loss monotonicity. This argument does not assume that the positional product follows a pointwise reduced product equation.

(iii) If the stopping event has not already occurred, use (197). In (38), the non-error primitive is

$$w_{13}(\tau_b) + \frac{\mathsf{G}}{T}\int_{\tau_b}^{\tau}(-v_{31})(\eta_0 + v_{31}w_{13})\,\mathrm{d}r.$$

It is nondecreasing before the stopping event. After the anchor sublevel is reached, its derivative is at least $C^{-1}d^{-1/4}\mathsf{G}/T$. Hence in an additional CT/G time this primitive reaches $4d^{-1/4}$, unless $-v_{31}w_{13} = 3\eta_0/4$ occurs first. Its distance from the actual w_{13} is uniformly at most $d^{-1/2+o(1)} = o(d^{-1/4})$.

Once the primitive reaches that level, it remains there or increases. Thus $w_{13} \geq 3d^{-1/4}$ and the spatial balance gives $-v_{31} \geq \sqrt{w_{13}^2 - \sigma^2/d} \geq w_{13}/2$ for sufficiently large d, since σ is fixed. Its derivative is consequently at least $C^{-1}\mathsf{G}/T$ times its own value, after absorbing the $o(d^{-1/4})$ difference from w_{13}. Exponential comparison makes it reach two in $C(T/\mathsf{G})\log d$ additional time. At that point $w_{13} \geq 3/2$ and $-v_{31}w_{13} \geq w_{13}\sqrt{w_{13}^2 - \sigma^2/d} > 3\eta_0/4$, a contradiction. The stopping event therefore occurs in the claimed time. The total time used in (i)-(iii) is strictly smaller than $H_d/2$ for all sufficiently large d, so the use of its finite-interval energy bound is not circular. Notice that no assertion that w_{13} retains its initial Taylor sign was required: the anchor first creates a drift larger than the cumulative projected error.

(iv) Let τ_* be the first time with $-v_{31}w_{13} = 3\eta_0/4$. The spatial balance gives

$$(3\eta_0/4)^2 = v_{31}^2 w_{13}^2 \leq v_{31}^2(v_{31}^2 + \sigma^2/d), \qquad |v_{31}(\tau_*)| \geq C^{-1}.$$

On the next interval of length τ_b, the scalar energy estimate gives

$$|v_{31}(\tau) - v_{31}(\tau_*)| + |w_{13}(\tau) - w_{13}(\tau_*)| \leq C\sqrt{\tau_b/T} = d^{-1/2+o(1)}.$$

The scalar sizes on $[0, H_d]$ are $d^{o(1)}$, so uniformly there $-v_{31}w_{13} = 3\eta_0/4 + o(1)$ and $|v_{31}| \geq 1/(2C)$. The fast dissipation inequality now implies

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathcal{L}_{\mathrm{GD}}}{d} \leq -\frac{d}{T}\exp\{-CT\log(e+T)\}(d/n)^{CT}\left[\frac{\mathcal{L}_{\mathrm{GD}}}{d} - \left(\frac{\eta_0^2}{32} + o(1)\right)\frac{\mathsf{G}}{T}\right].$$

The same rate-times-length bound used in (i) and variation of constants give at $\tau_0 = \tau_* + \tau_b = d^{o(1)}$

$$\frac{\mathcal{L}_{\mathrm{GD}}(\tau_0)}{d} \leq \left(\frac{\eta_0^2}{32} + o(1)\right)\frac{\mathsf{G}}{T} < \frac{9\eta_0^2\mathsf{G}}{32T}.$$

The time-zero term and $\mathsf{G}/M \leq 2d/n \leq 1/5$ give

$$|v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0) - \eta_0|^2 \leq \left(\frac{\eta_0^2}{16} + o(1)\right)\frac{\mathsf{G}}{M} < \frac{\eta_0^2}{4}.$$

Together with $-v_{31}w_{13} = 3\eta_0/4 + o(1)$ and $v_{31} < 0$, this proves (39). Every point with $-v_{31}w_{13} \leq \eta_0/4$ has loss at least $9\eta_0^2\mathsf{G}/(32T)$ by (33). Loss monotonicity excludes that set at all later times. Finally, $(\eta_0/4)^2 < (v_{31}w_{13})^2 \leq v_{31}^2(v_{31}^2 + \sigma^2/d)$ gives the uniform lower bound on $|v_{31}|$ in (40). □

G.5 The Stage-1 endpoint

Lemma G.14. *At every Stage-1 point,*

$$\begin{aligned}
&|v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0) - \eta_0|^2 + |v_{31}w_{13} + \eta_0|^2 \\
&\quad + \sum_{t=1}^{T-1}|v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t) + \eta_0(\widetilde{\mathbf{W}}_{55})_{1t} - \eta_0|^2 + \sum_{t=1}^{T-1}\sum_{\ell=2}^{t}|(\widetilde{\mathbf{W}}_{55})_{\ell t}|^2 \\
&\leq \exp\{CT\log(e+T)\}(n/d)^{CT}\frac{\mathcal{L}_{\mathrm{GD}}}{d} = d^{o(1)}\frac{\mathcal{L}_{\mathrm{GD}}}{d}.
\end{aligned} \tag{198}$$

Proof. Write A_t from (22) in the monomial basis. The $t = 0$ row first selects $v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_0) - \eta_0$. For $t \geq 1$, the degree-$(t+1)$ coefficient selects $v_{31}w_{13} + \eta_0$. After removing it, the successive degrees of

$$\lambda, \quad 1 - (1 - \eta_0\lambda)^2, \ldots, 1 - (1 - \eta_0\lambda)^t$$

select $v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)+\eta_0(\widetilde{\mathbf{W}}_{55})_{1t}-\eta_0$ and $(\widetilde{\mathbf{W}}_{55})_{2t},\dots,(\widetilde{\mathbf{W}}_{55})_{tt}$. This triangular coefficient map has inverse norm at most $\exp\{CT\log(e+T)\}$. The normalized finite moment Gram inverse through degree T contributes at most $(n/d)^{CT}$. Applying these two bounds to the sum of squared residual norms proves the first inequality. The equality $d^{o(1)}$ follows from (17). □

Lemma G.15. *At any Stage-1 time satisfying $\mathcal{L}_{\mathrm{GD}}\le d^{-1/4}$,*

$$\begin{aligned}|v_{31}|+|w_{13}|+|v_{33}|&\le d^{o(1)},\\ |w_{24}|+\|\mathbf{w}_{25}\|+\sum_{B\in\{44,45,54,55\}}\|\mathbf{W}_B\|_F&\le d^{o(1)},\\ \max_{1\le\ell\le t<T}|(\widetilde{\mathbf{W}}_{55})_{\ell t}|&\le d^{o(1)}.\end{aligned}\tag{199}$$

Proof. First consider the first time the loss reaches $d^{-1/4}$. Lemmas C.6 and C.7 bound this time by

$$d^{o(1)}+C\frac{T}{d/n}\log(d^{5/4})=d^{o(1)}.$$

The scalar energy identity gives, up to this time,

$$\sup_\tau(|v_{31}|+|w_{13}|+|v_{33}|)+\int(|\dot v_{31}|+|\dot w_{13}|+|\dot v_{33}|)\,\mathrm{d}\tau\le d^{o(1)},\qquad v_{33}\ge v_{33}(0)>0.$$

For $1\le\ell\le t<T$, the two coefficient maps are

$$\begin{aligned}&\big[\mathbb{1}_{\{\ell=1\}}(1\ \ \mathbf{e}_t^\mathsf{T})\big]_{\ell,t}\begin{pmatrix}w_{24}\\ \mathbf{w}_{25}^\mathsf{T}\end{pmatrix}=\big(\mathbb{1}_{\{\ell=1\}}(w_{24}+\mathbf{w}_{25}\mathbf{e}_t)\big)_{\ell,t},\\ &\big[(1\ \ \mathbf{e}_t^\mathsf{T}\ \ \mathbf{e}_\ell^\mathsf{T}\ (\mathbf{e}_t\otimes\mathbf{e}_\ell)^\mathsf{T})\big]_{\ell,t}\begin{pmatrix}w_{44}\\ \mathbf{w}_{45}^\mathsf{T}\\ \mathbf{w}_{54}\\ \mathrm{vec}(\mathbf{W}_{55})\end{pmatrix}\\ &\qquad=\big(w_{44}+\mathbf{w}_{45}\mathbf{e}_t+\mathbf{e}_\ell^\mathsf{T}\mathbf{w}_{54}+\mathbf{e}_\ell^\mathsf{T}\mathbf{W}_{55}\mathbf{e}_t\big)_{\ell,t}.\end{aligned}\tag{200}$$

Their second row Gram matrix satisfies

$$\big[(1+\mathbb{1}_{\{\ell=j\}})(1+\mathbb{1}_{\{t=s\}})\big]_{(\ell,t),(j,s)}\succeq\mathbf{I},\qquad\big\|\big[(1+\mathbb{1}_{\{\ell=j\}})(1+\mathbb{1}_{\{t=s\}})\big]\big\|_2\le CT^2.$$

Define the fixed matrix

$$\begin{aligned}\mathbf{K}:=&\left[\mathbf{I}-\frac12\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}(1\ \ \mathbf{e}_0^\mathsf{T})\right]\big[\mathbb{1}_{\{\ell=1\}}\begin{pmatrix}1\\ \mathbf{e}_t\end{pmatrix}\big]_{\ell,t}\\ &\cdot\big[(1+\mathbb{1}_{\{\ell=j\}})(1+\mathbb{1}_{\{t=s\}})\big]^{-1}_{(\ell,t),(j,s)}\\ &\cdot\big[(1\ \ \mathbf{e}_s^\mathsf{T}\ \ \mathbf{e}_j^\mathsf{T}\ (\mathbf{e}_s\otimes\mathbf{e}_j)^\mathsf{T})\big]_{j,s},\qquad\|\mathbf{K}\|_2\le CT^2.\end{aligned}\tag{201}$$

The chain rule and the zero initialization of $\mathbf{W}_{44},\mathbf{W}_{45},\mathbf{W}_{54},\mathbf{W}_{55}$ give

$$\begin{aligned}&\begin{pmatrix}w_{44}\\ \mathbf{w}_{45}^\mathsf{T}\\ \mathbf{w}_{54}\\ \mathrm{vec}(\mathbf{W}_{55})\end{pmatrix}\in\mathrm{range}\big[(1\ \ \mathbf{e}_t^\mathsf{T}\ \ \mathbf{e}_\ell^\mathsf{T}\ (\mathbf{e}_t\otimes\mathbf{e}_\ell)^\mathsf{T})\big]_{\ell,t}^\mathsf{T},\\ &\sum_{B\in\{44,45,54,55\}}\|\mathbf{W}_B\|_F^2\le\sum_{1\le\ell\le t<T}\big|w_{44}+\mathbf{w}_{45}\mathbf{e}_t+\mathbf{e}_\ell^\mathsf{T}\mathbf{w}_{54}+\mathbf{e}_\ell^\mathsf{T}\mathbf{W}_{55}\mathbf{e}_t\big|^2,\end{aligned}\tag{202}$$

$$\left[\mathbf{I}-\frac12\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}(1\ \ \mathbf{e}_0^\mathsf{T})\right]\begin{pmatrix}\dot w_{24}\\ \dot{\mathbf{w}}_{25}^\mathsf{T}\end{pmatrix}=\frac{v_{31}}{\eta_0 v_{33}}\mathbf{K}\begin{pmatrix}\dot w_{44}\\ \dot{\mathbf{w}}_{45}^\mathsf{T}\\ \dot{\mathbf{w}}_{54}\\ \mathrm{vec}(\dot{\mathbf{W}}_{55})\end{pmatrix}.\tag{203}$$

By (20) and $v_{31}(w_{24}+\mathbf{w}_{25}\mathbf{e}_0)=\widetilde{\theta}_1$,

$$\mathbf{K}\begin{pmatrix} w_{44}\\ \mathbf{w}_{45}^{\mathsf{T}}\\ \mathbf{w}_{54}\\ \mathrm{vec}(\mathbf{W}_{55})\end{pmatrix}+\frac{v_{31}}{\eta_0 v_{33}}\mathbf{K}\mathbf{K}^{\mathsf{T}}\left[\mathbf{I}-\frac{1}{2}\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}(1\ \ \mathbf{e}_0^{\mathsf{T}})\right]\begin{pmatrix}w_{24}\\ \mathbf{w}_{25}^{\mathsf{T}}\end{pmatrix}$$
$$=\mathbf{K}\begin{pmatrix} 0\\ \mathbf{0}_T\\ \mathbf{0}_T\\ \mathrm{vec}\left(\left[\frac{(\widetilde{\mathbf{\Theta}}_3)_{\ell t}-\frac{1}{2}\widetilde{\theta}_1\mathbb{1}_{\{\ell=1\}}}{\eta_0^{\mathbb{1}_{\{\ell=1\}}}v_{33}}\right]_{\ell,t}\right)\end{pmatrix}. \tag{204}$$

Equation (198) implies $\|\,\mathrm{RHS}\,(204)\|_2\le d^{o(1)}$. Differentiating (204) and using (203) yields

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left\{\left(\mathbf{I}+\frac{v_{31}^2}{\eta_0^2v_{33}^2}\mathbf{K}\mathbf{K}^{\mathsf{T}}\right)^{1/2}\left[\mathbf{I}-\frac{1}{2}\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}(1\ \ \mathbf{e}_0^{\mathsf{T}})\right]\begin{pmatrix}w_{24}\\ \mathbf{w}_{25}^{\mathsf{T}}\end{pmatrix}\right\}$$
$$=\frac{v_{31}}{\eta_0 v_{33}}\left(\mathbf{I}+\frac{v_{31}^2}{\eta_0^2v_{33}^2}\mathbf{K}\mathbf{K}^{\mathsf{T}}\right)^{-1/2}\frac{\mathrm{d}}{\mathrm{d}\tau}\{\mathrm{RHS}\,(204)\}.$$

Integration by parts gives

$$\left[\left(\mathbf{I}+\frac{v_{31}^2}{\eta_0^2v_{33}^2}\mathbf{K}\mathbf{K}^{\mathsf{T}}\right)^{1/2}\left[\mathbf{I}-\frac{1}{2}\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}(1\ \ \mathbf{e}_0^{\mathsf{T}})\right]\begin{pmatrix}w_{24}\\ \mathbf{w}_{25}^{\mathsf{T}}\end{pmatrix}\right.$$
$$\left.-\frac{v_{31}}{\eta_0 v_{33}}\left(\mathbf{I}+\frac{v_{31}^2}{\eta_0^2v_{33}^2}\mathbf{K}\mathbf{K}^{\mathsf{T}}\right)^{-1/2}\{\mathrm{RHS}\,(204)\}\right]_0^{\tau} \tag{205}$$
$$=-\int_0^{\tau}\left(\frac{\mathrm{d}}{\mathrm{d}s}\frac{v_{31}}{\eta_0 v_{33}}\right)\left(\mathbf{I}+\frac{v_{31}^2}{\eta_0^2v_{33}^2}\mathbf{K}\mathbf{K}^{\mathsf{T}}\right)^{-3/2}\{\mathrm{RHS}\,(204)\}\,\mathrm{d}s.$$

Since

$$\left\|\left(\mathbf{I}+\frac{v_{31}^2}{\eta_0^2v_{33}^2}\mathbf{K}\mathbf{K}^{\mathsf{T}}\right)^{-\alpha}\right\|_2\le 1\quad(\alpha\ge 0),\qquad \sup_{\tau}\left|\frac{v_{31}}{\eta_0 v_{33}}\right|+\int\left|\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{v_{31}}{\eta_0 v_{33}}\right|\mathrm{d}\tau\le d^{o(1)},$$

we obtain

$$\sup_{\tau}\left\|\left[\mathbf{I}-\frac{1}{2}\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}(1\ \ \mathbf{e}_0^{\mathsf{T}})\right]\begin{pmatrix}w_{24}(\tau)\\ \mathbf{w}_{25}(\tau)^{\mathsf{T}}\end{pmatrix}\right\|_2\le d^{o(1)}. \tag{206}$$

At the first hitting time, (198) and (30) give

$$|v_{31}|\ge c,\qquad \left\|\frac{\widetilde{\theta}_1}{2v_{31}}\begin{pmatrix}1\\ \mathbf{e}_0\end{pmatrix}\right\|_2\le d^{o(1)}.$$

Thus (202) and (20) imply

$$|w_{24}|+\|\mathbf{w}_{25}\|_2+\sum_{B\in\{44,45,54,55\}}\|\mathbf{W}_B\|_F\le d^{o(1)}.$$

In fact, the coefficient estimate and the two identities in (30)–(31) also give, at this endpoint,

$$v_{31}^2=\eta_0+o(1),\qquad w_{13}^2=\eta_0+o(1),\qquad v_{33}^2=v_{33}(0)^2+d^{-1+o(1)}. \tag{207}$$

The signs are $v_{31}<0<w_{13}$.

For any subsequent time τ, the raw energy identity and (43) give

$$\int_{\tau}^{\infty}\left(\|\dot{\mathbf{V}}(s)\|_F^2+\|\dot{\mathbf{W}}(s)\|_F^2\right)^{1/2}\mathrm{d}s\le C\sqrt{\frac{T}{d/n}\mathcal{L}_{\mathrm{GD}}(\tau)}\le d^{-1/8+o(1)}.$$

This follows by integrating $-\mathcal{L}'_{\mathrm{GD}}/\sqrt{\mathcal{L}_{\mathrm{GD}}}$ and using (43). Thus the raw bounds remain valid at every later Stage-1 time, and the balance identities again give (207). □

G.6 Raw realization of auxiliary parameters

Lemma G.16. *Every finite choice of the coefficients*

$$\widetilde{\theta}_1, \quad \widetilde{\theta}_2, \quad \beta_t := v_{31}(w_{24} + \mathbf{w}_{25}\mathbf{e}_t) + \widetilde{\theta}_1(\widetilde{\mathbf{W}}_{55})_{1t} \quad (1 \le t < T), \quad (\widetilde{\mathbf{W}}_{55})_{\ell t}\ (2 \le \ell \le t < T)$$

has a raw representative with $\mathbf{W}_{33} = \mathbf{0}_{d\times d}$. *Moreover, the representative can be chosen on the inherited Stage-2 leaf*

$$v_{31}^2 - w_{13}^2 - \frac{w_{24}^2 + \|\mathbf{w}_{25}\|^2}{d} = -\frac{\sigma^2}{d},$$
$$dv_{33}^2 - \sum_{B\in\{44,45,54,55\}} \|\mathbf{W}_B\|_F^2 = d(v_{33}^{\mathrm{init}})^2,$$
$$\sum_{t=0}^{T-1} (\mathbf{w}_{25})_t - w_{24} = -\sigma.$$

For $\widetilde{\theta}_1 \neq 0$, *the conversion to* (46) *is*

$$\beta_t = \widetilde{\theta}_1(\widetilde{\mathbf{\Theta}}_4)_{1t} \quad (1 \le t < T), \qquad (\widetilde{\mathbf{W}}_{55})_{\ell t} = (\widetilde{\mathbf{\Theta}}_4)_{\ell t} \quad (2 \le \ell \le t < T).$$

Proof. Set $\mathbf{W}_{33} = \mathbf{0}_{d\times d}$ and all unused blocks to zero. First set $(\widetilde{\mathbf{W}}_{55})_{1t} = 0$, which is allowed because it occurs only through β_t. Choose $v_{31} < 0$, set $w_{13} = \widetilde{\theta}_2/v_{31}$, and define

$$w_{24} := \frac{\widetilde{\theta}_1 + \sum_{t=1}^{T-1}\beta_t + \sigma v_{31}}{v_{31}(T+1)},$$
$$(\mathbf{w}_{25})_0 := \frac{\widetilde{\theta}_1}{v_{31}} - w_{24}, \qquad (\mathbf{w}_{25})_t := \frac{\beta_t}{v_{31}} - w_{24} \quad (1 \le t < T).$$

Then the positional kernel has value $-\sigma$, $v_{31}(w_{24} + (\mathbf{w}_{25})_0) = \widetilde{\theta}_1$, $v_{31}(w_{24} + (\mathbf{w}_{25})_t) = \beta_t$ for $t \ge 1$, and $v_{31}w_{13} = \widetilde{\theta}_2$. Put

$$\widetilde{\theta}_1^2 + \sum_{t=1}^{T-1}\beta_t^2 - \frac{\left(\widetilde{\theta}_1 + \sum_{t=1}^{T-1}\beta_t\right)^2}{T+1} > 0 \quad \text{unless } \widetilde{\theta}_1 = \beta_1 = \cdots = \beta_{T-1} = 0,$$

because the displayed quadratic form is positive definite. The positional square norm is

$$w_{24}^2 + \|\mathbf{w}_{25}\|^2 = \frac{1}{v_{31}^2}\left[\widetilde{\theta}_1^2 + \sum_{t=1}^{T-1}\beta_t^2 - \frac{\left(\widetilde{\theta}_1 + \sum_{t=1}^{T-1}\beta_t\right)^2}{T+1}\right] + \frac{\sigma^2}{T+1}.$$

If $\widetilde{\theta}_2 \neq 0$ or $(\widetilde{\theta}_1, \beta_1, \ldots, \beta_{T-1}) \neq 0$, the spatial invariant as a function of $\rho = v_{31}^2 > 0$ is

$$v_{31}^2 - w_{13}^2 - \frac{w_{24}^2 + \|\mathbf{w}_{25}\|^2}{d} = \rho - \frac{(-\widetilde{\theta}_2)^2 + \frac{1}{d}\left[\widetilde{\theta}_1^2 + \sum_{t=1}^{T-1}\beta_t^2 - \frac{\left(\widetilde{\theta}_1 + \sum_{t=1}^{T-1}\beta_t\right)^2}{T+1}\right]}{\rho} - \frac{\sigma^2}{d(T+1)}.$$

Its derivative is strictly positive, and its limits at zero and infinity are $-\infty$ and $+\infty$. There is therefore a unique $\rho > 0$, hence a unique $v_{31} = -\sqrt{\rho} < 0$, for which the invariant equals $-\sigma^2/d$. If $\widetilde{\theta}_2 = 0$ and $\widetilde{\theta}_1 = \beta_1 = \cdots = \beta_{T-1} = 0$, use

$$v_{31} = 0, \qquad w_{13} = \sqrt{\frac{T\sigma^2}{d(T+1)}}, \qquad w_{24} = \frac{\sigma}{T+1}, \qquad (\mathbf{w}_{25})_t = -\frac{\sigma}{T+1},$$

which has zero positional products, positional kernel $-\sigma$, and spatial invariant equal to $-\sigma^2/d$.

To realize the coefficients $(\widetilde{\mathbf{W}}_{55})_{\ell t}$, choose $v_{33} > 0$, set $(\mathbf{W}_{55})_{\ell t} = (\widetilde{\mathbf{W}}_{55})_{\ell t}/v_{33}$ for $2 \le \ell \le t < T$, set $w_{44} = \mathbf{w}_{45} = \mathbf{w}_{54} = 0$, and set all other entries of $\mathbf{W}_{55}$ to zero. The auxiliary $(\widetilde{\mathbf{W}}_{55})_{\ell t}$'s are then prescribed. If $\sum_{t=1}^{T-1}\sum_{\ell=2}^{t}|(\widetilde{\mathbf{W}}_{55})_{\ell t}|^2 > 0$, the left side of (52) is

$$dv_{33}^2 - \frac{\sum_{t=1}^{T-1}\sum_{\ell=2}^{t}|(\widetilde{\mathbf{W}}_{55})_{\ell t}|^2}{v_{33}^2},$$

which is strictly increasing from $-\infty$ to $+\infty$; if the sum is zero, choose $v_{33} = v_{33}^{\mathrm{init}}$. Thus v_{33} can always be chosen to give $d(v_{33}^{\mathrm{init}})^2$, while remaining positive. The resulting raw recurrence is

$$p_0 = 0, \qquad p_1 = \widetilde{\theta}_1 \lambda, \qquad p_{t+1} = (1 + \widetilde{\theta}_2 \lambda) p_t + \beta_t \lambda + \sum_{\ell=2}^{t} (\widetilde{\mathbf{W}}_{55})_{\ell t} p_\ell .$$

For $\widetilde{\theta}_1 \neq 0$, the displayed conversion gives (47). On the nonzero spatial region, the strict derivatives in the two scalar invariant equations and the implicit-function theorem give a continuous section of raw representatives on this algebraic leaf. The isolated spatial origin is handled by the special representative above; no continuity through that origin or connectivity of an arbitrary raw gauge fibre is asserted. □