ADAPTIVE OPTIMIZATION IMPROVES SCALING IN QUADRATIC NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the scaling laws of adaptive training algorithms for quadratic neural networks, and characterize scaling exponents parameterized by target complexity and model size. We demonstrate that an adaptive algorithm utilizing accumulated prediction errors can significantly improve the scaling behavior compared to fixed-learning-rate methods. Notably, this approach achieves statistically optimal exponents at any time-step without requiring manual learning rate tuning. Finally, we empirically demonstrate that the acceleration phenomenon due to adaptive learning rate generalizes: similar scaling improvements are observed under general activation functions and when employing practical adaptive algorithms that accumulate squared gradients.

1 INTRODUCTION

When pretraining large neural networks, loss scales predictably as a power law with respect to dataset size, parameter count, and total compute. This consistent behavior allows practitioners to reliably project the performance of computationally expensive models from small-scale preliminary runs (Kaplan et al., 2020; Bahri et al., 2024). Driven by its empirical significance, the theoretical community has increasingly sought to derive scaling laws from first principles (Michaud et al., 2024; Nam et al., 2024; Paquette et al., 2024; Lin et al., 2024; Defilippis et al., 2025).

While early theoretical analyses primarily focused on linear models (Paquette et al., 2024; Lin et al., 2024), recent work has expanded toward solvable nonlinear architectures capable of capturing deep learning phenomena like feature learning (Ren et al., 2025; Ben Arous et al., 2025; Defilippis et al., 2025). However, existing scaling analyses for nonlinear networks predominantly analyze standard SGD with a fixed learning rate (Ren et al., 2025; Ben Arous et al., 2025), which requires data-dependent learning-rate tuning to recover optimal scaling exponents. On the other hand, real-world practitioners rely almost exclusively on adaptive optimizers like Adam, as they dynamically adjust step sizes based on the local geometry of the loss landscape. This often leads to superior scaling across compute regimes (Glentis et al., 2026; Zhao et al., 2024). This motivates our core question:

Can we rigorously establish that adaptive optimization yields superior scaling laws compared to SGD with a fixed learning rate in the feature learning regime?

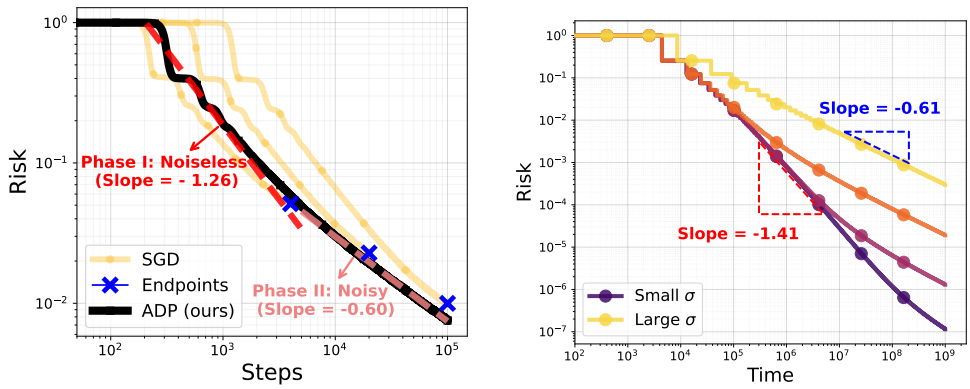
We answer this question in the affirmative. Specifically, we mathematically demonstrate the scaling benefits of adaptive optimization by analyzing the feature learning dynamics of a two-layer neural network with quadratic activations. In our formulation, the target function and input distribution are modeled as:

$$y = f_*(\boldsymbol{x}) + \sigma\epsilon \;\text{ with }\; f_*(\boldsymbol{x}) = \textstyle\sum_{j=1}^{r} \lambda_j \phi\left(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle\right) \text{ and } \boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d), \tag{1.1}$$

where $\phi(z) = \frac{z^2-1}{\sqrt{2}}$ is the 2nd Hermite polynomial; ϵ is zero-mean, unit-variance independent sub-exponential noise; $\sigma > 0$ dictates the label-noise strength; $\{\boldsymbol{\theta}_j\}_{j=1}^r \subset \mathbb{R}^d$ are orthogonal signal directions, and $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$ denote their associated signal strengths. Our goal is to learn this target network using a two-layer student network of width r_s trained via adaptive gradient methods, establishing exact scaling exponents for the resulting representation learning dynamics.

Several classical high-dimensional inference problems, including phase retrieval ($r = 1$) (Chen & Candes, 2015; Tan & Vershynin, 2023) and online PCA ($r > 1$) (Oja & Karhunen, 1985; Allen-Zhu

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.



(a) Online SGD dynamics

(b) Closed-form population dynamics

Figure 1: Risk trajectory for learning (1.1) with signal exponent $\alpha = 1.25$. (a) Online SGD dynamics (sampling without replacement) compared against fixed-LR SGD baselines. (b) Exact closed-form solutions for the proposed adaptive algorithm under the population dynamics in (3.1). In (a), we set $\sigma = 0.5$, and the SGD baselines are evaluated across different training durations $T \in \{4 \times 10^3, 2 \times 10^4, 10^5\}$, with the learning rate optimally tuned for each specific horizon (see Appendix A for experimental details). Crosses indicate the final risk of the fixed-LR SGD baselines at the end of training. The adaptive algorithm is trained only once for $T = 10^5$. In (b), we evaluate the exact dynamics under noise levels $\sigma \in \{0.001, 0.01, 0.1, 1\}$. In both settings, we consistently observe two distinct linear phases whose exponents match with those of the noiseless and noisy ERM.

& Li, 2017), can be formulated as learning quadratic networks. More broadly, these belong to the family of additive multi-index models (Oko et al., 2024; Defilippis et al., 2026; Girardin et al., 2026; Ren & Lee, 2024; Ben Arous et al., 2024), where the target function depends on a small number of latent directions that are nonlinearly transformed and additively combined, as in (1.1).

Prior work derived the scaling rate for this exact setting under a power-law assumption on the signal strengths ($\lambda_i \propto i^{-\alpha}$) (Girardin et al., 2026; Ren et al., 2025; Ben Arous et al., 2025) . This recovers α-dependent scaling as feature directions are learned sequentially based on their signal strength. The learning rate in this setting, however, must be scaled proportionally with the target signal to ensure that the accumulated noise remains smaller than the signal (referred to as *unstable discretization* in Ren et al. (2025)). Our work builds on this insight: we show that accumulating the learning trajectory to adapt the learning rate modulates the signal-to-noise ratio dynamically and recover the optimal learning schedule naturally. The exact rates and comparisons are detailed in the following section.

1.1 Our Contributions

Our main contributions are summarized below (see also Table 1):

- *Scaling behavior of adaptive algorithms:* In Section 3, we analyze both the population gradient flow and online SGD dynamics under an adaptive learning rate schedule that accumulates prediction errors. We provide the exact asymptotic limit of the excess risk as a function of training time t and network width r_s. Notably, our results reveal two distinct phases: an initial *Noiseless phase* with an exponent of $1 - 2\alpha$ (recovering the exponent of noiseless ERM), followed by a slower *noisy phase* with an exponent of $\frac{1-2\alpha}{2\alpha}$ (matching the noisy ERM). These dynamics are illustrated in Figure 1.
- *Improvement over SGD with a fixed LR:* To demonstrate the advantage of adaptive training, we evaluate it against SGD with a tuned, fixed learning rate across various training durations. As shown in Figure 1, a single adaptive run matches the performance of optimally tuned SGD across all time horizons. In Theorem 1, we formally prove that the proposed adaptive schedule dynamically matches the optimal SGD step size at any given time horizon. Consequently, our adaptive algorithm achieves the best possible scaling performance of SGD with a fixed LR without requiring manual tuning, which we empirically verify in Figures 1 and 2.
- *Connection to practical methods.* We empirically validate our findings beyond what was covered by our theoretical framework. Specifically, we compare our proposed method against practical adaptive

Algorithm	Risk scaling	Requires tuning?
ERM Girardin et al. (2026)	$\begin{cases} (\frac{d}{N})^{(1-2\alpha)} + r_s^{1-2\alpha} & \sigma = 0 \\ (\frac{d}{N})^{\frac{1-2\alpha}{2\alpha}} + r_s^{1-2\alpha} & \sigma > 0 \end{cases}$	–
SGD (with a fixed LR) Ren et al. (2025); Ben Arous et al. (2025)	$(\eta\bar{t})^{\frac{1-2\alpha}{\alpha}} + r_s^{1-2\alpha}$	Yes
Adaptive (Ours)	$(1 + \bar{t}^{2\alpha} \vee \frac{\bar{t}}{\sigma^2})^{\frac{1-2\alpha}{2\alpha}} + r_s^{1-2\alpha}$	**No**

Table 1: Scaling laws for learning the quadratic neural network in (1.1) with signal strengths $\lambda_i \propto i^{-\alpha}$ using the empirical risk minimizer (ERM), SGD with a fixed learning rate (LR), and the adaptive scheme proposed in (3.6). Here, r_s denotes the width of the student network, and σ denotes the standard deviation of the label noise.

- For ERM, N denotes the sample size. We display the sample complexity for the regimes in Girardin et al. (2026) that are applicable to our setting (low regularization, sufficient samples).
- For SGD, $\bar{t} \sim t/(d\,\mathrm{polylog}(d))$, where t is the number of gradient steps (which equals the sample size). For fixed-LR SGD, η denotes the learning rate that must be tuned as a function of $\bar{t}$. See Figure 1 for a comparison of the scaling rates achieved by the adaptive scheme versus tuned fixed-LR SGD.

algorithms like Adam and RMSprop, which accumulate squared gradients rather than prediction errors. Furthermore, we extend this evaluation to networks utilizing ReLU and absolute value activations. We observe that our adaptive algorithm exhibits risk trajectories that closely match those of these practical optimizers (see Figure 2), which suggests that the adaptive mechanism analyzed here shares key behavioral characteristics with widely used practical methods.

1.2 Additional Related Works

Adaptive optimization. Adaptive methods have been extensively studied since the advent of modern deep learning, following empirical evidence of their ability to improve training performance (Tieleman & Hinton, 2012; Kingma & Ba, 2014; Loshchilov & Hutter, 2017). Existing literature on these methods generally falls into two categories: an optimization perspective, which seeks to understand why they converge faster (Duchi et al., 2011; D'efossez et al., 2020), and a generalization perspective, which explores whether and why models trained with them generalize more effectively (Wilson et al., 2017; Li et al., 2025; Tang et al., 2025). Studies focusing on optimization have yielded both positive results regarding their adaptivity to landscape geometry (Ward et al., 2018; Li & Orabona, 2019), as well as negative results highlighting convergence failures even in simplistic settings (Reddi et al., 2018). From the generalization perspective, theoretical work has sought to analyze their behavior in task-dependent settings, often framing their benefits in terms of robustness to noise (Zhang et al., 2020) or through the lens of implicit bias (Xie & Li, 2024; Zhang et al., 2024).

Recently, there has been a growing interest in understanding the high-dimensional asymptotics of training algorithms to bridge these two perspectives (Ben Arous et al., 2022; Collins-Woodfin et al., 2023; 2024; Jagannath et al., 2025). The work closest to ours is Collins-Woodfin et al. (2024), which analyzed an AdaGrad-style algorithm for generalized linear models. However, their analysis only permits time horizons proportional to the ambient dimension, which cannot capture non-convex settings with high-dimensional saddle points, such as phase retrieval (Ben Arous et al., 2021). Our work characterizes the adaptive scaling laws beyond linear time scales, and shows the first instance of an analogous result in a non-linear feature learning setting.

Feature learning with multi-index teacher models. Learning low-dimensional structure from high-dimensional data is a fundamental problem in statistics (Li & Duan, 1989). The simplest case, where $r = 1$ (the *single-index model*), has been extensively studied in the feature learning literature (Ben Arous et al., 2021; Ba et al., 2022; Mousavi-Hosseini et al., 2022; Ba et al., 2023; Mousavi-Hosseini et al., 2023; Moniri et al., 2023; Mahankali et al., 2023; Berthier et al., 2024; Damian et al., 2024; Glasgow et al., 2025). More generally, settings where $r > 1$ fall within the broader class of *multi-index models*. While recent analyses have established sharp theoretical guarantees for these architectures, they predominantly rely on fixed learning rate schedules (Abbe et al., 2022; 2023; Bietti et al., 2023; Ben Arous et al., 2021; 2024; Ferbach et al., 2025). Consequently,

a theoretical understanding of whether—and how—adaptive optimization helps feature learning remains elusive. In this work, we demonstrate that adaptive optimization is beneficial for feature learning even under stochastic dynamics, moving beyond the scope and applicability of earlier results in both the feature learning and adaptive optimization literatures.

2 Problem Setting

Notation. We use $\langle \cdot, \cdot \rangle$ and $\|\cdot\|_2$ to denote the Euclidean inner product and norm, respectively. For matrices, $\|\cdot\|_2$ denotes the operator norm. We use $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ to suppress absolute constants in upper and lower bounds, and $\tilde{\mathcal{O}}(\cdot)$ to hide polylogarithmic factors in d. We write $f \eqsim g$ to indicate that both $f = \mathcal{O}(g)$ and $f = \Omega(g)$ hold. We use $o_d(1)$ to denote terms that vanish as $d \to \infty$, and write $f(d) \ll g(d)$ (respectively, $f(d) \gg g(d)$) if $f \in o(g)$ (respectively, $g \in o(f)$).

2.1 Student-teacher Setting

Teacher Network. We consider the problem of learning a teacher network with a quadratic (second-order Hermite) activation function, defined as:

$$y = f_*(\boldsymbol{x}) + \sigma\epsilon \quad \text{with} \quad f_*(\boldsymbol{x}) \coloneqq \sum_{j=1}^{r} \lambda_j \cdot \tfrac{1}{\sqrt{2}}\big(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle^2 - 1\big) \quad \text{and} \quad \boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d).$$

Here, $\boldsymbol{x} \in \mathbb{R}^d$ is the input, ϵ is zero-mean, unit-variance sub-exponential noise, $\sigma \geq 0$ controls the label-noise level, and r is the teacher network width. The set $\{\boldsymbol{\theta}_i\}_{i=1}^r \subset \mathbb{R}^d$ consists of unknown, orthonormal signal vectors. The contributions of these signal directions are determined by the unknown second-layer coefficients $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$. We assume these coefficients follow a power-law decay $\lambda_i \eqsim i^{-\alpha}$ for $\alpha > 0.5$. We consider the regime where $r \asymp d^\beta$ for $\beta \in (0, 1]$.

Without loss of generality, we assume $\sum_{i=1}^r \lambda_i^2 = 1$ so that the target function has a unit variance, i.e., $\mathbb{E}[f_*(\boldsymbol{x})^2] = 1$. For notational convenience, we collect the teacher directions and signal strengths in a single matrix $\boldsymbol{\Psi}_*$, which allows us to equivalently express the target function as a trace:

$$\boldsymbol{\Psi}_* \coloneqq \sum_{j=1}^{r} \lambda_j \boldsymbol{\theta}_j \boldsymbol{\theta}_j^\top \quad \text{where} \quad f_*(\boldsymbol{x}) = \operatorname{Tr}\left(\boldsymbol{\Psi}_* \tfrac{1}{\sqrt{2}}(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d)\right).$$

This trace representation naturally justifies our orthonormality assumption on the signal vectors $\{\boldsymbol{\theta}_i\}_{i=1}^r$, as it simply corresponds to the spectral decomposition of $\boldsymbol{\Psi}_*$.

Student Network. We learn the target model using a quadratic student network defined as:

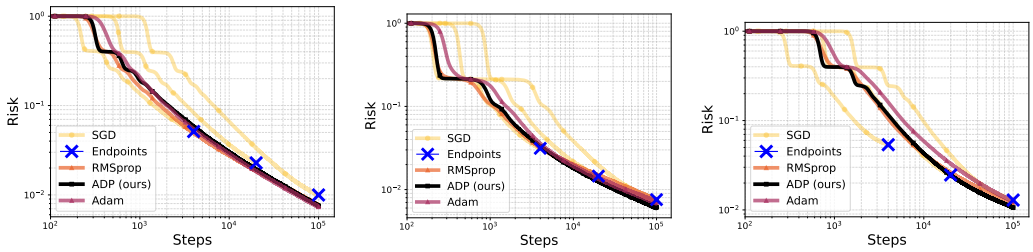
$$\hat{y}(\boldsymbol{x}, \boldsymbol{W}) = \frac{1}{r_s} \sum_{i=1}^{r_s} \tfrac{1}{\sqrt{2}}\big(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle^2 - \|\boldsymbol{w}_i\|_2^2\big), \tag{2.1}$$

where r_s is the width of the student network, and $\{\boldsymbol{w}_i\}_{i=1}^{r_s} \subset \mathbb{R}^d$ denotes the set of trainable weights. We collect these weights as the columns of the matrix $\boldsymbol{W} \in \mathbb{R}^{d \times r_s}$. When clear from context, we omit the explicit dependence on $\boldsymbol{x}$ and simply write $\hat{y}(\boldsymbol{W})$. For analytical convenience, we subtract the norm $\|\boldsymbol{w}_i\|_2^2$ to center the network predictions, i.e., $\mathbb{E}_{\boldsymbol{x}}[\hat{y}(\boldsymbol{x}, \boldsymbol{W})] = 0$. We note that training bias terms with a larger learning rate yields the exact same centering effect in practice (Yu et al., 2026).

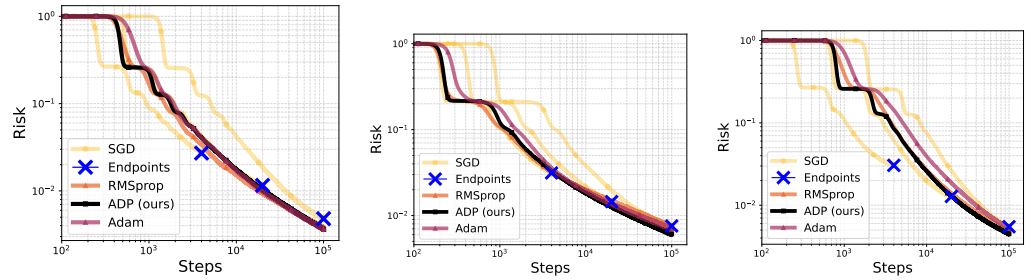
Training objective. The training objective is to minimize the expected squared loss. The population loss admits a closed-form expression given by:

$$\mathcal{P}(\boldsymbol{W}) \coloneqq \mathbb{E}_{(\boldsymbol{x},y)}\left[(y - \hat{y}(\boldsymbol{x}, \boldsymbol{W}))^2\right] = \left\| \frac{1}{r_s} \boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{\Psi}_* \right\|_{\mathrm{F}}^2 + \sigma^2. \tag{2.2}$$

For notational convenience, we denote the excess risk at time t by $\mathcal{R}(t) \coloneqq \mathcal{P}(\boldsymbol{W}_t) - \sigma^2$.



(a) $\phi = \text{quad}$, $\alpha = 1.0$ (b) $\phi = \text{relu}$, $\alpha = 1.0$ (c) $\phi = \text{abs}$, $\alpha = 1.0$



(d) $\phi = \text{quad}$, $\alpha = 1.25$ (e) $\phi = \text{relu}$, $\alpha = 1.25$ (f) $\phi = \text{abs}$, $\alpha = 1.25$

Figure 2: Risk trajectory for activations $\phi \in \{\text{quad}, \text{relu}, \text{abs}\}$ and $\alpha \in \{1.0, 1.25\}$ with $\sigma = 0.5$. The SGD baselines are evaluated across different training durations $T \in \{4 \times 10^3, 2 \times 10^4, 10^5\}$, with the learning rate optimally tuned for each specific horizon , where crosses indicate the final risk of SGD at the end of training. Adam, RMSprop and our algorithm are trained for $T = 10^5$. See Appendix A for more details.

3 Main Results

3.1 Warm Up: Continuous Adaptive Dynamics

To motivate our algorithmic approach, we first analyze an adaptive continuous-time dynamic for minimizing the population objective $\mathcal{P}(\boldsymbol{W})$ defined in (2.2). The dynamics are governed by:

$$\partial_t \boldsymbol{W}_t = -\gamma(t)\nabla\mathcal{P}(\boldsymbol{W}_t), \quad \text{where} \quad \gamma(t) := \frac{0.25 r_s}{\sqrt{\frac{1}{\log d}\int_0^t w_{s,t}\mathcal{P}(\boldsymbol{W}_s)\mathrm{d}s + \epsilon}}, \tag{3.1}$$

where the population gradient is explicitly given by

$$\nabla\mathcal{P}(\boldsymbol{W}_t) = -\frac{4}{r_s}\Big(\boldsymbol{\Psi}_* - \frac{1}{r_s}\boldsymbol{W}_t\boldsymbol{W}_t^\top\Big)\boldsymbol{W}_t.$$

Here, the $w_{s,t}$ term in $\gamma(t)$ denotes a temporal weight function used to discount past information and the additive term ϵ is used to have a well-defined $\gamma(t)$ when t is very small. Since feature recovery occurs on a $\log d$ time scale (see (3.3)), the $(\log d)^{-1}$ prefactor is required to normalize the integral in the denominator. Furthermore, the $0.25 r_s$ constant factor in the learning rate is introduced purely for notational convenience to cleanly cancel the $\frac{4}{r_s}$ coefficient in the gradient. The following theorem characterizes the resulting risk curve and the associated learning rate dynamics.

Theorem 1. *Consider the initialization $\boldsymbol{W}_0 \in \mathbb{R}^{d\times r_s}$ with independent entries $\boldsymbol{W}_{0,ij} \sim \mathcal{N}(0, 1/d)$. Let the weight function scale as $w_{s,t} \simeq (s/t)^q$ for any $q \geq 0 \vee (2\alpha - 2)$ and width $r_s \geq \Omega_\alpha(1)$. Then, the population (3.1) dynamics satisfy the following with probability $1 - o(1/d^2) - \Omega(1/r_s^2)$:*

1. ***Exact risk curve.*** *The risk follows*

$$\mathcal{R}(t\log d) = 1 - \sum_{i=1}^{r_s}\lambda_i^2 \mathbb{1}\{F_*(t) \geq \tfrac{1}{\lambda_i}\} + o_d(1) \tag{3.2}$$

where F_* is a monotone increasing function, satisfying the nonlinear differential equation:

$$\partial_t^2 F_*(t) = -\frac{1}{2}\Big(\mathcal{R}(t\log d) + \sigma^2 + q\epsilon t^{-1}\Big)\big(\partial_t F_*(t)\big)^3 + \frac{q}{2t}\partial_t F_*(t) \tag{3.3}$$

with initial conditions $F_*(0) = 0$ and $\partial_t F_*(0) = \frac{1}{\sqrt{\epsilon}}$.

2. ***Asymptotic risk.*** *The differential equation in (3.3) implies a growth rate $F_*(t) \simeq t^\alpha \wedge \frac{\sqrt{t}}{\sigma}$, which yields the asymptotic bound:*

$$\mathcal{R}(t \log d) \simeq \left(1 + t^{2\alpha} \vee \frac{t}{\sigma^2}\right)^{\frac{1-2\alpha}{2\alpha}} + \sum_{i=r_s+1}^{r} \lambda_i^2.$$

3. ***Adaptive learning rate.*** *Consider the time interval corresponding to the recovery of the j-th teacher direction in (3.2), defined as*

$$\mathcal{I}_i := \{t \geq 0 : \tfrac{1}{\lambda_i} \leq F_*(t) < \tfrac{1}{\lambda_{i+1}}\}. \tag{3.4}$$

For $t \in \mathcal{I}_i$, the adaptive learning rate scales as

$$\gamma(t) \simeq \begin{cases} \lambda_j, & \sigma > 0 \\ \lambda_j^{\frac{1-\alpha}{\alpha}} & \sigma = 0. \end{cases} \tag{3.5}$$

The proof of Theorem 1 is discussed in Section 4. We make the following observations:

- *Modulated emergent behavior.* The exact risk curve in (3.2) exhibits a superposition of emergent learning behaviors similar to prior literature Ren et al. (2025); Ben Arous et al. (2025), but crucially, the time is now modulated by $F_*(t)$. The presence of F_* is a direct consequence of the adaptive learning rate $\gamma(t)$, and it is governed by a differential equation because this adaptivity—i.e., the dependence of $\gamma(t)$ on the objective $\mathcal{P}$—creates a negative feedback mechanism. The growth rate of F_* can be determined by balancing the exponents in (3.3), where the dominant term between the risk $\mathcal{R}(t)$ and the noise variance σ^2 in the first coefficient dictates the overall scaling exponent (see Section 4 for an extended discussion).
- *Effectively noiseless vs. noisy regimes.* The theorem identifies two distinct decay phases in the asymptotic risk. The initial phase exhibits a decay exponent of $1 - 2\alpha$, matching the noiseless ERM setting, while the subsequent phase slows to an exponent of $\frac{1-2\alpha}{2\alpha}$, recovering the standard rate of the noisy ERM setting (see Table 1). In the exact noiseless limit ($\sigma \to 0$), this faster rate dictates the entire trajectory. For any small but finite $\sigma > 0$, the noiseless rate strictly governs the dynamics until a critical time horizon $t \simeq \sigma^{\frac{-2}{2\alpha-1}}$, after which the slower noisy rate dominates. We illustrate these distinct phase transitions in Figure 1(b).
- *Learning-rate scheduling.* Standard SGD with a fixed learning rate requires tuning the step size with the target signal λ_j to maintain the necessary signal-to-noise ratio when recovering the first j teacher directions (Ren et al., 2025). Consequently, an optimal adaptive algorithm must dynamically adjust its learning rate according to the signal strength of the feature currently being recovered. As established in the final item of Theorem 1, our proposed adaptive strategy automatically satisfies this requirement. Specifically, in the noisy setting, it maintains a learning rate $\gamma(t) \simeq \lambda_j$, which recovers the optimal schedule derived in Ren et al. (2025). Furthermore, our analysis identifies a new optimal schedule for the noiseless regime: $\gamma(t) \simeq \lambda_j^{\frac{1-\alpha}{\alpha}}$. As we detail in Section 4 (in the last step), both cases fundamentally correspond to balancing the ratio of the signal to the gradient noise. Finally, numerical simulations in Figure 3 illustrate the relationship between the adaptive learning rate $\gamma(t)$ and the modulated time scale $F_*(t)$. In the noisy setting, this scaling remains remarkably robust across various choices of α and σ. In the noiseless regime, we observe similarly consistent behavior, yielding an increasing or decreasing learning rate depending on whether $\alpha > 1$ or $\alpha < 1$.

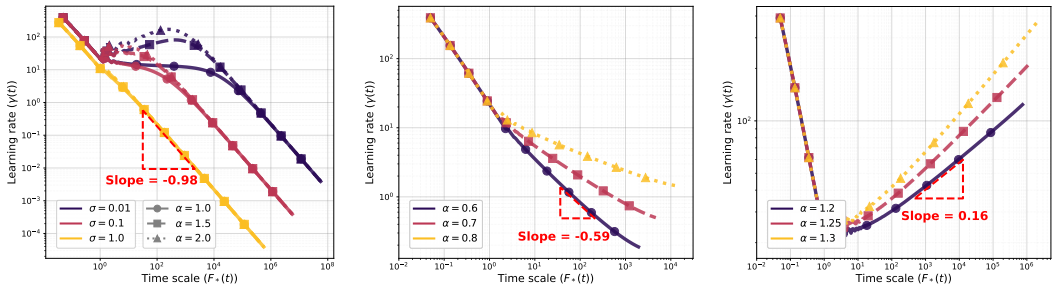
3.2 Discrete Dynamics: Online Adaptive Algorithm

We now analyze the finite-sample, discrete-time counterpart to the continuous adaptive dynamics in (3.1). In particular, we consider an online (one-pass) discretization. Given a fresh sample $(\boldsymbol{x}_{t+1}, y_{t+1})$, we define the instantaneous squared loss:

$$\mathcal{L}_{t+1}(\boldsymbol{W}) := \left(y_{t+1} - \hat{y}(\boldsymbol{x}_{t+1}; \boldsymbol{W})\right)^2,$$

where $\hat{y}(\boldsymbol{x}_{t+1}; \boldsymbol{W})$ denotes the network output defined in (2.1). The discrete update rules are given by:

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \gamma_t \nabla \mathcal{L}_{t+1}(\boldsymbol{W}_t), \quad \text{where} \quad \gamma_t = \frac{\eta \cdot 0.25 r_s}{\sqrt{\frac{\nu_t}{\log d}} + \epsilon}, \tag{ADP}$$



(a) (b) (c)

Figure 3: The adaptive learning rate $\gamma(t)$ plotted against the modulated time scale $F_*(t)$ for (a) the noisy setting, (b) the noiseless setting with $\alpha < 1$, and (c) the noiseless setting with $\alpha > 1$. In (a), we use $\alpha \in \{1.0, 1.5, 2.0\}$ and noise levels $\sigma \in \{0.01, 0.1, 1\}$. The slope rapidly stabilizes to 1, which verifies the learning rate schedule derived in Theorem 1. In (b) and (c), the slope is consistent with the $\frac{1-\alpha}{\alpha}$ prediction of Theorem 1, where $\alpha < 1$ yields a decaying learning rate, while $\alpha > 1$ results in an increasing learning rate.

and the loss moving average is updated via

$$\nu_{t+1} = \eta \mathcal{L}_{t+1}(\boldsymbol{W}_t) + (1 - \beta_t)\nu_t. \tag{3.6}$$

In this algorithm, η represents the base learning rate, and ν_t serves as the discrete analogue of the population loss integral in (3.1). The following theorem characterizes the risk curve and the learning rate for the discrete-time.

Theorem 2. *Let $q \in \mathbb{N}$ satisfy the conditions in Theorem 1, and assume $r_s \gg \log^2 d$. We set the base learning rate η, adaptive parameter β_t, and stability constant ϵ as:*

$$\eta \simeq \frac{1}{d\,\mathrm{polylog}(d)}, \quad \beta_t = \frac{q-1}{t+q}, \quad \textit{and} \quad \epsilon \simeq O(1).$$

Then, the discrete (ADP) dynamics satisfy the following with probability $1 - o(1/d^2) - \Omega(1/r_s^2)$:

1. ***Scaling laws in discrete time.*** *We have*

$$\mathcal{R}(\eta t \log d) \simeq \left(1 + (\eta t)^{2\alpha} \vee \frac{\eta t}{\sigma^2}\right)^{\frac{1-2\alpha}{2\alpha}} + \sum_{i=r_s+1}^{r} \lambda_i^2 + o_d(1).$$

2. ***Learning rate.*** *Given that $\eta t \in \mathcal{I}_i$, where $\mathcal{I}_i$ is defined in (3.4), $\frac{\gamma_t}{\eta}$ satisfy the scheduling in (3.5).*

The complete proof of Theorem 2 is provided in Appendix C.1. We make the following remarks:

- We emphasize that (ADP) represents a direct discretization of the continuous adaptive dynamics with a fixed time step η. Because this same step size is used to integrate ν_t, η naturally appears as the coefficient of the instantaneous loss L in (3.6). The parameter β_t simulates the temporal weight function $w_{s,t}$ via the product expansion $\prod_{j=s}^{t}(1-\beta_j) \simeq (s/t)^q$, allowing the discrete sum ν_t to approximate the continuous integral over the population risk $P = \mathbb{E}[L]$:

$$\nu_t = \eta \sum_{s=1}^{t} \prod_{j=s}^{t} (1-\beta_j) L(\boldsymbol{W}_s) \approx \int_0^{\eta t} w_{s,\eta t} P(\boldsymbol{W}_s)\, ds$$

- Following the first item of Theorem 2, we establish a (nearly) information-theoretically optimal sample complexity of $N \asymp d\,\mathrm{polylog}(d)$, recovering the optimal scaling exponents for both the noisy ($\sigma > 0$) and noiseless ($\sigma \to 0$) regimes. Notably, prior work employing a similar learning-rate schedule achieved strictly suboptimal rates in their respective setting (Ren et al., 2025).
- Finally, the second item in Theorem 2 suggests that γ_t modulates the baseline learning rate η analogously to the continuous-time schedule in (3.5), which ensures that the effective step size dynamically tracks the signal strength of the active feature. Just as in the continuous population dynamics, this automatic alignment maintains the optimal signal-to-noise ratio required for sequential feature recovery without necessitating manual step-size tuning.

4 Overview of Proof Techniques

For the proof sketch, we focus on the population dynamics in (3.1), as the discrete-time dynamics closely follow this continuous-time limit. This correspondence can be rigorously established using the proof techniques from Ben Arous et al. (2025). In this section, we primarily highlight how the analysis diverges for the adaptive learning rate case. Furthermore, because our argument relies on a time-rescaling technique, we denote the weight matrix as $\boldsymbol{W}(t)$ rather than $\boldsymbol{W}_t$ for notational clarity.

Step 1: Reduction to a sufficient statistic. First, we observe that the risk in (2.2) depends on the weights entirely through the normalized Gram matrix $\boldsymbol{G}(t) \coloneqq \frac{1}{r_s}\boldsymbol{W}(t)\boldsymbol{W}(t)^\top$. Applying the chain rule to the gradient flow, the evolution of $\boldsymbol{G}(t)$ is given by a Riccati-style differential equation:

$$\partial_t \boldsymbol{G}(t) = \frac{4\gamma(t)}{r_s}\left(\boldsymbol{\Psi}_*\boldsymbol{G}(t) + \boldsymbol{G}(t)\boldsymbol{\Psi}_* - 2\boldsymbol{G}(t)^2\right). \tag{4.1}$$

Crucially, the right-hand side is modulated purely by the scalar function $t \mapsto \frac{4\gamma(t)}{r_s}$.

Step 2: Time-scale change and the high-dimensional limit. Because the learning rate dictates the speed of the flow in (4.1), we can decouple the trajectory of the system from its velocity by warping time. We define an effective time scale $F(t) = \int_0^t \frac{4\gamma(s)}{r_s}\, ds$. To capture the high-dimensional behavior ($d \to \infty$), we introduce logarithmically rescaled quantities, denoted by the subscript $*$, such as $F_*(t) = F(t\log d)/\log d$. By evaluating the system in this warped time, $\boldsymbol{H}(t) = \boldsymbol{G}(F^{-1}(t))$, the dynamics shed the $\gamma(t)$ dependency entirely:

$$\partial_t \boldsymbol{H}(t) = \boldsymbol{\Psi}_*\boldsymbol{H}(t) + \boldsymbol{H}(t)\boldsymbol{\Psi}_* - 2\boldsymbol{H}(t)^2.$$

This uncoupled system allows us to cleanly establish (via (Ben Arous et al., 2025, Theorem 1)) that the risk evolution under $\boldsymbol{H}$ follows a sequential learning phenomenon, which reads

$$\|\boldsymbol{H}(t\log d) - \boldsymbol{\Psi}_*\|_F^2 = 1 - \sum_{i=1}^{r_s} \lambda_i^2 \mathbb{1}\{t \geq \tfrac{1}{\lambda_i}\} + o_d(1).$$

By mapping $\boldsymbol{H}$ back through $F_*(t)$, we recover the true temporal risk trajectory $\mathcal{R}(t)$.

Step 3: Deriving the nonlinear ODE for F_*. To characterize the trajectory of the effective time scale $F_*(t)$, we must analyze the integral defining our adaptive learning rate schedule. We define $V(t)$ as the accumulated, polynomially-discounted history of the risk and noise:

$$V(t) = \int_0^t \left(\frac{s}{t}\right)^q \left(\mathcal{R}(s\log d) + \sigma^2\right) ds.$$

Differentiating $V(t)$ with respect to t using the Leibniz integral rule produces two competing terms: a decay term emerging from the $\frac{1}{t^q}$ factor, and a source term from the upper limit of the integral:

$$\partial_t V(t) = -\frac{q}{t}V(t) + \mathcal{R}(t\log d) + \sigma^2.$$

By definition, the effective time scale $F_*(t)$ is the integral of the learning rate, implying its velocity is directly tied to $V(t)$:

$$\partial_t F_*(t) = \gamma_*(t) = \frac{1}{\sqrt{V(t)+\epsilon}}.$$

To eliminate the integral $V(t)$ and form a closed-form differential equation for F_*, we take the second derivative of $F_*(t)$. Applying the chain rule, we obtain:

$$\begin{aligned}\partial_t^2 F_*(t) &= -\frac{1}{2}\big(V(t)+\epsilon\big)^{-\frac{3}{2}}\partial_t V(t) \\ &= -\frac{1}{2}\big(V(t)+\epsilon\big)^{-\frac{3}{2}}\left(-\frac{q}{t}V(t) + \mathcal{R}(t\log d) + \sigma^2\right).\end{aligned}$$

To express this entirely in terms of F_*, we use the algebraic identity $\big(V(t)+\epsilon\big)^{-\frac{3}{2}} = \big(\partial_t F_*(t)\big)^3$ and substitute $V(t) = \big(\partial_t F_*(t)\big)^{-2} - \epsilon$. Rearranging the terms perfectly recovers the fundamental ODE formulated in (3.3):

$$\partial_t^2 F_*(t) = -\frac{1}{2}\left(\mathcal{R}(t\log d) + \sigma^2 + \frac{q}{t}\epsilon\right)\big(\partial_t F_*(t)\big)^3 + \frac{q}{2t}\partial_t F_*(t).$$

Step 4: Asymptotic growth rate via dominant balance. Because the derived ODE is highly nonlinear, we extract the asymptotic behavior of $F_*(t)$ using the method of dominant balance. We assume a power-law ansatz of the form $F_*(t) \simeq Ct^u$ for $u > 0$, which implies $\partial_t F_*(t) \simeq Cut^{u-1}$ and $\partial_t^2 F_*(t) \simeq Cu(u-1)t^{u-2}$.

The dynamics fundamentally split into two distinct regimes depending on the presence of noise:

- **Noise-Dominated Regime ($\sigma > 0$):** The constant noise variance σ^2 eventually dominates the decaying risk $\mathcal{R}$ and the $O(1/t)$ terms. The ODE simplifies to balancing the second derivative against the cubed first derivative scaled by noise:

$$\partial_t^2 F_*(t) \simeq -\frac{1}{2}\sigma^2 \big(\partial_t F_*(t)\big)^3.$$

  Substituting our ansatz yields $t^{u-2} \propto t^{3u-3}$, which requires $u - 2 = 3u - 3 \implies u = \frac{1}{2}$. Solving for the constant gives $C \simeq \frac{1}{\sigma}$, meaning $F_*(t) \simeq \frac{\sqrt{t}}{\sigma}$.

- **Signal-Dominated Regime ($\sigma \to 0$):** In the noiseless limit, the risk term dictates the dynamics. From prior spectral bounds, we know the risk scales with the effective time as $\mathcal{R}(t \log d) \simeq (F_*(t))^{\frac{1-2\alpha}{\alpha}}$. Substituting this dominant term back into the ODE yields:

$$\partial_t^2 F_*(t) \simeq -\frac{1}{2}(F_*(t))^{\frac{1-2\alpha}{\alpha}} \big(\partial_t F_*(t)\big)^3.$$

  Substituting $F_*(t) \simeq Ct^u$ here requires matching the powers of t. Equating the exponents gives $u - 2 = u\left(\frac{1-2\alpha}{\alpha}\right) + 3(u-1)$, which recovers $u = \alpha$, with the constant $C \simeq 1$, giving $F_*(t) \simeq t^\alpha$.

Step 5: Optimal learning rate scheduling. Finally, we bridge this continuous limit back to the discrete stochastic updates. In the stochastic regime, the optimal learning rate must balance the target signal strength λ_k against the gradient noise variance, which scales as $\mathbb{E}\big[\|\nabla L_{t+1}(\boldsymbol{W}_t)\|_F^2\big] \simeq d\mathcal{P}(\boldsymbol{G}_t)$. Enforcing this signal-to-noise balance dictates an effective learning rate of $\gamma_* \propto \lambda_k/(d\mathcal{P}(\boldsymbol{G}_t))$.

To see how this shapes the schedule during the learning of the k-th feature (i.e., for $t \in \mathcal{I}_k$ as defined in (3.4)), we apply the risk scaling $\mathcal{P}(\boldsymbol{G}_t) \simeq k^{1-2\alpha} + \sigma^2$ established in (3.2). Substituting this variance bound back into γ_* perfectly recovers the λ_k-dependent scaling rules, and theoretically justify the adaptive scheduling proposed in (3.5).

5 Conclusion

Existing scaling law analyses for nonlinear networks predominantly rely on a constant learning rate that must be manually tuned to the target signal strength, whereas practitioners overwhelmingly favor adaptive algorithms. In this work, we bridged this gap by analyzing SGD with adaptive learning rates for learning quadratic networks. We established precise scaling laws that characterize both the population gradient flow and its empirical, discrete-time approximation. Furthermore, we characterized the optimal learning-rate scheduling and demonstrated how the proposed algorithm adaptively recovers it. Finally, we empirically verified our theoretical results and showed that Adam and RMSprop exhibit highly similar learning curves.

Limitations and future directions. (i) For mathematical convenience, our theoretical focus was restricted to two-layer networks with quadratic activations. While our experiments show that similar scaling behaviors hold for general activation functions, extending our rigorous asymptotic analysis to arbitrary nonlinearities remains a natural direction for future work. (ii) For the discrete-time analysis, our proof requires an additional $\mathrm{polylog}(d)$ factor to rigorously match the continuous-time dynamics. However, our experiments closely matches with end-training risks for tuned fixed-rate SGD, which often operates at the critical $O(1/d)$ learning rate limit. Extending our theoretical guarantees to this critical learning-rate regime is an interesting open problem. (iii) Finally, making the connection between our proposed adaptive scheme and widely used optimizers, such as Adam and RMSprop, mathematically rigorous is another interesting direction for future research.

AI Assistance Disclosure

During the preparation of this manuscript, the authors utilized large language models (LLMs) exclusively for two purposes: to polish the prose for clarity and grammar, and as an auxiliary tool to double-check and verify manual mathematical derivations. No novel mathematical results, theorems, or core intellectual contributions were generated by AI. The authors have thoroughly reviewed all AI-assisted text and mathematical verifications, and take full responsibility for the final content and the correctness of the proofs presented in this work.

References

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pp. 4782–4887. PMLR, 2022.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2552–2623. PMLR, 2023.

Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 487–492. IEEE, 2017.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HlIAoCHDWW.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27), 2024. ISSN 1091-6490. doi: 10.1073/pnas.2311878121. URL http://dx.doi.org/10.1073/pnas.2311878121.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35: 25349–25362, 2022.

Gérard Ben Arous, Cédric Gerbelot, and Vanessa Piccolo. High-dimensional optimization for multi-spiked tensor pca. *arXiv preprint arXiv:2408.06401*, 2024.

Gérard Ben Arous, Murat A. Erdogdu, Nuri Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: Sgd dynamics and scaling laws, 2025. URL https://arxiv.org/abs/2508.03688.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, pp. 1–84, 2024.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning Gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Advances in Neural Information Processing Systems*, 28, 2015.

Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models. *arXiv preprint arXiv:2308.08977*, 2023.

Elizabeth Collins-Woodfin, Inbar Seroussi, Begoña García Malaxechebarría, Andrew W. Mackenzie, et al. The high line: Exact risk and learning rate curves of stochastic adaptive learning rate algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 6500–6548, 2024.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36, 2024.

Leonardo Defilippis, Yizhou Xu, Julius Girardin, Emanuele Troiani, Vittorio Erba, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime, 2025. URL https://arxiv.org/abs/2509.24882.

Leonardo Defilippis, Florent Krzakala, Bruno Loureiro, and Antoine Maillard. A noise sensitivity exponent controls large statistical-to-computational gaps in single- and multi-index models, 2026. URL https://arxiv.org/abs/2603.17896.

Alexandre D'efossez, Léon Bottou, Francis R. Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Trans. Mach. Learn. Res.*, 2022, 2020.

John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011.

Damien Ferbach, Katie Everett, Gauthier Gidel, Elliot Paquette, and Courtney Paquette. Dimension-adapted momentum outscales sgd, 2025. URL https://arxiv.org/abs/2505.16098.

Julius Girardin, Emanuele Troiani, Yizhou Xu, Vittorio Erba, Florent Krzakala, and Lenka Zdeborová. How width and data shape generalization scaling laws in quadratic neural networks. *ArXiv*, abs/2606.28242, 2026.

Margalit Glasgow, Denny Wu, and Joan Bruna. Propagation of chaos in one-hidden-layer neural networks beyond logarithmic time. *arXiv preprint arXiv:2504.13110*, 2025.

Athanasios Glentis, Dawei Li, Chung-Yiu Yau, and Mingyi Hong. Revisiting the adam-sgd gap in llm pre-training: The role of large effective learning rates. *ArXiv*, abs/2605.17787, 2026.

Aukosh Jagannath, Taj Jones-McCormick, and Varnan Sarangian. High-dimensional limit theorems for sgd: Momentum and adaptive step-sizes. *arXiv preprint arXiv:2511.03952*, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bingrui Li, Wei Huang, Andi Han, Zhanpeng Zhou, Taiji Suzuki, Jun Zhu, and Jianfei Chen. On the optimization and generalization of two-layer transformers with sign gradient descent. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=97rOQDPmk2.

Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *Annals of Statistics*, 17:1009–1052, 1989. URL https://api.semanticscholar.org/CorpusID:122222812.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 983–992. PMLR, 16–18 Apr 2019.

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *arXiv preprint arXiv:2406.08466*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

Arvind Mahankali, Haochen Zhang, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond ntk with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. *Advances in Neural Information Processing Systems*, 36, 2023.

Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2024.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. In *The Eleventh International Conference on Learning Representations*, 2022.

Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A. Erdogdu. Gradient-based feature learning under structured data. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

Yoonsoo Nam, Nayara Fonseca, Seok Hyeong Lee, and Ard Louis. An exactly solvable model for emergence and scaling laws. *arXiv preprint arXiv:2404.17563*, 2024.

Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1): 69–84, 1985.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations. In *Conference on Learning Theory*. PMLR, 2024.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

Yunwei Ren and Jason D Lee. Learning orthogonal multi-index models: A fine-grained information exponent analysis. *arXiv preprint arXiv:2410.09678*, 2024.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D Lee. Emergence and scaling laws in sgd learning of shallow neural networks. *arXiv preprint arXiv:2504.19983*, 2025.

Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *Journal of Machine Learning Research*, 24(58): 1–47, 2023.

Xuan Tang, Han Zhang, Yuan Cao, and Difan Zou. Understanding the generalization of stochastic gradient adam in learning neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://openreview.net/forum?id=ETgPUJfQE1.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, 2012.

Rachel A. Ward, Xiaoxia Wu, and Léon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. In *International Conference on Machine Learning*, 2018.

Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:3273477.

Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞-norm constrained optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54488–54510. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xie24e.html.

Qingyuan Yu, Nuri Mert Vural, Xin T. Tong, and Murat A. Erdogdu. Feature learning in high-dimensions under structured covariance: Scaling laws in quadratic networks. In *ICML Workshop on High-dimensional Learning Dynamics (HiLD 2026)*, 2026. URL https://openreview.net/pdf?id=Up11AQvPhV.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21774–21798, 2024.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 15383–15393, 2020.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham M. Kakade. Deconstructing what makes a good optimizer for language models. *ArXiv*, abs/2407.07972, 2024.

CONTENTS

1 Introduction **1**

1.1 Our Contributions 2

1.2 Additional Related Works 3

2 Problem Setting **4**

2.1 Student-teacher Setting 4

3 Main Results **5**

3.1 Warm Up: Continuous Adaptive Dynamics 5

3.2 Discrete Dynamics: Online Adaptive Algorithm 6

4 Overview of Proof Techniques **8**

5 Conclusion **9**

A Experimental Details **15**

B Proofs for Main Results **15**

C Proof for Theorem 2 **15**

C.1 Preliminaries 15

D Supporting Statements for Theorem 3 **17**

D.1 Definitions and Bounds 18

D.2 Bounding System Decomposition 19

D.2.1 Upper Bounds 19

D.2.2 Lower Bound 21

D.3 Analysis of Deterministic Bounding Systems 22

D.3.1 Upper Deterministic System 22

D.3.2 Lower Deterministic System 23

D.4 Noise Characterization 25

D.4.1 Variance Bounds 26

D.4.2 Characterization of Learning Rate 33

E Auxillary Propositions **35**

E.1 Matrix Identities 35

E.2 Some moment bounds and concentration inequalities 35

E.3 Miscellaneous 37

A Experimental Details

Experimental Setup. In Figure 1(a) and Figure 2, we plot the mean squared error (MSE) loss for SGD with a fixed learning rate, our proposed adaptive learning rate in (ADP), RMSprop, and Adam in the online setting (sampling without replacement). We set the dimensions to $d = r_s = 1024$, the teacher rank to $r = 512$, and the batch size to 1024. For completeness, the standard update rules for RMSprop and Adam are given as:

$$\textbf{RMSprop:} \quad \begin{aligned} \boldsymbol{v}_t &= \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2, \\ \boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \frac{\eta}{\sqrt{\boldsymbol{v}_t} + \epsilon} \odot \boldsymbol{g}_t \end{aligned}$$

$$\textbf{Adam:} \quad \begin{aligned} \boldsymbol{m}_t &= \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t, \\ \boldsymbol{v}_t &= \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2, \\ \boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \frac{\eta}{\sqrt{\hat{\boldsymbol{v}}_t} + \epsilon} \odot \hat{\boldsymbol{m}}_t \quad \text{where} \quad \hat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1-\beta_1^t}, \;\; \hat{\boldsymbol{v}}_t = \frac{\boldsymbol{v}_t}{1-\beta_2^t}. \end{aligned}$$

Hyperparameters. We evaluated the SGD baselines across three different time horizons $T \in \{4 \times 10^3, 2 \times 10^4, 10^5\}$, while the adaptive algorithms were trained for a single horizon of $T = 10^5$. For both SGD and our adaptive algorithm (ADP), we swept the hyperparameters for each T and evaluated the models using the optimal learning rate. For RMSprop and Adam, we used the same base learning rate η and second-order moment parameter β_2 as in (ADP), as this configuration provided the best empirical results. For Adam specifically, we set the first-order momentum parameter to $\beta_1 = 0.99$, which provided the best performance.

General activations. For the experiments in Figure 2 involving general activations, we assume both the teacher and student networks employ the same activation function, which we normalize to have zero mean and unit ℓ_2 norm. The student network uses a 2-homogeneous parameterization:

$$\hat{y}(\boldsymbol{W}) = \frac{1}{r_s}\sum_{i=1}^{r_s} \|\boldsymbol{w}_i\|_2^2 \cdot \phi\left(\left\langle \tfrac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|_2}, \boldsymbol{x}\right\rangle\right) \quad \text{where} \quad \phi(x) = \begin{cases} \frac{x^2-1}{\sqrt{2}}, & \text{quad} \\ \frac{x \vee 0 - 1/\sqrt{2\pi}}{\sqrt{1/2 - 1/(2\pi)}}, & \text{relu} \\ \frac{|x| - \sqrt{2/\pi}}{\sqrt{1-2/\pi}}, & \text{abs.} \end{cases}$$

B Proofs for Main Results

Proof organization. As we detailed the proof of Theorem 1 in Section 4, we do not provide a new proof here. Appendix C.1 provide a precise statement for Theorem 2 (see Theorem 3). The subsequent sections provide the supporting arguments for Theorem 3.

Additional Notation and Definitions. We denote the symmetric part of a square matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ as $\mathrm{sym}(\boldsymbol{M}) := \frac{1}{2}(\boldsymbol{M} + \boldsymbol{M}^T)$. For symmetric matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$, we write $\boldsymbol{A} \prec \boldsymbol{B}$ or $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive definite or positive semidefinite, respectively, and we use $\|\cdot\|_F$ for the Frobenius norm.

C Proof for Theorem 2

C.1 Preliminaries

Let $\mathrm{He}_2(\boldsymbol{x}) = \frac{1}{\sqrt{2}}(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d)$. We have

$$y_{k+1} = f_*(\boldsymbol{x}_{k+1}) + \sigma\epsilon_{k+1} \;\; \text{where} \;\; f_*(\boldsymbol{x}) = \mathrm{Tr}\left(\boldsymbol{\Psi}_* \mathrm{He}_2(\boldsymbol{x})\right)$$

where ϵ is a unit norm subexponential independent noise and $\boldsymbol{\Psi}_* \in \mathbb{R}^{d\times d}$ has a positive semi-definite matrix with unit Frobenius norm.

We define

- Width-r_s student network:

$$\hat{y}_{k+1} = \mathrm{Tr}\left(\tfrac{1}{r_s}\boldsymbol{W}_k\boldsymbol{W}_k^\top \mathrm{He}_2(\boldsymbol{x}_{k+1})\right), \quad \text{where} \quad \boldsymbol{W}_k \in \mathbb{R}^{d\times r_s}.$$

- Filtration $\mathcal{F}_k \coloneqq \sigma\big(\boldsymbol{W}_k, (\boldsymbol{x}_i, y_i)_{i\le k}\big)$ and conditional expectation $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_k]$.
- We use the loss $L(\boldsymbol{W}_k) \coloneqq (y_{k+1} - \hat{y}_{k+1})^2$.

By using the hyperparameters in Theorem 2, (ADP) reads:

$$\boldsymbol{W}_{k+1} = \boldsymbol{W}_k + \underbrace{\frac{\eta}{\sqrt{\frac{\nu_k}{\log d} + \epsilon}}}_{\coloneqq \gamma_k} (y_{k+1} - \hat{y}_{k+1})\mathrm{He}_2(\boldsymbol{x}_{k+1})\boldsymbol{W}_k$$

$$\nu_{k+1} = \frac{1}{d\rho_d} L(\boldsymbol{W}_k) + (1-\beta_k)\nu_k.$$

We also define

$$F(k) = \sum_{s=0}^{k-1} \gamma_s \quad \text{and} \quad \overline{\gamma}_k \coloneqq \frac{1}{\sqrt{\frac{\nu_k}{\log d} + \epsilon}}.$$

The following theorem formalizes Theorem 2:

Theorem 3. *Let $q \ge 2 \vee (2\alpha - 2)$. We consider $r_s \gg \log^2 d$ and*

$$\eta = \frac{1}{d\rho_d} \quad \textit{and} \quad \rho_d \ll \log^{5.5} d \quad \textit{and} \quad \beta_k = \frac{\lceil q \rceil - 1}{k + \lceil q \rceil}$$

With probability $1 - o_d(1/d^2) - \Omega(1/r_s^2)$, the following holds:

- *The risk curve satisfies*

$$R(kd\rho_d \log d) \eqsim \left(\left(1+k\right)^{1-2\alpha} + \left(1 + \frac{k}{\sigma^2}\right)^{\frac{1-2\alpha}{2\alpha}}\right) \vee \sum_{j=r_s+1}^{d} \lambda_j^2$$

  for $k \le \rho_d^c$ for a sufficiently small $c \in O(1)$.

- *Consider the stopping time*

$$\mathcal{T}_i \coloneqq \inf\{k \in \mathbb{N} : \tfrac{1}{\lambda_i} > F(k)\}.$$

  For $k \in [\mathcal{T}_i, \mathcal{T}_{i+1})$, $\overline{\gamma}_k$ satisfies the scheduling provided in (3.5)

Proof. By using the upper and lower bounding systems in (D.4) and (D.5), we can show that

$$R(kd\rho_d \log d) \eqsim \left(1 + \frac{F(kd\rho_d \log d)}{\log d}\right)^{\frac{1-2\alpha}{\alpha}} \wedge \sum_{j=r_s+1}^{d} \lambda_j^2.$$

We will use asymptotic characterization of F in (D.8), which reads

$$\frac{F(\overline{k}d\rho_d \log d)}{\log d} \eqsim \left((1+\overline{k})^\alpha - 1\right) \wedge \sigma\sqrt{\overline{k}} \tag{C.1}$$

we prove the first statement. For the learning rate, by Bernstein bound, we can show that with high probability for $k = \overline{k}d\rho_d \log d$,

$$\sqrt{\frac{\nu_k}{\log_d} + \epsilon} \eqsim \sqrt{\frac{1}{d\rho_d \log d}\sum_{s=0}^{k-1}\left(\frac{s}{k}\right)^q \mathcal{P}(\boldsymbol{G}_s) + \epsilon} \eqsim (1+\overline{k})^{1-\alpha} + \sigma\sqrt{\overline{k}} \tag{C.2}$$

where we used the characterization in the first item for the last equality. By relating (C.1) with (C.2), we have the second statement. □

D SUPPORTING STATEMENTS FOR THEOREM 3

In this part, we will prove the statement we used in the proof of Theorem 3. We have

$$\begin{aligned}\boldsymbol{W}_{k+1} &= \boldsymbol{W}_k + \frac{\overline{\gamma}_k}{d\rho_d}\big(y_{k+1} - \hat{y}_{k+1}\big)\mathrm{He}_2(\boldsymbol{x}_{k+1})\boldsymbol{W}_k \\ &= \boldsymbol{W}_k + \frac{\overline{\gamma}_k}{d\rho_d}\big(\boldsymbol{\Psi}_* - \frac{1}{r_s}\boldsymbol{W}_k\boldsymbol{W}_k^\top\big)\boldsymbol{W}_k + \frac{\overline{\gamma}_k}{d\rho_d}\Big(\big(y_{k+1} - \hat{y}_{k+1}\big)\mathrm{He}_2(\boldsymbol{x}_{k+1}) \\ &\quad - \mathbb{E}_k[\big(y_{k+1} - \hat{y}_{k+1}\big)\mathrm{He}_2(\boldsymbol{x}_{k+1})]\Big)\boldsymbol{W}_k\end{aligned}$$

We define $\boldsymbol{G}_k \coloneqq \frac{1}{r_s}\boldsymbol{W}_k\boldsymbol{W}_k^\top$ and

$$\mathsf{X}_{k+1} \coloneqq \big(y_{k+1} - \hat{y}_{k+1}\big)\mathrm{He}_2(\boldsymbol{x}_{k+1}) - \mathbb{E}[\big(y_{k+1} - \hat{y}_{k+1}\big)\mathrm{He}_2(\boldsymbol{x}_{k+1})].$$

We have

$$\begin{aligned}\boldsymbol{G}_{k+1} &= \boldsymbol{G}_k + \frac{\overline{\gamma}_k}{d\rho_d}(\boldsymbol{\Psi}_*\boldsymbol{G}_k + \boldsymbol{G}_k\boldsymbol{\Psi}_* - 2\boldsymbol{G}_k^2) + \frac{\overline{\gamma}_k^2}{d^2\rho_d^2}(\boldsymbol{\Psi}_* - \boldsymbol{G}_k)\boldsymbol{G}_k(\boldsymbol{\Psi}_* - \boldsymbol{G}_k) \\ &\quad + \frac{\overline{\gamma}_k}{d\rho_d}\big(\mathsf{X}_{k+1}\boldsymbol{G}_k + \boldsymbol{G}_k\mathsf{X}_{k+1}\big) + \frac{\overline{\gamma}_k^2}{d^2\rho_d^2}\big(\mathsf{X}_{k+1}\boldsymbol{G}_k(\boldsymbol{\Psi}_* - \boldsymbol{G}_k) + (\boldsymbol{\Psi}_* - \boldsymbol{G}_k)\boldsymbol{G}_k\mathsf{X}_{k+1}\big) \\ &\quad + \frac{\overline{\gamma}_k^2}{d^2\rho_d^2}\mathsf{X}_{k+1}\boldsymbol{G}_k\mathsf{X}_{k+1}\end{aligned}$$

We start with the following proposition:

Proposition 1. *Let $\Delta_k \coloneqq \boldsymbol{\Psi}_* - \boldsymbol{G}_k$. We have*

$$\begin{aligned}\mathbb{E}_k\big[\mathsf{X}_{k+1}\boldsymbol{G}_k\mathsf{X}_{k+1}\big] &= 3\Delta_k\boldsymbol{G}_k\Delta_k + 2\Delta_k^2\boldsymbol{G}_k + 2\boldsymbol{G}_k\Delta_k^2 + 2\,\mathrm{Tr}(\boldsymbol{G}_k)\Delta_k^2 + 2\,\mathrm{Tr}(\Delta_k\boldsymbol{G}_k)\Delta_k \\ &\quad + \tfrac{1}{2}\,\mathrm{Tr}(\Delta_k^2)\boldsymbol{G}_k + \big[2\,\mathrm{Tr}(\Delta_k^2\boldsymbol{G}_k) + \tfrac{1}{2}\,\mathrm{Tr}(\Delta_k^2)\,\mathrm{Tr}(\boldsymbol{G}_k)\big]\,\boldsymbol{I}_d \\ &\quad + \tfrac{\sigma^2}{2}(\boldsymbol{G}_k + \mathrm{Tr}(\boldsymbol{G}_k)\boldsymbol{I}_d)\end{aligned}$$

We further define

- $\gamma_k \coloneqq \frac{\overline{\gamma}_k}{d\rho_d}$ and $r_u \coloneqq \log^{\frac{1}{\alpha}} d$
- $\mathcal{R}(\boldsymbol{G}_k) = \mathrm{Tr}(\Delta_k^2)$ and $\mathcal{P}(\boldsymbol{G}_k) \coloneqq (\mathrm{Tr}(\Delta_k^2) + \sigma^2)$
- $\mathsf{R}_{k+1}(\boldsymbol{G}_k) \coloneqq \mathsf{X}_{k+1}\boldsymbol{G}_k\mathsf{X}_{k+1}$
- $\boldsymbol{S}_k \coloneqq \mathbb{E}_k[\mathsf{R}_{k+1}(\boldsymbol{G}_k)] + (\boldsymbol{\Psi}_* - \boldsymbol{G}_k)\boldsymbol{G}_k(\boldsymbol{\Psi}_* - \boldsymbol{G}_k)$
- $\boldsymbol{M}_{k+1} \coloneqq (\boldsymbol{G}_k\mathsf{X}_{k+1} + \mathsf{X}_{k+1}\boldsymbol{G}_k) + \gamma_k\big((\boldsymbol{\Psi}_* - \boldsymbol{G}_k)\boldsymbol{G}_k\mathsf{X}_{k+1} + \mathsf{X}_{k+1}\boldsymbol{G}_k(\boldsymbol{\Psi}_* - \boldsymbol{G}_k)\big) + \gamma_k\big(\mathsf{R}_{k+1}(\boldsymbol{G}_k) - \mathbb{E}_k[\mathsf{R}_{k+1}(\boldsymbol{G}_k)]\big)$

We have

$$\boldsymbol{G}_{k+1} = \boldsymbol{G}_k + \gamma_k\big(\boldsymbol{\Psi}_*\boldsymbol{G}_k + \boldsymbol{G}_k\boldsymbol{\Psi}_* - 2\boldsymbol{G}_k^2\big) + \gamma_k\boldsymbol{S}_k + \gamma_k\boldsymbol{M}_{k+1}.$$

We define the block matrix systems:

$$\boldsymbol{G}_k = \begin{bmatrix}\boldsymbol{G}_{k,11} & \boldsymbol{G}_{k,12}\\ \boldsymbol{G}_{k,12}^\top & \boldsymbol{G}_{k,22}\end{bmatrix},\quad \boldsymbol{\Psi}_* = \begin{bmatrix}\boldsymbol{\Psi}_{11} & 0\\ 0 & \boldsymbol{\Psi}_{22}\end{bmatrix},\quad \boldsymbol{S}_k = \begin{bmatrix}\boldsymbol{S}_{k,11} & \boldsymbol{S}_{k,12}\\ \boldsymbol{S}_{k,12}^\top & \boldsymbol{S}_{k,22}\end{bmatrix},$$

$$\boldsymbol{M}_k = \begin{bmatrix}\boldsymbol{M}_{k,11} & \boldsymbol{M}_{k,12}\\ \boldsymbol{M}_{k,12}^\top & \boldsymbol{M}_{k,22}\end{bmatrix},\quad \mathsf{X}_k = \begin{bmatrix}\mathsf{X}_{k,11} & \mathsf{X}_{k,12}\\ \mathsf{X}_{k,12}^\top & \mathsf{X}_{k,22}\end{bmatrix},$$

where we have

- $\boldsymbol{G}_{k,11}, \boldsymbol{\Psi}_{11}, \boldsymbol{S}_{k,11}, \boldsymbol{M}_{k,11}, \mathsf{X}_{k,11} \in \mathbb{R}^{r_u \times r_u}$,
- $\boldsymbol{G}_{k,22}, \boldsymbol{\Psi}_{22}, \boldsymbol{S}_{k,22}, \boldsymbol{M}_{k,22}, \mathsf{X}_{k,22} \in \mathbb{R}^{(d-r_u)\times(d-r_u)}$.

The upper and lower diagonal terms have the following iterations:

$$\boldsymbol{G}_{k+1,11} = \boldsymbol{G}_{k,11} + \gamma_k\big(\boldsymbol{\Psi}_{11}\boldsymbol{G}_{k,11} + \boldsymbol{G}_{k,11}\boldsymbol{\Psi}_{11} - 2\boldsymbol{G}_{k,11}^2 - 2\boldsymbol{G}_{k,12}\boldsymbol{G}_{k,12}^\top\big) + \gamma_k^2\boldsymbol{S}_{k,11} + \gamma_k\boldsymbol{M}_{k+1,11}.$$

and

$$\boldsymbol{G}_{k+1,22} = \boldsymbol{G}_{k,22} + \gamma_k\big(\boldsymbol{\Psi}_{22}\boldsymbol{G}_{k,22} + \boldsymbol{G}_{k,22}\boldsymbol{\Psi}_{22} - 2\boldsymbol{G}_{k,22}^2 - 2\boldsymbol{G}_{k,12}^\top\boldsymbol{G}_{k,12}\big) + \gamma_k^2\boldsymbol{S}_{k,22} + \gamma_k\boldsymbol{M}_{k+1,22}.$$

Lemma 1. *We have*

$$0 \preceq \boldsymbol{S}_k \preceq C\mathcal{P}(\boldsymbol{G}_k)\operatorname{Tr}(\boldsymbol{G}_k)$$

Proof. $\boldsymbol{S}_k$ is positive semi-definite, therefore, the lower bound follows. Moreover,

$$\mathbb{E}_k[\mathsf{R}_{k+1}(\boldsymbol{G}_k)] \preceq \mathbb{E}[\big(y_{k+1} - \hat{y}_{k+1}\big)^2\mathrm{He}_2(\boldsymbol{x}_{k+1})\boldsymbol{G}_k\mathrm{He}_2(\boldsymbol{x}_{k+1})] \overset{(a)}{\preceq} C\mathcal{P}(\boldsymbol{G}_k)\operatorname{Tr}(\boldsymbol{G}_k)\boldsymbol{I}_d,$$

where we used Proposition (13) in (a). □

D.1 Definitions and Bounds

In the following, we consider the concentration bound (Ben Arous et al., 2025, Lemma 1) for the bounds on the initial matrix $\boldsymbol{G}_0$, where the union bound gives the probability bound considered in Theorem 3. We make the following definitions:

- **Noise sequence.** $\boldsymbol{N}_0 = 0$ and $\boldsymbol{N}_{k+1} = \boldsymbol{N}_k + \gamma_k\boldsymbol{M}_{k+1}$.
- **Bounding sequences.** We define the following:
  - *Upper bound for tails.* For $\bar{\boldsymbol{G}}_{0,22} = \frac{1.1}{d}\boldsymbol{I}_{d-r_u}$, we define

$$\bar{\boldsymbol{G}}_{t+1,22} = \bar{\boldsymbol{G}}_{k,22} + \tfrac{\gamma_k}{1+1.1\gamma_k}\big(\bar{\boldsymbol{G}}_{k,22}\boldsymbol{\Psi}_{22} + \boldsymbol{\Psi}_{22}\bar{\boldsymbol{G}}_{k,22} - 2\bar{\boldsymbol{G}}_{k,22}^2\big) + \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u}$$

  - *Upper bound for head.* For $\bar{\boldsymbol{G}}_0 = \frac{1.1}{d}\boldsymbol{I}_{r_u}$, we define

$$\bar{\boldsymbol{G}}_{k+1} = \bar{\boldsymbol{G}}_k + \tfrac{\gamma_k}{1+1.1\gamma_k}\Big(\bar{\boldsymbol{G}}_k(\boldsymbol{\Psi}_{11} + \tfrac{1}{\rho_d^{1/3}}\boldsymbol{I}_{r_u}) + (\boldsymbol{\Psi}_{11} + \tfrac{1}{\rho_d^{1/3}}\boldsymbol{I}_{r_u})\bar{\boldsymbol{G}}_k - 2\bar{\boldsymbol{G}}_k^2\Big) + \tfrac{C\gamma_k^2}{(1+1.1\gamma_k)^2}\boldsymbol{\Psi}_{11}^2 + \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u}$$

  - *Lower bound for head.* For $\underline{\boldsymbol{G}}_0 = \frac{0.9}{d}\boldsymbol{I}_{r_u}$, we define

$$\underline{\boldsymbol{G}}_{k+1} = \underline{\boldsymbol{G}}_k + \tfrac{\gamma_k}{1-1.1\gamma_k}(\underline{\boldsymbol{G}}_k(\boldsymbol{\Psi}_{11} - \tfrac{C^*}{\rho_d^{1/3}}\boldsymbol{I}_{r_u} - \delta\boldsymbol{I}_{r_u}) + (\boldsymbol{\Psi}_{11} - \tfrac{C^*}{\rho_d^{1/3}}\boldsymbol{I}_{r_u} - \delta\boldsymbol{I}_{r_u})\underline{\boldsymbol{G}}_k - 2\underline{\boldsymbol{G}}_k^2)$$

$$- \tfrac{C\gamma_k^2}{(1-1.1\gamma_k)^2}(\boldsymbol{\Psi}_{11} - \delta\boldsymbol{I}_{r_u})^2 - \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u}$$

- **Stopping times.** We consider the intersection of the following events:
  - $\mathcal{E}_{\text{bnd},1} \equiv \big\{\mathcal{P}(\boldsymbol{G}_k) \le 2 \text{ AND } 1 \le \gamma_k \le \rho_d^{1/3}\big\}$
  - $\mathcal{E}_{\text{bnd},2} \equiv \big\{\|\boldsymbol{G}_k\|_2 \le 1.1 \text{ AND } \|\boldsymbol{G}_k\|_F \le 4 \text{ AND } \operatorname{Tr}(\boldsymbol{G}_k) \le C(1 + r_u^{1-\alpha})$ AND $\|\gamma_k\boldsymbol{S}_k\|_2 \le \frac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\big\}$
  - $\mathcal{E}_{\text{dyn}} \equiv \{\frac{C}{\gamma_k} \ge \frac{P(\boldsymbol{G}_k)F(k)}{\log d}$ AND $\nu_k = (1 \pm o_d(1))\frac{1}{d\rho_d}\sum_{s=0}^{k-1}(\frac{s}{k})^q P(\boldsymbol{G}_k)\}$
  - $\mathcal{E}_{\text{noise},1} \equiv \big\{\frac{-1}{\rho_d^{1/3}}\bar{\boldsymbol{G}}_k \preceq \boldsymbol{N}_{k,11} \preceq \frac{1}{\rho_d^{1/3}}\bar{\boldsymbol{G}}_k$ AND $\frac{-1}{\rho_d^{1/3}}\bar{\boldsymbol{G}}_k \preceq \bar{\boldsymbol{V}}_k\big(\boldsymbol{I}_{r_u} + \frac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k\big)^{-1}\boldsymbol{N}_{k,11} + \boldsymbol{N}_{k,11}\big(\boldsymbol{I}_{r_u} + \frac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k\big)^{-1}\bar{\boldsymbol{V}}_k \preceq \frac{1}{\rho_d^{1/3}}\bar{\boldsymbol{G}}_k\big\}$

– $\mathcal{E}_{\text{noise},2} \equiv \big\{\|\bar{\boldsymbol{V}}_{k,22}\|_2 \le 2r_u^{-\alpha}$ AND $\|\boldsymbol{N}_{k,22}\|_2 \le \frac{C(1+r_u^{1-\alpha})}{\rho_d^{1/3}d}$
AND $\|\bar{\boldsymbol{V}}_{k,22}\boldsymbol{N}_{k,22} + \boldsymbol{N}_{k,22}\bar{\boldsymbol{V}}_{k,22}\|_2 \le \frac{4C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\big\}$

We define

$$\mathcal{T}_{\text{bad}}(\omega) \coloneqq \inf\{k \ge 0 \ : \ \omega \notin \mathcal{E}_{\text{bnd},1} \cap \mathcal{E}_{\text{bnd},2} \cap \mathcal{E}_{\text{dyn}} \cap \mathcal{E}_{\text{noise},1} \cap \mathcal{E}_{\text{noise},2}\}$$

and

$$\mathcal{T}_{\text{hit}} \coloneqq \mathcal{T}_{\text{bad}} \wedge \inf\big\{k \ge 0 \ : \ F(k) > \tfrac{r_u^{\alpha}}{2}\big\}.$$

D.2 Bounding System Decomposition

D.2.1 Upper Bounds

Upper Bound for $\mathbf{G}_{k,22}$

We have

$$\begin{aligned}\boldsymbol{G}_{k+1,22} \preceq \boldsymbol{G}_{k,22} + \gamma_k(\boldsymbol{\Psi}_{22}\boldsymbol{G}_{k,22} + \boldsymbol{G}_{k,22}\boldsymbol{\Psi}_{22} - 2\boldsymbol{G}_{k,22}^2) \\ + \gamma_k \tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u} + \gamma_k \boldsymbol{M}_{k+1,22}\end{aligned}$$

We use $\boldsymbol{V}_{k,22} \coloneqq 2\boldsymbol{G}_{k,22} - \boldsymbol{\Psi}_{22}$. For $t < \mathcal{T}_{\text{hit}}$, we have

$$\begin{aligned}\boldsymbol{V}_{k+1,22} &= \boldsymbol{V}_{k,22} + \gamma_k\boldsymbol{\Psi}_{22}^2 - \gamma_k\boldsymbol{V}_{k,22}^2 + \gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u} + \gamma_k 2\boldsymbol{M}_{k+1,22} \\ &\preceq \boldsymbol{V}_{k,22}\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\boldsymbol{V}_{k,22}\Big)^{-1} + \gamma_k\boldsymbol{\Psi}_{22}^2 + \gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u} + \gamma_k 2\boldsymbol{M}_{k+1,22}\end{aligned}$$

Proposition 2. *We have* $\boldsymbol{G}_{k\wedge\mathcal{T}_{hit},22} \preceq \bar{\boldsymbol{G}}_{k\wedge\mathcal{T}_{hit},22} + \boldsymbol{N}_{k\wedge\mathcal{T}_{hit},22}$.

Proof. We let $\bar{\boldsymbol{V}}_{k,22} \coloneqq 2\bar{\boldsymbol{G}}_{k,22} - \boldsymbol{\Psi}_{22}$. We use proof by induction. Consider $k < \mathcal{T}_{\text{hit}}$ and assume that the statement holds for t. We have

$$\begin{aligned}\boldsymbol{V}_{k+1,22} &\preceq (\bar{\boldsymbol{V}}_{k,22} + 2\boldsymbol{N}_{k,22})\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}(\bar{\boldsymbol{V}}_{k,22} + 2\boldsymbol{N}_{k,22})\Big)^{-1} + \gamma_k\boldsymbol{\Psi}_{22}^2 \\ &\quad + \gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u} + \gamma_k\boldsymbol{M}_{k+1,22} \\ &= \bar{\boldsymbol{V}}_{k,22}\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1} + \gamma_k\boldsymbol{\Psi}_{22}^2 + \gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u} + \gamma_k\boldsymbol{M}_{k+1,22} \\ &\quad + \Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1}2\boldsymbol{N}_{k,22}\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}(\bar{\boldsymbol{V}}_{k,22} + 2\boldsymbol{N}_{k,22})\Big)^{-1}. \qquad \text{(D.1)}\end{aligned}$$

By using Proposition 14, we can show that

$$\begin{aligned}&\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1}2\boldsymbol{N}_{k,22}\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}(\bar{\boldsymbol{V}}_{k,22} + 2\boldsymbol{N}_{k,22})\Big)^{-1} \\ &\qquad \preceq \Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1}2\boldsymbol{N}_{k,22}\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1}\end{aligned}$$

where

$$\begin{aligned}&\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1}2\boldsymbol{N}_{k,22}\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1} \\ &= 2\boldsymbol{N}_{k,22} - \tfrac{\gamma_k}{1+1.1\gamma_k}2\boldsymbol{N}_{k,22}\bar{\boldsymbol{V}}_k(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k)^{-1} - \tfrac{\gamma_k}{1+1.1\gamma_k}2\bar{\boldsymbol{V}}_k(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k)^{-1}\boldsymbol{N}_{k,22} \\ &\quad + \tfrac{2\gamma_k^2}{(1+1.1\gamma_k)^2}\bar{\boldsymbol{V}}_k(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k)^{-1}\boldsymbol{N}_{k,22}\bar{\boldsymbol{V}}_k(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k)^{-1} \\ &\preceq 2\boldsymbol{N}_{k,22} + \tfrac{\gamma_k}{1+1.1\gamma_k}\tfrac{8C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u} + \gamma_k^2\tfrac{Cr_u^{-2\alpha}(1+r_u^{1-\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u}\end{aligned}$$

Then, we have

$$(\text{D.1}) \preceq \bar{\boldsymbol{V}}_{k,22}\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1} + \gamma_k\boldsymbol{\Psi}_{22}^2 + \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u} + 2\boldsymbol{N}_{k+1,22}$$

Therefore, we have

$$\begin{aligned}
\boldsymbol{G}_{k+1,22} &\preceq \bar{\boldsymbol{G}}_{k,22} + \tfrac{\gamma_k}{1+1.1\gamma_k}\big(\bar{\boldsymbol{G}}_{k,22}\boldsymbol{\Psi}_{22} + \boldsymbol{\Psi}_{22}\bar{\boldsymbol{G}}_{k,22} - 2\bar{\boldsymbol{G}}_{k,22}^2\big) \\
&\quad + \tfrac{0.5\gamma_k^2}{(1+1.1\gamma_k)^2}\bar{\boldsymbol{V}}_{k,22}^3(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22})^{-1} + \tfrac{1.1\gamma_k^2}{1+1.1\gamma_k}\boldsymbol{\Psi}_{22}^2 \\
&\quad + \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u} + \boldsymbol{N}_{k+1,22} \\
&\overset{(a)}{\preceq} \bar{\boldsymbol{G}}_{k,22} + \tfrac{\gamma_k}{1+1.1\gamma_k}\big(\bar{\boldsymbol{G}}_{k,22}\boldsymbol{\Psi}_{22} + \boldsymbol{\Psi}_{22}\bar{\boldsymbol{G}}_{k,22} - 2\bar{\boldsymbol{G}}_{k,22}^2\big) \\
&\quad + \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{d-r_u} + \boldsymbol{N}_{k+1,22} \\
&= \bar{\boldsymbol{G}}_{t+1,22} + \boldsymbol{N}_{k+1,22},
\end{aligned}$$

where we used $\gamma_k \le \rho_d^{2/3}(r_u^\alpha + r_u)$ in (a).

□

Corollary 1. *Under the assumption that* $\rho_d^{1/3-\epsilon} \gg 1 + r_u^{1-\alpha}$ *and* $d \ge \Omega_\epsilon(1)$*, we have* $\bar{\boldsymbol{G}}_{k,22} \preceq \frac{e^2}{d}\boldsymbol{I}_{d-r_u}$ *for* $k \le \inf\{k \ge 0 : \sum_{s=0}^{k-1}\gamma_s > \frac{r_u^\alpha}{2}\}$*. Therefore,* $\boldsymbol{G}_{k\wedge\mathcal{T}_{hit},22} \preceq \frac{1.1e^2}{d}\boldsymbol{I}_{d-r_u}$

Proof. The result follows Proposition from 15. □

Upper Bound for $\mathbf{G}_{k,11}$

We recall

$$\begin{aligned}
\boldsymbol{G}_{k+1,11} &= \boldsymbol{G}_{k,11} + \gamma_k\big(\boldsymbol{\Psi}_{11}\boldsymbol{G}_{k,11} + \boldsymbol{G}_{k,11}\boldsymbol{\Psi}_{11} - 2\boldsymbol{G}_{k,11}^2 - 2\boldsymbol{G}_{k,12}\boldsymbol{G}_{k,12}^\top\big) \\
&\quad + \gamma_k^2\boldsymbol{S}_{k,11} + \gamma_k\boldsymbol{M}_{k+1,11} \\
&\preceq \boldsymbol{G}_{k,11} + \gamma_k\big(\boldsymbol{\Psi}_{11}\boldsymbol{G}_{k,11} + \boldsymbol{G}_{k,11}\boldsymbol{\Psi}_{11} - 2\boldsymbol{G}_{k,11}^2\big) + \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} + \gamma_k\boldsymbol{M}_{k+1,11}
\end{aligned}$$

By using $\boldsymbol{V}_k^+ = 2\boldsymbol{G}_{k,11} - \boldsymbol{\Psi}_{11}$, we have for $t < \mathcal{T}_{\text{hit}}$

$$\boldsymbol{V}_{k+1}^+ \preceq \boldsymbol{V}_k^+(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\boldsymbol{V}_k^+)^{-1} + \gamma_k\boldsymbol{\Psi}_{11}^2 + \gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} + 2\gamma_k\boldsymbol{M}_{k+1,11}.$$

Proposition 3. *We have* $\boldsymbol{G}_{k\wedge\mathcal{T}_{hit},11} \preceq \bar{\boldsymbol{G}}_{k\wedge\mathcal{T}_{hit}} + \boldsymbol{N}_{k\wedge\mathcal{T}_{hit},11}$.

Proof. We let $\bar{\boldsymbol{V}}_k := 2\bar{\boldsymbol{G}}_k - \boldsymbol{\Psi}_{11}$. We use proof by induction. Consider $k < \mathcal{T}_{\text{hit}}$ and assume that the statement holds for k. We have

$$\begin{aligned}
\boldsymbol{V}_{k+1}^+ &\preceq (\bar{\boldsymbol{V}}_k + 2\boldsymbol{N}_{k,11})\Big(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}(\bar{\boldsymbol{V}}_k + 2\boldsymbol{N}_{k,11})\Big)^{-1} + \gamma_k\boldsymbol{\Psi}_{11}^2 \\
&\quad + \gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} + \gamma_k\boldsymbol{M}_{k+1,11} \\
&= \bar{\boldsymbol{V}}_k\Big(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k\Big)^{-1} + \gamma_k\boldsymbol{\Psi}_{11}^2 + \gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} + \gamma_k\boldsymbol{M}_{k+1,11} \\
&\quad + \Big(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k\Big)^{-1}2\boldsymbol{N}_{k,11}\Big(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}(\bar{\boldsymbol{V}}_k + 2\boldsymbol{N}_{k,11})\Big)^{-1}. \qquad (\text{D.2})
\end{aligned}$$

By using Proposition 14, we can show that

$$\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1}2\boldsymbol{N}_{k,22}\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}(\bar{\boldsymbol{V}}_{k,22} + 2\boldsymbol{N}_{k,22})\Big)^{-1}$$

$$\preceq \Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1} 2\boldsymbol{N}_{k,22}\Big(\boldsymbol{I}_{d-r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_{k,22}\Big)^{-1}$$

where

$$\begin{aligned}
&\Big(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k\Big)^{-1} 2\boldsymbol{N}_{k,11}\Big(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k\Big)^{-1} \\
&= 2\boldsymbol{N}_{k,11} - \tfrac{\gamma_k}{1+1.1\gamma_k} 2\boldsymbol{N}_{k,11}\bar{\boldsymbol{V}}_k(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k)^{-1} - \tfrac{\gamma_k}{1+1.1\gamma_k} 2\bar{\boldsymbol{V}}_k(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k)^{-1}\boldsymbol{N}_{k,11} \\
&+ \tfrac{2\gamma_k^2}{(1+1.1\gamma_k)^2}\bar{\boldsymbol{V}}_k\Big(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k\Big)^{-1}\boldsymbol{N}_{k,11}\bar{\boldsymbol{V}}_k\Big(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k\Big)^{-1} \\
&\preceq 2\boldsymbol{N}_{k,11} + \tfrac{\gamma_k}{1+1.1\gamma_k}\tfrac{2}{\rho_d^{1/3}}\bar{\boldsymbol{G}}_k + \tfrac{\gamma_k^2}{(1+1.1\gamma_k)^2}\tfrac{2}{\rho_d^{1/3}}\bar{\boldsymbol{V}}_k^2.
\end{aligned}$$

By (D.2), we have

$$\begin{aligned}
\boldsymbol{V}_{k+1}^+ &\preceq \bar{\boldsymbol{V}}_k\Big(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k\Big)^{-1} + \gamma_k\boldsymbol{\Psi}_{11}^2 + \gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} + \frac{\gamma_k}{1+1.1\gamma_k}\frac{2}{\rho_d^{1/3}}\bar{\boldsymbol{G}}_k \\
&\quad + \frac{\gamma_k^2}{(1+1.1\gamma_k)^2}\frac{2}{\rho_d^{1/3}}\bar{\boldsymbol{V}}_k^2 + 2\boldsymbol{N}_{k+1,11}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\boldsymbol{G}_{k+1,11} &\preceq \bar{\boldsymbol{G}}_k + \tfrac{\gamma_k}{1+1.1\gamma_k}\Big(\bar{\boldsymbol{G}}_k(\boldsymbol{\Psi}_{11} + \tfrac{1}{\rho_d^{1/3}}\boldsymbol{I}_{r_u}) + (\boldsymbol{\Psi}_{11} + \tfrac{1}{\rho_d^{1/3}}\boldsymbol{I}_{r_u})\bar{\boldsymbol{G}}_k - 2\bar{\boldsymbol{G}}_k^2\Big) \\
&\quad + \tfrac{0.5\gamma_k^2}{(1+1.1\gamma_k)^2}\bar{\boldsymbol{V}}_k^3(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1+1.1\gamma_k}\bar{\boldsymbol{V}}_k)^{-1} + \tfrac{\gamma_k^2}{(1+1.1\gamma_k)^2}\tfrac{2}{\rho_d^{1/3}}\bar{\boldsymbol{V}}_k^2 + \tfrac{1.1\gamma_k^2}{1+1.1\gamma_k}\boldsymbol{\Psi}_{11}^2 \\
&\quad + \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} + \boldsymbol{N}_{k+1,11} \\
&\preceq \bar{\boldsymbol{G}}_k + \tfrac{\gamma_k}{1+1.1\gamma_k}\Big(\bar{\boldsymbol{G}}_k(\boldsymbol{\Psi}_{11} + \tfrac{1}{\rho_d^{1/3}}\boldsymbol{I}_{r_u}) + (\boldsymbol{\Psi}_{11} + \tfrac{1}{\rho_d^{1/3}}\boldsymbol{I}_{r_u})\bar{\boldsymbol{G}}_k - 2\bar{\boldsymbol{G}}_k^2\Big) \\
&\quad + \tfrac{C\gamma_k^2}{(1+1.1\gamma_k)^2}\boldsymbol{\Psi}_{11}^2 + \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} + \boldsymbol{N}_{k+1,11} \\
&= \bar{\boldsymbol{G}}_{k+1} + \boldsymbol{N}_{k+1,11}
\end{aligned}$$

□

D.2.2 Lower Bound

We define $\delta := \frac{1.1e^2}{d}$. Recall we have

$$\begin{aligned}
\boldsymbol{G}_{k+1,11} &\succeq \boldsymbol{G}_{k,11} + \gamma_k\Big(\boldsymbol{\Psi}_{11}\boldsymbol{G}_{k,11} + \boldsymbol{G}_{k,11}\boldsymbol{\Psi}_{11} - 2\boldsymbol{G}_{k,11}^2 - 2\delta\boldsymbol{G}_{k,11}\Big) \\
&\quad - \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} + \gamma_k\boldsymbol{M}_{k+1,11}
\end{aligned}$$

Using $\boldsymbol{V}_k^- := 2\boldsymbol{G}_{k,11} - (\boldsymbol{\Psi}_{11} - \delta\boldsymbol{I}_{r_u})$, we have the following for $t < \mathcal{T}_{\text{hit}}$

$$\boldsymbol{V}_{k+1}^- \succeq \boldsymbol{V}_k^-(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1-1.1\gamma_k}\boldsymbol{V}_k^-)^{-1} + \gamma_k(\boldsymbol{\Psi}_{11} - \delta\boldsymbol{I}_{r_u})^2 - \gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} + 2\gamma_k\boldsymbol{M}_{k+1,11}.$$

Proposition 4. *We have* $\boldsymbol{G}_{k\wedge\mathcal{T}_{hit},11} \succeq \underline{\boldsymbol{G}}_{k\wedge\mathcal{T}_{hit}} + \boldsymbol{N}_{k\wedge\mathcal{T}_{hit},11}$

Proof. Define the deterministic lower comparison process $\underline{\boldsymbol{V}}_k := 2\underline{\boldsymbol{G}}_k - (\boldsymbol{\Psi}_{11} - \delta\boldsymbol{I}_{r_u})$.

We prove by induction. Consider $k < \mathcal{T}_{\text{hit}}$ and suppose that the statement holds for k. We have

$$\begin{aligned}
\boldsymbol{V}_{k+1}^- &\succeq (\underline{\boldsymbol{V}}_k + 2\boldsymbol{N}_{k,11})\Big(\boldsymbol{I}_{r_u} + \tfrac{\gamma_k}{1-1.1\gamma_k}(\underline{\boldsymbol{V}}_k + 2\boldsymbol{N}_{k,11})\Big)^{-1} + \gamma_k(\boldsymbol{\Psi}_{11} - \delta\boldsymbol{I}_{r_u})^2 \\
&\quad - \gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} + 2\gamma_k\boldsymbol{M}_{k+1,11}.
\end{aligned}$$

$$=\underline{\boldsymbol{V}}_k\,(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}+\gamma_k(\boldsymbol{\Psi}_{11}-\delta\boldsymbol{I}_{r_u})^2-\gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u} \tag{D.3}$$
$$+2\gamma_k\boldsymbol{M}_{k+1,11}+(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}2\boldsymbol{N}_{k,11}(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}(\underline{\boldsymbol{V}}_k\,+2\boldsymbol{N}_{k,11}))^{-1}$$

For the last term, applying an expansion gives

$$\begin{aligned}
&(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}2\boldsymbol{N}_{k,11}\Big(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}(\underline{\boldsymbol{V}}_k\,+2\boldsymbol{N}_{k,11})\Big)^{-1}\\
&\quad=(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}2\boldsymbol{N}_{k,11}(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}\\
&\quad-\tfrac{4\gamma_k}{1-1.1\gamma_k}\Big((\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}\boldsymbol{N}_{k,11}(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}\Big)\times\\
&\quad\Big(\boldsymbol{N}_{k,11}(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}(\underline{\boldsymbol{V}}_k\,+2\boldsymbol{N}_{k,11}))^{-1}\Big)\\
&\quad\succeq(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}2\boldsymbol{N}_{k,11}(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}-\tfrac{C\gamma_k}{(1-1.1\gamma_k)\rho_d^{2/3}}\underline{\boldsymbol{G}}_k
\end{aligned}$$

Using stopping times on $\boldsymbol{N}_{k,11}$, $\bar{\boldsymbol{V}}_k$,

$$\begin{aligned}
&(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}2\boldsymbol{N}_{k,11}(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}\\
&\qquad\succeq 2\boldsymbol{N}_{k,11}-\tfrac{2\gamma_k}{(1-1.1\gamma_k)\rho_d^{1/3}}\underline{\boldsymbol{G}}_k-\tfrac{2\gamma_k^2}{(1-1.1\gamma_k)^2\rho_d^{1/3}}\underline{\boldsymbol{V}}_k{}^2.
\end{aligned}$$

Hence, by (D.3),

$$\begin{aligned}
\boldsymbol{V}_{k+1}^-\succeq&\underline{\boldsymbol{V}}_k\,(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}+\gamma_k(\boldsymbol{\Psi}_{11}-\delta\boldsymbol{I}_{r_u})^2-\tfrac{2\gamma_k}{(1-1.1\gamma_k)\rho_d^{1/3}}\underline{\boldsymbol{G}}_k-\tfrac{C\gamma_k}{(1-1.1\gamma_k)\rho_d^{2/3}}\underline{\boldsymbol{G}}_k\\
&-\gamma_k\tfrac{2C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u}-\tfrac{2\gamma_k^2}{(1-1.1\gamma_k)^2\rho_d^{1/3}}\underline{\boldsymbol{V}}_k{}^2+2\boldsymbol{N}_{k+1,11}
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\boldsymbol{G}_{k+1,11}\succeq&\underline{\boldsymbol{G}}_k+\tfrac{\gamma_k}{1-1.1\gamma_k}\Big(\underline{\boldsymbol{G}}_k(\boldsymbol{\Psi}_{11}-\tfrac{C^*}{\rho_d^{1/3}}\boldsymbol{I}_{r_u}-\delta\boldsymbol{I}_{r_u})+(\boldsymbol{\Psi}_{11}-\tfrac{C^*}{\rho_d^{1/3}}\boldsymbol{I}_{r_u}-\delta\boldsymbol{I}_{r_u})\underline{\boldsymbol{G}}_k-2\underline{\boldsymbol{G}}_k^2\Big)\\
&+\tfrac{\gamma_k^2}{2(1-1.1\gamma_k)^2}\underline{\boldsymbol{V}}_k{}^3(\boldsymbol{I}_{r_u}+\tfrac{\gamma_k}{1-1.1\gamma_k}\underline{\boldsymbol{V}}_k\,)^{-1}-\tfrac{1.1\gamma_k^2}{1-1.1\gamma_k}(\boldsymbol{\Psi}_{11}-\delta\boldsymbol{I}_{r_u})^2\\
&-\tfrac{2\gamma_k^2}{(1-1.1\gamma_k)^2\rho_d^{1/3}}\underline{\boldsymbol{V}}_k{}^2-\gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u}-\tfrac{C\gamma_k}{\rho_d^{2/3}}\boldsymbol{I}_{r_u}+\boldsymbol{N}_{k+1,11}\\
\succeq&\underline{\boldsymbol{G}}_k+\tfrac{\gamma_k}{1-1.1\gamma_k}\Big(\underline{\boldsymbol{G}}_k(\boldsymbol{\Psi}_{11}-\tfrac{C^*}{\rho_d^{1/3}}\boldsymbol{I}_{r_u}-\delta\boldsymbol{I}_{r_u})+(\boldsymbol{\Psi}_{11}-\tfrac{C^*}{\rho_d^{1/3}}\boldsymbol{I}_{r_u}-\delta\boldsymbol{I}_{r_u})\underline{\boldsymbol{G}}_k-2\underline{\boldsymbol{G}}_k^2\Big)\\
&-\tfrac{C\gamma_k^2}{(1-1.1\gamma_k)^2}(\boldsymbol{\Psi}_{11}-\delta\boldsymbol{I}_{r_u})^2-\gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u}-\tfrac{C\gamma_k}{\rho_d^{2/3}}\boldsymbol{I}_{r_u}+\boldsymbol{N}_{k+1,11}\\
=&\underline{\boldsymbol{G}}_{k+1}+\boldsymbol{N}_{k+1,11}
\end{aligned}$$

□

D.3 Analysis of Deterministic Bounding Systems

D.3.1 Upper Deterministic System

Recall that our upper deterministic system is defined as

$$\begin{aligned}
\bar{\boldsymbol{G}}_{k+1}=&\bar{\boldsymbol{G}}_k+\frac{\gamma_k}{1+1.1\gamma_k}\Big(\bar{\boldsymbol{G}}_k(\boldsymbol{\Psi}_{11}+\tfrac{1}{\rho_d^{1/3}}\boldsymbol{I}_{r_u})+(\boldsymbol{\Psi}_{11}+\tfrac{1}{\rho_d^{1/3}}\boldsymbol{I}_{r_u})\bar{\boldsymbol{G}}_k-2\bar{\boldsymbol{G}}_k^2\Big)\\
&+\tfrac{C\gamma_k^2}{(1+1.1\gamma_k)^2}\boldsymbol{\Psi}_{11}^2+\gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u}
\end{aligned}$$

Define, for this subsection, letting C represent any constant

$$\begin{cases}
c_k:=\frac{C\gamma_k}{1+1.1\gamma_k}\\
E:=\frac{C(r_u^{-\alpha}+r_u^{1-2\alpha})(1+1.1\gamma_k)}{d\rho_d^{1/3}}\\
\mu_i:=\lambda_i+\frac{1}{\rho_d^{1/3}}
\end{cases}$$

The result of this section is to provide a deterministic bound for this sequence, shown in the following proposition:

Proposition 5. *Under the conditions*

$$\begin{cases} \bar{\boldsymbol{G}}_0 = \frac{1.1}{d}\boldsymbol{I}_{r_u} \\ k \le Cd\rho_d^{1+\varepsilon}\log d \\ r_u \asymp \log^{\frac{1}{\alpha}} d \\ \rho_d \asymp \log^{5.5} d \end{cases}$$

We have

$$\bar{\boldsymbol{G}}_{k,ii} \le 1.1\mu_i \wedge \frac{1.2}{d}\exp\left((2+\frac{1}{d})\mu_i\sum_{s=0}^{k-1}\gamma_s\right).$$

Proof. Consider the diagonal entries $\bar{g}_k := \bar{\boldsymbol{G}}_{k,ii}$ of the deterministic system:

$$\bar{g}_{k+1} = \bar{g}_k + 2c_k\bar{g}_k(\mu_i - \bar{g}_k) + c_k^2\lambda_i^2 + c_kE$$

Define, for $D = O(1)$,

$$h_k := \frac{\bar{g}_k}{\mu_i} + r$$
$$r := \frac{-1+\sqrt{1+\frac{2DE}{\mu_i^2}}}{2}$$

where absorbing the c_k^2 term in this regime and rearranging yields

$$h_{k+1} = h_k + 2\mu_i c_k h_k(\sqrt{1+\frac{2DE}{\mu_i^2}} - h_k)$$

Using (16) on h_k and rearranging, we get an upper bound

$$h_k \le \left(\frac{1}{\sqrt{1+\frac{2DE}{\mu_i^2}}} + \frac{d\mu_i}{1.1}\exp\left(-2\sqrt{1+\frac{2DE}{\mu_i^2}}\sum_{s=0}^{k-1}\frac{\mu_i\gamma_k}{1-2\sqrt{1+\frac{2DE}{\mu_i^2}}\gamma_k\mu_i}\right)\right)^{-1} \tag{D.4}$$

$$\bar{g}_k \le 1.1\mu_i \wedge \frac{1.2}{d}\exp((2+o(\frac{1}{d}))\mu_i\sum_{s=0}^{k-1}\gamma_s)$$

□

For the lower bound of the upper deterministic system, we will lower bound $\underline{\boldsymbol{G}}_k$ instead in the next section.

D.3.2 Lower Deterministic System

Recall that our lower deterministic system is defined as

$$\underline{\boldsymbol{G}}_{k+1} = \underline{\boldsymbol{G}}_k + \tfrac{\gamma_k}{1-1.1\gamma_k}(\underline{\boldsymbol{G}}_k(\boldsymbol{\Psi}_{11} - \tfrac{C^*}{\rho_d^{1/3}}\boldsymbol{I}_{r_u} - \delta\boldsymbol{I}_{r_u}) + (\boldsymbol{\Psi}_{11} - \tfrac{C^*}{\rho_d^{1/3}}\boldsymbol{I}_{r_u} - \delta\boldsymbol{I}_{r_u})\underline{\boldsymbol{G}}_k - 2\underline{\boldsymbol{G}}_k^2)$$
$$- \tfrac{C\gamma_k^2}{(1-1.1\gamma_k)^2}(\boldsymbol{\Psi}_{11} - \delta\boldsymbol{I}_{r_u})^2 - \gamma_k\tfrac{C(r_u^{-\alpha}+r_u^{1-2\alpha})}{\rho_d^{1/3}d}\boldsymbol{I}_{r_u}$$

Define, for this subsection, letting C represent any constant

$$\begin{cases} c_k := \frac{C\gamma_k}{1-1.1\gamma_k} \\ E := \frac{C(r_u^{-\alpha}+r_u^{1-2\alpha})(1-1.1\gamma_k)}{d\rho_d^{1/3}} \\ \mu_i := \lambda_i - \frac{C^*}{\rho_d^{1/3}} - \delta \\ l_i := \lambda_i - \delta \end{cases}$$

Proposition 6. *Under the conditions*

$$\begin{cases} \underline{\boldsymbol{G}}_0 = \frac{0.9}{d}\boldsymbol{I}_{r_u} \\ k \le Cd\rho_d^{1+\varepsilon}\log d \\ r_u \asymp \log^{\frac{1}{\alpha}} d \\ \rho_d \asymp \log^{5.5} d \end{cases}$$

We have

$$\underline{\boldsymbol{G}}_{k,ii} \ge 0.4\mu_i \wedge \frac{0.4}{d}\exp\left((2-\frac{1}{d})\mu_i\sum_{s=0}^{k-1}\gamma_k\right)$$

Proof. Consider the diagonal entries $\underline{g}_k := \underline{\boldsymbol{G}}_{k,ii}$:

$$\underline{g}_{k+1} = \underline{g}_k + 2c_k\underline{g}_k(\mu_i - \underline{g}_k) - c_k^2 l_i^2 - c_k E$$

Define, for $D = O(1)$,

$$y_k := \frac{\underline{g}_k}{\mu_i} + r$$
$$r := \frac{-1+\sqrt{1-\frac{2DE}{\mu_i^2}}}{2}$$

to get the iteration

$$y_{k+1} = y_k + 2c_k\mu_i y_k(\sqrt{1-\frac{2DE}{\mu_i^2}} - y_k)$$

By (16) on y_k, we get a lower bound

$$y_k \ge \left(\frac{1}{\sqrt{1-\frac{2DE}{\mu_i^2}}} + \frac{d\mu_i}{0.9}\exp\left(-2\sqrt{1-\frac{2DE}{\mu_i^2}}\sum_{s=0}^{k-1}\frac{\mu_i\gamma_k}{1+2\sqrt{1-\frac{2DE}{\mu_i^2}}\gamma_k\mu_i}\right)\right)^{-1} \tag{D.5}$$

which, after rearranging, yields

$$\underline{g}_k \ge 0.4\mu_i \wedge \frac{0.4}{d}\exp\left((2-o(\frac{1}{d}))\mu_i\sum_{s=0}^{k-1}\gamma_k\right)$$

□

Note that we can show "closeness" of the two deterministic systems by bounding their ratio by the two bounds in (D.5) and (D.4) in the following corollary:

Corollary 2. *For $k < \mathcal{T}_{hit}$, $i \le r_u$, we have*

$$\underline{\boldsymbol{G}}_{k,ii}^{-1}\bar{\boldsymbol{G}}_{k,ii} \le 1.2$$

Proof. We have that

$$\frac{\bar{g}_k}{\underline{g}_k} \asymp \frac{h_k}{y_k} \le \max\left\{\frac{\sqrt{1+\frac{2DE}{\mu_i^2}}}{\sqrt{1-\frac{2DE}{\mu_i^2}}}, \exp\left(O(\frac{1}{d\rho_d})\right)\right\} \le 1.2$$

□

D.4 Noise Characterization

This section will prove the noise concentration stopping times by relating the upper deterministic system to the noise sequence. We start by defining $\mathbb{P}_k[\cdot] = \mathbb{P}[\cdot|\mathcal{F}_k]$. Moreover, we let

- $\{\mathsf{X}_{k+1}\}_{t\geq 0}$ and $\{\mathsf{R}_{k+1}\}_{t\geq 0}$ be the martingale difference sequence defined in Section D.
- Let $\boldsymbol{K}_1, \boldsymbol{K}_2 \in \mathbb{R}^{d\times m}$ $\mathcal{F}_k$-measurable (possibly random) matrices.
- Let $\{\mathsf{G}_k\}_{t\geq 0}$ is a $\mathcal{F}_k$-measurable random matrix sequence.

Recall that our noise terms are defined as

$$\begin{aligned}\boldsymbol{M}_{k+1} = (\boldsymbol{G}_k\mathsf{X}_{k+1} + \mathsf{X}_{k+1}\boldsymbol{G}_k) + \gamma_k\big((\boldsymbol{\Psi}_* - \boldsymbol{G}_k)\boldsymbol{G}_k\mathsf{X}_{k+1} + \mathsf{X}_{k+1}\boldsymbol{G}_k(\boldsymbol{\Psi}_* - \boldsymbol{G}_k)\big)\\ + \gamma_k\big(\mathsf{R}_{k+1}(\boldsymbol{G}_k) - \mathbb{E}_k[\mathsf{R}_{k+1}(\boldsymbol{G}_k)]\big)\end{aligned} \tag{D.6}$$

We define $e_k := y_{k+1} - \hat{y}_{k+1}$, $\mathcal{P}(\boldsymbol{G}_k) := \|\boldsymbol{\Psi}_* - \boldsymbol{G}_k\|_F^2 + \sigma^2$ and

$$\mathcal{A}_{k+1}(\boldsymbol{K}_1, \boldsymbol{K}_2) \equiv \left\{\left\|\boldsymbol{K}_1^\top \mathsf{X}_{k+1}\mathsf{G}_k\boldsymbol{K}_2\right\|_2 \leq L^2(\sigma + \|\boldsymbol{\Psi}_* - \boldsymbol{G}_k\|_F)\|\boldsymbol{K}_1\|_F\|\mathsf{G}_k\boldsymbol{K}_2\|_F\right\}$$

$$\begin{aligned}\mathcal{B}_{k+1}(\boldsymbol{K}_1, \boldsymbol{K}_2) \equiv \Big\{\left\|\boldsymbol{K}_1^\top \mathsf{R}_{k+1}(\mathsf{G}_k)\boldsymbol{K}_2\right\|_2 \leq L^3(\sigma + \|\boldsymbol{\Psi}_* - \boldsymbol{G}_k\|_F)^2 \times\\ \left(\sqrt{\mathrm{Tr}(\mathsf{G}_k)^2 + \|\mathsf{G}_k\|_F^2}\|\boldsymbol{K}_1\|_F\|\boldsymbol{K}_2\|_F\right)\Big\}.\end{aligned}$$

Proposition 7. *There exists a universal constant $C, c > 0$ such that for $L \geq \Omega(1)$, the following holds:*

1. *We have $\mathbb{P}_k\Big[\mathcal{A}_{k+1}(\boldsymbol{K}_1, \boldsymbol{K}_2) \cap \big\{|e_{k+1}| \leq L(\sigma + \|\boldsymbol{\Psi}_* - \boldsymbol{G}_k\|_F)\big\}\Big] \geq 1 - 2e^{-cL}$, and*

$$\mathbb{E}_k\left[(\mathrm{sym}\left(\boldsymbol{K}_1^\top \mathsf{X}_{k+1}\mathsf{G}_k\boldsymbol{K}_2\right))^2\right] \preceq C\mathcal{P}(\boldsymbol{G}_k)\Big(\|\boldsymbol{K}_2^\top \mathsf{G}_k\|_F^2\boldsymbol{K}_1^\top\boldsymbol{K}_1 + \|\boldsymbol{K}_1\|_F^2\boldsymbol{K}_2^\top\mathsf{G}_k^2\boldsymbol{K}_2\Big).$$

2. *We have $\mathbb{P}_k\Big[\mathcal{B}_{k+1}(\boldsymbol{K}_1, \boldsymbol{K}_2) \cap \big\{|e_{k+1}| \leq L(\sigma + \|\boldsymbol{\Psi}_* - \boldsymbol{G}_k\|_F)\big\}\Big] \geq 1 - 5e^{-cL}$. Moreover,*

$$\begin{aligned}\mathbb{E}_k\left[(\mathrm{sym}\left(\boldsymbol{K}_1^\top \mathsf{R}_{k+1}(\mathsf{G}_k)\boldsymbol{K}_2\right))^2\right] \preceq C\mathcal{P}(\boldsymbol{G}_k)^2\|\boldsymbol{K}_2\|_F^2\big(\mathrm{Tr}(\mathsf{G}_k)^2 + \|\mathsf{G}_k\|_F^2 + \|\mathsf{G}_k\boldsymbol{K}_2\|_F^2\big)\boldsymbol{K}_1^\top\boldsymbol{K}_1\\ + C\mathcal{P}(\boldsymbol{G}_k)^2\|\boldsymbol{K}_1\|_F^2\big(\mathrm{Tr}(\mathsf{G}_k)^2 + \|\mathsf{G}_k\|_F^2 + \|\mathsf{G}_k\boldsymbol{K}_1\|_F^2\big)\boldsymbol{K}_2^\top\boldsymbol{K}_2\end{aligned}$$

Proof. First, we derive a concentration bound for $|e_{k+1}|$. We observe that

$$e_{k+1} = \mathrm{Tr}\left((\boldsymbol{\Psi}_* - \boldsymbol{G}_k)(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1} - \boldsymbol{I}_d)\right) + \sigma\epsilon_{k+1}$$

where ϵ_{k+1} is a unit norm subexponential independent noise. By Corollary 3,

$$\mathbb{E}_k\left[|e_{k+1}|^p\right] \leq C^p p^p(\sigma + \|\boldsymbol{\Psi}_* - \boldsymbol{G}_k\|_F)^p \text{ for } p \geq 2,$$

which, by Proposition 11, implies

$$\mathbb{P}_k\Big[|e_{k+1}| \geq Cu(\sigma + \|\boldsymbol{\Psi}_* - \boldsymbol{G}_k\|_F)\Big] \leq e^{-u/e} \text{ for } u \geq 2e, \tag{D.7}$$

In the following, we prove each item separately.

First item. We define

$$\begin{aligned}\Gamma_{k+1} &= \left(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top - \boldsymbol{I}_d\right)\\ \boldsymbol{K}_1^\top\mathsf{X}_{k+1}\mathsf{G}_k\boldsymbol{K}_2 &= e_{k+1}\boldsymbol{K}_1^\top\Gamma_{k+1}\mathsf{G}_k\boldsymbol{K}_2.\end{aligned}$$

By Proposition 12, we have

$$\mathbb{P}\Big[\|\boldsymbol{K}_1^\top\Gamma_{k+1}\mathsf{G}_k\boldsymbol{K}_2\|_2 \geq Cu\|\boldsymbol{K}_1\|_F\|\mathsf{G}_k\boldsymbol{K}_2\|_F\Big] \leq e^{-u/e} \text{ for } u \geq 2e.$$

By (D.7), we have the probability bound. For the variance bound, we have

$$\begin{aligned}
&\mathbb{E}_k\Big[\operatorname{sym}\big(\boldsymbol{K}_1^\top e_{k+1}\Gamma_{k+1}\mathsf{G}_k\boldsymbol{K}_2\big)^2\Big] \\
&\quad\preceq \boldsymbol{K}_1^\top\mathbb{E}_k\left[e_{k+1}^2\Gamma_{k+1}\mathsf{G}_k\boldsymbol{K}_2\boldsymbol{K}_2^\top\mathsf{G}_k\Gamma_{k+1}\right]\boldsymbol{K}_1 \\
&\quad+\boldsymbol{K}_2^\top\mathsf{G}_k\mathbb{E}_k\left[e_{k+1}^2\Gamma_{k+1}\boldsymbol{K}_1\boldsymbol{K}_1^\top\Gamma_{k+1}\right]\boldsymbol{K}_2 \\
&\quad\preceq C(\|\boldsymbol{\Psi}_*-\boldsymbol{G}_k\|_F^2+\sigma^2)\Big(\|\boldsymbol{K}_2^\top\mathsf{G}_k\|_F^2\boldsymbol{K}_1^\top\boldsymbol{K}_1+\|\boldsymbol{K}_1\|_F^2\boldsymbol{K}_2^\top\mathsf{G}_k^2\boldsymbol{K}_2\Big).
\end{aligned}$$

where we used Proposition 13 in the last inequality.

Second item. We define

$$\boldsymbol{K}_1^\top\mathsf{R}_{k+1}(\mathsf{G}_k)\boldsymbol{K}_2=e_{k+1}^2\boldsymbol{K}_1^\top(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top-\boldsymbol{I}_d)\mathsf{G}_k(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top-\boldsymbol{I}_d)\boldsymbol{K}_2$$

By Proposition 12, for $u\geq 2e$, we have

$$\begin{aligned}
\mathbb{P}\Big[\|\boldsymbol{K}_1^\top\Gamma_{k+1}\mathsf{G}_k\Gamma_{k+1}\boldsymbol{K}_2\|_2\geq Cu\sqrt{\operatorname{Tr}(\mathsf{G}_k)^2+\|\mathsf{G}_k\|_F^2}\|\boldsymbol{K}_1\|_F\|\boldsymbol{K}_2\|_F\Big] \\
\leq 4e^{-u/e}
\end{aligned}$$

By (D.7), we have the probability bound. For the variance bound, we have

$$\begin{aligned}
&\mathbb{E}_k\Big[\operatorname{sym}\big(e_{k+1}^2\boldsymbol{K}_1^\top\Gamma_{k+1}\mathsf{G}_k\Gamma_{k+1}\boldsymbol{K}_2\big)^2\Big] \\
&\quad\preceq \boldsymbol{K}_1^\top\mathbb{E}_k\left[e_{k+1}^4\Gamma_{k+1}\mathsf{G}_k\Gamma_{k+1}\boldsymbol{K}_2\boldsymbol{K}_2^\top\Gamma_{k+1}\mathsf{G}_k\Gamma_{k+1}\right]\boldsymbol{K}_1 \\
&\quad+\boldsymbol{K}_2^\top\mathbb{E}_k\left[e_{k+1}^4\Gamma_{k+1}\mathsf{G}_k\Gamma_{k+1}\boldsymbol{K}_1\boldsymbol{K}_1^\top\Gamma_{k+1}\mathsf{G}_k\Gamma_{k+1}\right]\boldsymbol{K}_2 \\
&\quad\preceq C(\|\boldsymbol{\Psi}_*-\boldsymbol{G}_k\|_F^2+\sigma^2)^2\|\boldsymbol{K}_2\|_F^2\big(\operatorname{Tr}(\mathsf{G}_k)^2+\|\mathsf{G}_k\|_F^2+\|\mathsf{G}_k\boldsymbol{K}_2\|_F^2\big)\boldsymbol{K}_1^\top\boldsymbol{K}_1 \\
&\quad+C(\|\boldsymbol{\Psi}_*-\boldsymbol{G}_k\|_F^2+\sigma^2)^2\|\boldsymbol{K}_1\|_F^2\big(\operatorname{Tr}(\mathsf{G}_k)^2+\|\mathsf{G}_k\|_F^2+\|\mathsf{G}_k\boldsymbol{K}_1\|_F^2\big)\boldsymbol{K}_2^\top\boldsymbol{K}_2
\end{aligned}$$

where we used Proposition 13 in the last inequality.

□

D.4.1 Variance Bounds

The goal of this section is to prove that the stopping time for the noise sequence $\boldsymbol{N}_{k,11}$ holds. We seek to use quadratic variance to bound $\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\boldsymbol{N}_{k,11}\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}$

Proposition 8. *Let*

$$S_k:=\sum_{j=1}^{k}\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\gamma_j\boldsymbol{M}_{j+1,11}\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}$$

Define

$$\begin{cases}\mu_i:=\lambda_i+\frac{1}{\rho_d^{1/3}}\\ F(j):=\sum_{s=0}^{j-1}\gamma_s\end{cases}$$

and consider the regime

$$\begin{cases}r_u\eqsim\log^{\frac{1}{\alpha}}d\\ \rho_d\eqsim\log^{5.5}d\end{cases}$$

Assume the following conditions hold:

- $$\bar{\boldsymbol{G}}_{k,ii}\preceq 1.1\mu_i\wedge\frac{1.2}{d}\exp\left((2+\frac{1}{d})\mu_iF(j)\right)$$

- $$\bar{\boldsymbol{G}}_{k,ii} \geq \underline{\boldsymbol{G}}_{k,ii} \geq 0.4\mu_i \wedge \frac{0.4}{d}\exp\left((2-\frac{1}{d})\mu_i F(k)\right)$$

Then, we have

$$\textit{Quad}(S_k) := \sum_{j=1}^{k} \mathbb{E}_{j-1}\left[\left(\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\gamma_j \boldsymbol{M}_{j+1,11}\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\right)^2\right] \overset{d\to\infty}{\longrightarrow} 0$$

Proof. Define

$$K_1 = K_2 := \begin{bmatrix} \bar{\boldsymbol{G}}_k^{-\frac{1}{2}} \\ 0 \end{bmatrix}$$

We proceed by decomposing (D.6) into its three terms. The first term is

$$\sum_{j=1}^{k} K_1^T \gamma_j (\boldsymbol{G}_j \mathsf{X}_{j+1} + \mathsf{X}_{j+1}\boldsymbol{G}_j) K_2$$

and we seek to bound its quadratic variance

$$\sum_{j=1}^{k} \gamma_j^2 \mathbb{E}[\text{sym}(K_1^T \boldsymbol{G}_j \mathsf{X}_{j+1} K_2)^2]$$

where, recalling the bounds set in (7), we have

$$\mathbb{E}[\text{sym}(K_1^T \boldsymbol{G}_j \mathsf{X}_{j+1} K_2)^2] \leq C\mathcal{P}(\boldsymbol{G}_j)\text{Tr}(K_1^T K_1)\text{Tr}(K_2^T \boldsymbol{G}_j^2 K_2)$$

Observe that in our case, expanding out the block form of $\boldsymbol{G}_j$

$$K_1^T K_1 = \bar{\boldsymbol{G}}_k^{-1}$$
$$K_2^T \boldsymbol{G}_j^2 K_2 \preceq \bar{\boldsymbol{G}}_k^{-\frac{1}{2}} \boldsymbol{G}_{j,11}^2 \bar{\boldsymbol{G}}_k^{-\frac{1}{2}} + \frac{1.1e^2}{d}\bar{\boldsymbol{G}}_k^{-\frac{1}{2}} \boldsymbol{G}_{j,11} \bar{\boldsymbol{G}}_k^{-\frac{1}{2}}$$

So

$$\mathbb{E}[\text{sym}(K_1^T \boldsymbol{G}_j \mathsf{X}_{j+1} K_2)^2] \leq C\mathcal{P}(\boldsymbol{G}_j)\text{Tr}(\bar{\boldsymbol{G}}_j^{-1})\left[\text{Tr}(\bar{\boldsymbol{G}}_k^{-\frac{1}{2}} \boldsymbol{G}_{j,11}^2 \bar{\boldsymbol{G}}_k^{-\frac{1}{2}}) + \frac{1}{d}\text{Tr}(\bar{\boldsymbol{G}}_k^{-\frac{1}{2}} \boldsymbol{G}_{j,11} \bar{\boldsymbol{G}}_k^{-\frac{1}{2}})\right]$$

Since $\bar{\boldsymbol{G}}_j$ is increasing, its inverse is decreasing and can be bounded by $\bar{\boldsymbol{G}}_0^{-1} = \frac{d}{1.1}\boldsymbol{I}_{r_u}$. Furthermore, for $k \leq \mathcal{T}_{\text{bad}}$,

$$\boldsymbol{G}_{j,11} \preceq (1 + \frac{1}{\rho_d^{1/3}})\bar{\boldsymbol{G}}_j \implies \bar{\boldsymbol{G}}_k^{-\frac{1}{2}} \boldsymbol{G}_{j,11} \bar{\boldsymbol{G}}_k^{-\frac{1}{2}} \preceq (1 + \frac{1}{\rho_d^{1/3}})\boldsymbol{I}_{r_u}$$

Since the term is PSD, its eigenvalues are bounded in the interval $[0, 1 + \frac{1}{\rho_d}]$, and

$$\bar{\boldsymbol{G}}_k^{-\frac{1}{2}} \boldsymbol{G}_{j,11} \bar{\boldsymbol{G}}_k^{-1} \boldsymbol{G}_{j,11} \bar{\boldsymbol{G}}_k^{-\frac{1}{2}} \preceq (1 + \frac{1}{\rho_d^{1/3}})\bar{\boldsymbol{G}}_k^{-\frac{1}{2}} \boldsymbol{G}_{j,11} \bar{\boldsymbol{G}}_k^{-\frac{1}{2}}$$
$$\boldsymbol{G}_{j,11} \bar{\boldsymbol{G}}_k^{-1} \boldsymbol{G}_{j,11} \preceq (1 + \frac{1}{\rho_d^{1/3}})\boldsymbol{G}_{j,11} \preceq (1 + \frac{1}{\rho_d^{1/3}})^2 \bar{\boldsymbol{G}}_j$$

We also observe that

$$\frac{1}{d}\text{Tr}(\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\boldsymbol{G}_{j,11}\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}) \leq (1+\frac{1}{\rho_d^{1/3}})\frac{r_u}{d}\boldsymbol{I}_{r_u}$$

Thus our variance becomes

$$\begin{aligned}\sum_{j=1}^{k}\gamma_j^2\mathbb{E}[\text{sym}(K_1^T\boldsymbol{G}_j\mathsf{X}_{j+1}K_2)^2] &\leq Cdr_u\sum_{j=1}^{k}\gamma_j^2\mathcal{P}(\boldsymbol{G}_j)(\text{Tr}(\bar{\boldsymbol{G}}_j)+\frac{r_u}{d}) \\ &\leq \underbrace{Cdr_u\sum_{j=1}^{k}\gamma_j^2\mathcal{P}(\boldsymbol{G}_j)\text{Tr}(\bar{\boldsymbol{G}}_j)}_{(A)} + \underbrace{Cr_u^2\sum_{j=1}^{k}\gamma_j^2\mathcal{P}(\boldsymbol{G}_j)}_{(B)}\end{aligned}$$

Note that $\bar{\boldsymbol{G}}_{k,ii}$ converges at a different timestep for each i. The crossing time k_{i*} must satisfy

$$F(k_{i*}) \eqsim \frac{\log(d\mu_i)}{2\mu_i}$$

Thus, we have the cases

$$\bar{\boldsymbol{G}}_{k,ii} \preceq \begin{cases}1.1\mu_i & k \leq k_{i*} \\ \frac{1.2}{d}\exp\left((2+\frac{1}{d})\mu_i F(j)\right) & k > k_{i*}\end{cases}$$

Note that in the time regime $F(k) \eqsim \bar{C}\log d$, there exists a set of cutoff times $\mathcal{U}_{\mathcal{T}}\{i : \bar{C}\log d << F(k_{i*}), i \leq r_u\}$. Denote $\bar{i} := \inf(\mathcal{U}_{\mathcal{T}})$, and for $\bar{i} \leq i \leq r_u$, $\bar{\boldsymbol{G}}_{k,ii}$ will not converge. Indeed, for high dimensions, $\bar{i} < r_u$ as $F(k_{i*})$ is on the order of $\log d$.

We can break up the sum (A) into these regimes:

- (I) - exponential phase for head - $i \leq \bar{i}, j \leq k_{i*}$
- (II) - converged phase for head - $i \leq \bar{i}, j \geq k_{i*}$
- (III) - exponential phase for tails - $i > \bar{i}, j \leq k$

For the following, let C represent any $O(1)$ factor. We have,

$$\begin{aligned}(A) = Cdr_u\sum_{j=1}^{k}\gamma_j^2\mathcal{P}(\boldsymbol{G}_j)\text{Tr}(\bar{\boldsymbol{G}}_j) &= Cdr_u\sum_{i=1}^{r_u}\sum_{j=1}^{k}\gamma_j^2\mathcal{P}(\boldsymbol{G}_j)\bar{\boldsymbol{G}}_{j,ii} \\ = \underbrace{Cdr_u\sum_{i=1}^{\bar{i}}\sum_{j=1}^{k_i*}\gamma_j^2\mathcal{P}(\boldsymbol{G}_j)\bar{\boldsymbol{G}}_{j,ii}}_{(I)} &+ \underbrace{Cdr_u\sum_{i=1}^{\bar{i}}\sum_{j=k_{i*}+1}^{k}\gamma_j^2\mathcal{P}(\boldsymbol{G}_j)\bar{\boldsymbol{G}}_{j,ii}}_{(II)} \\ &+ \underbrace{Cdr_u\sum_{i=\bar{i}+1}^{r_u}\sum_{j=1}^{k}\gamma_j^2\mathcal{P}(\boldsymbol{G}_j)\bar{\boldsymbol{G}}_{j,ii}}_{(III)}\end{aligned}$$

Recall that from the stopping times $\mathcal{E}_{dyn}, \mathcal{E}_{bnd,1}$, we can write

$$\frac{\mathcal{P}(\boldsymbol{G}_j)}{\sqrt{\frac{\nu_j}{\log d}+\epsilon}} \leq \frac{2}{\sqrt{\epsilon}} \wedge \frac{C\log d}{F(j)}$$

Define k' to be the time of regime change:

$$F(k') = \frac{C}{2}\log d\sqrt{\epsilon} \eqsim \log d$$

We will simplify each sum $(I), (II), (III)$ separately, starting with (I):

$$\begin{aligned}
(I) &= Cdr_u \sum_{i=1}^{\bar{i}} \sum_{j=1}^{k_{i*}} \gamma_j^2 \mathcal{P}(\boldsymbol{G}_j) \bar{\boldsymbol{G}}_{j,ii} \\
&= \frac{Cr_u}{\rho_d} \sum_{i=1}^{\bar{i}} \sum_{j=1}^{k_{i*}} \gamma_j \frac{\mathcal{P}(\boldsymbol{G}_j)}{\sqrt{\frac{\nu_j}{\log d} + \epsilon}} \bar{\boldsymbol{G}}_{j,ii} \\
&= \frac{Cr_u \log d}{\rho_d} \sum_{i=1}^{\bar{i}} \sum_{j=k'+1\wedge k_{i*}}^{k_{i*}} \frac{\gamma_j \bar{\boldsymbol{G}}_{j,ii}}{F(j)} \wedge \frac{Cr_u}{\rho_d\sqrt{\epsilon}} \sum_{i=1}^{\bar{i}} \sum_{j=1}^{k'\wedge k_{i*}} \gamma_j \bar{\boldsymbol{G}}_{j,ii} \\
&= \frac{Cr_u \log d}{d\rho_d} \sum_{i=1}^{\bar{i}} \sum_{j=k'+1\wedge k_{i*}}^{k_{i*}} \frac{\gamma_j \exp\left((2+\frac{1}{d})\mu_i F(j)\right)}{F(j)} \\
&\qquad \wedge \frac{Cr_u}{d\rho_d\sqrt{\epsilon}} \sum_{i=1}^{\bar{i}} \sum_{j=1}^{k'\wedge k_{i*}} \gamma_j \exp\left((2+\frac{1}{d})\mu_i F(j)\right)
\end{aligned}$$

Since $F(k'+1), F(k_{i*}) > 1$, each summand is monotone and we can bound by their respective integrals with timestep γ_j:

$$\begin{aligned}
(I) &\le \frac{Cr_u \log d}{d\rho_d} \sum_{i=1}^{\bar{i}} \int_{F(k'+1\wedge k_{i*})}^{F(k_{i*}+1)} \frac{\exp\left((2+\frac{1}{d})\mu_i x\right)}{x} \mathrm{d}x \\
&\qquad \wedge \frac{Cr_u}{d\rho_d\sqrt{\epsilon}} \sum_{i=1}^{\bar{i}} \int_{F(1)}^{F(k'\wedge k_{i*}+1)} \exp\left((2+\frac{1}{d})\mu_i x\right) \mathrm{d}x \\
&\le \frac{Cr_u \log d}{d\rho_d} \sum_{i=1}^{\bar{i}} \frac{\exp((2+\frac{1}{d})\mu_i F(k_{i*}+1))}{\mu_i F(k_{i*}+1)} \wedge \frac{Cr_u}{d\rho_d\sqrt{\epsilon}} \sum_{i=1}^{\bar{i}} \frac{\exp((2+\frac{1}{d})\mu_i F(k_{i*}\wedge k'+1))}{\mu_i}
\end{aligned}$$

where we used $\mu_i F(k_{i*}+1) \gtrsim r_u^{-\alpha}\log d \eqsim 1$ to ensure monotonicity and simplify the integral

$$\int_{F(k'+1\wedge k_{i*})}^{F(k_{i*})} \frac{\exp\left((2+\frac{1}{d})\mu_i x\right)}{x}\mathrm{d}x \le \int_1^{(2+\frac{1}{d})\mu_i F(k_{i*}+1)} \frac{\exp(y)}{y}\mathrm{d}y \le \frac{\exp((2+\frac{1}{d})\mu_i F(k_{i*}+1))}{\mu_i F(k_{i*}+1)}$$

Next,

$$\begin{aligned}
(II) &= Cdr_u \sum_{i=1}^{\bar{i}} \sum_{j=k_{i*}+1}^{k} \gamma_j^2 \mathcal{P}(\boldsymbol{G}_j) \bar{\boldsymbol{G}}_{j,ii} \\
&\le Cdr_u \sum_{i=1}^{\bar{i}} \sum_{j=k_{i*}+1}^{k} \mu_i \gamma_j^2 \mathcal{P}(\boldsymbol{G}_j) \\
&= \frac{Cr_u}{\rho_d} \sum_{i=1}^{\bar{i}} \sum_{j=k_{i*}+1}^{k} \mu_i \gamma_j \frac{\mathcal{P}(\boldsymbol{G}_j)}{\sqrt{\frac{\nu_j}{\log d} + \epsilon}} \\
&\le \frac{Cr_u}{\rho_d\sqrt{\epsilon}} \sum_{i=1}^{\bar{i}} \sum_{j=k_{i*}+1}^{k_{i*}+1\vee k'} \mu_i\gamma_j \wedge \frac{Cr_u \log d}{\rho_d} \sum_{i=1}^{\bar{i}} \sum_{j=k_{i*}+1\vee k'}^{k} \frac{\mu_i\gamma_j}{F(j)}
\end{aligned}$$

$$\leq \frac{Cr_u}{\rho_d\sqrt{\epsilon}}(F(k_{i*}+1)\vee F(k')) \wedge \frac{Cr_u \log d}{\rho_d}\log(F(k+1))$$

where we applied

$$\begin{aligned}\sum_{j=k_{i*}+1\vee k'}^{k}\frac{\gamma_j}{F(j)} &\leq \sum_{j=k_{i*}+1\vee k'}^{k}\left(1+\frac{\gamma_j}{F(j)}\right)\log(1+\frac{\gamma_j}{F(j)})\\ &\leq \left(1+\max_{j<k}\frac{\gamma_j}{F(k_{i*}+1\vee k')}\right)\sum_{j=k_{i*}+1\vee k'}^{k}\log(\frac{F(j+1)}{F(j)})\\ &\leq \left(1+\max_{j<k}\frac{\gamma_j}{F(k_{i*}+1\vee k')}\right)\log(\frac{F(k+1)}{F(k_{i*}+1\vee k')})\end{aligned}$$

and simplified the sum over i since $\sum_{i=1}^{\bar{i}}\mu_i \eqsim 1$. Since $F(k_{i*}+1\vee k') \eqsim \log d$ and $\gamma_j \lesssim \frac{1}{d\rho_d^{2/3}}$, the leading coefficient is $O(1)$.

Next,

$$\begin{aligned}(III) &= Cdr_u\sum_{i=\bar{i}+1}^{r_u}\sum_{j=1}^{k}\gamma_j^2\mathcal{P}(\boldsymbol{G}_j)\bar{\boldsymbol{G}}_{j,ii}\\ &= \frac{Cr_u}{\rho_d}\sum_{i=\bar{i}+1}^{r_u}\sum_{j=1}^{k}\gamma_j\frac{\mathcal{P}(\boldsymbol{G}_j)}{\sqrt{\frac{\nu_j}{\log d}+\epsilon}}\bar{\boldsymbol{G}}_{j,ii}\\ &\leq \frac{Cr_u}{\rho_d\sqrt{\epsilon}}\sum_{i=\bar{i}+1}^{r_u}\sum_{j=1}^{k'}\gamma_j\bar{\boldsymbol{G}}_{j,ii} \wedge \frac{Cr_u\log d}{\rho_d}\sum_{i=\bar{i}+1}^{r_u}\sum_{j=k'}^{k}\frac{\gamma_j\bar{\boldsymbol{G}}_{j,ii}}{F(j)}\\ &\leq \frac{Cr_u\log d}{d\rho_d}\sum_{i=\bar{i}+1}^{r_u}\sum_{j=k'}^{k}\frac{\gamma_j\exp((2+\frac{1}{d})\mu_i F(j))}{F(j)}\\ &\quad\wedge\frac{Cr_u}{d\rho_d\sqrt{\epsilon}}\sum_{i=\bar{i}+1}^{r_u}\sum_{j=1}^{k'}\gamma_j\exp\left((2+\frac{1}{d})\mu_i F(j)\right)\\ &\leq \frac{Cr_u\log d}{d\rho_d}\sum_{i=\bar{i}+1}^{r_u}\int_{F(k')}^{F(k+1)}\frac{\exp((2+\frac{1}{d})\mu_i x)}{x}\mathrm{d}x\\ &\quad\wedge\frac{Cr_u}{d\rho_d\sqrt{\epsilon}}\sum_{i=\bar{i}+1}^{r_u}\int_{F(1)}^{F(k')}\exp\left((2+\frac{1}{d})\mu_i x\right)\mathrm{d}x\\ &\leq \frac{Cr_u\log d}{d\rho_d}\sum_{i=\bar{i}+1}^{r_u}\frac{\exp((2+\frac{1}{d})\mu_i F(k+1))}{\mu_i F(k+1)} \wedge \frac{Cr_u}{d\rho_d\sqrt{\epsilon}}\sum_{i=\bar{i}+1}^{r_u}\frac{\exp((2+\frac{1}{d})\mu_i F(k')}{\mu_i}\end{aligned}$$

And (B) evaluates to

$$\begin{aligned}(B) &= Cr_u^2\sum_{j=1}^{k}\gamma_j^2\mathcal{P}(\boldsymbol{G}_j)\\ &= \frac{Cr_u^2}{d\rho_d}\sum_{j=1}^{k}\gamma_j\frac{\mathcal{P}(\boldsymbol{G}_j)}{\sqrt{\frac{\nu_j}{\log d}+\epsilon}}\\ &\leq \frac{Cr_u^2\log d}{d\rho_d}\sum_{j=k'}^{k}\frac{\gamma_j}{F(j)} \wedge \frac{Cr_u^2F(k')}{d\rho_d\sqrt{\epsilon}}\\ &\leq \frac{Cr_u^2\log d}{d\rho_d}\log(F(k+1)) \wedge \frac{Cr_u^2F(k')}{d\rho_d\sqrt{\epsilon}}\end{aligned}$$

Finally, we have

$$
\begin{aligned}
(I)+(II)+(III)+(B) \leq \frac{Cr_u}{d\rho_d}\sum_{i=1}^{\bar{i}}\frac{\exp((2+\frac{1}{d})\mu_i F(k_{i*}+1))}{\mu_i} \\
&\wedge \frac{Cr_u}{d\rho_d\sqrt{\epsilon}}\sum_{i=1}^{\bar{i}}\frac{\exp((2+\frac{1}{d})\mu_i F(k_{i*}\wedge k'+1))}{\mu_i} \\
&+ \frac{Cr_u}{\rho_d\sqrt{\epsilon}}(F(k_{i*}+1)\vee F(k')) \wedge \frac{Cr_u\log d}{\rho_d}\log F(k+1) \\
&+ \frac{Cr_u\log d}{d\rho_d}\sum_{i=\bar{i}+1}^{r_u}\frac{\exp((2+\frac{1}{d})\mu_i F(k+1))}{\mu_i} \\
&\wedge \frac{Cr_u}{d\rho_d\sqrt{\epsilon}}\sum_{i=\bar{i}+1}^{r_u}\frac{\exp((2+\frac{1}{d})\mu_i F(k')}{\mu_i} \\
&+ \frac{Cr_u^2\log d}{d\rho_d}\log F(k+1) \wedge \frac{Cr_u^2 F(k')}{d\rho_d\sqrt{\epsilon}}
\end{aligned}
$$

Since $F(k)=o(d)$, we have for our lower and upper bounds on $\bar{\boldsymbol{G}}_{k,ii}$:

$$
\exp\left((2-\frac{1}{d})\mu_i F(k)\right) \eqsim \exp\left((2+\frac{1}{d})\mu_i F(k)\right)
$$

and we simplify

$$
\begin{aligned}
(I)+(II)+(III)+(B) &\lesssim \frac{Cr_u}{\rho_d}\sum_{i=1}^{\bar{i}}\frac{\bar{\boldsymbol{G}}_{k_{i*}+1,ii}}{\mu_i} \wedge \frac{Cr_u}{\rho_d}\sum_{i=1}^{\bar{i}}\frac{\bar{\boldsymbol{G}}_{k_{i*}\wedge k'+1,ii}}{\mu_i} \\
&\quad + \frac{C\log(F(k+1))r_u\log d}{\rho_d} \\
&\quad + \frac{Cr_u\log d}{\rho_d}\sum_{i=\bar{i}+1}^{r_u}\frac{\bar{\boldsymbol{G}}_{k+1,ii}}{\mu_i} \wedge \frac{Cr_u}{\rho_d}\sum_{i=\bar{i}+1}^{r_u}\frac{\bar{\boldsymbol{G}}_{k',ii}}{\mu_i} \\
&\quad + \frac{C\log(F(k+1))r_u^2\log d}{d\rho_d} \wedge \frac{Cr_u^2\log d}{d\rho_d} \\
&\lesssim \frac{C\log(F(k+1))r_u\log d}{\rho_d} + \frac{Cr_u^2\log d}{\rho_d} + \frac{C\log(F(k+1))r_u^2\log d}{d\rho_d}
\end{aligned}
$$

where each $\bar{\boldsymbol{G}}_{k,ii}$-numerator can be upper bounded by $1.1\mu_i$.

The second summation term of (D.6) is

$$
\sum_{j=1}^{k} K_1^T\gamma_j^2((\boldsymbol{\Psi}_*-\boldsymbol{G}_j)\boldsymbol{G}_j\mathsf{X}_{j+1}+\mathsf{X}_{j+1}\boldsymbol{G}_j(\boldsymbol{\Psi}_*-\boldsymbol{G}_j))K_2
$$

We can define

$$
A_1 := K_1 \qquad\qquad A_2 := (\boldsymbol{\Psi}_*-\boldsymbol{G}_j)K_2
$$

to apply Proposition (7) to bound its quadratic variance:

$$\begin{aligned}\sum_{j=1}^{k}\gamma_j^4\mathbb{E}[\mathrm{sym}(K_1^T\mathsf{X}_{j+1}\boldsymbol{G}_j(\boldsymbol{\Psi}_*-\boldsymbol{G}_j)K_2)^2] &\le C\sum_{j=1}^{k}\gamma_j^4\mathcal{P}(\boldsymbol{G}_j)\mathrm{Tr}(\bar{\boldsymbol{G}}_k^{-1})\mathrm{Tr}(A_2^T\boldsymbol{G}_j^2A_2)\\ &\le C\sum_{j=1}^{k}\gamma_j^4\mathcal{P}(\boldsymbol{G}_j)\mathrm{Tr}(\bar{\boldsymbol{G}}_k^{-1})\Big(\mathrm{Tr}(\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\boldsymbol{G}_{j,11}^2\bar{\boldsymbol{G}}_k^{-\frac{1}{2}})\\ &\quad+\frac{1}{d}\mathrm{Tr}(\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\boldsymbol{G}_{j,11}\bar{\boldsymbol{G}}_k^{-\frac{1}{2}})+\frac{1}{d^2}\mathrm{Tr}(\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\boldsymbol{G}_{j,11}\bar{\boldsymbol{G}}_k^{-\frac{1}{2}})\Big)\end{aligned}$$

where we used $\mathcal{E}_{bnd,1}$ to bound $\|\boldsymbol{\Psi}_*-\boldsymbol{G}_j\|_F^2$ These same trace terms were bounded in the previous noise term, so applying $\mathcal{E}_{bnd,1}$ to γ_k, we have

$$\begin{aligned}\sum_{j=1}^{k}\gamma_j^4\mathbb{E}[\mathrm{sym}(K_1^T\mathsf{X}_{j+1}\boldsymbol{G}_j(\boldsymbol{\Psi}_*-\boldsymbol{G}_j)K_2)^2] &\le \sum_{j=1}^{k}\gamma_j^4\Big(\mathbb{E}[\mathrm{sym}(K_1^T\mathsf{X}_{j+1}\boldsymbol{G}_jK_2)^2]\\ &\quad+\frac{\mathcal{P}(\boldsymbol{G}_j)\mathrm{Tr}(\boldsymbol{G}_k^{-1})}{d^2}\mathrm{Tr}(\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\boldsymbol{G}_{j,11}\bar{\boldsymbol{G}}_k^{-\frac{1}{2}})\Big)\\ &\lesssim \frac{1}{d^2\rho_d^{2/3}}\sum_{j=1}^{k}\gamma_j^2\mathbb{E}[\mathrm{sym}(K_1^T\mathsf{X}_{j+1}\boldsymbol{G}_jK_2)^2]\\ &\lesssim \frac{1}{d^2\rho_d^{2/3}}\end{aligned}$$

Finally, the last summation term of (D.6) is

$$\sum_{j=1}^{k}\gamma_j^2K_1^T(\mathsf{R}_{j+1}(\boldsymbol{G}_j)-\mathbb{E}_j[\mathsf{R}_{j+1}(\boldsymbol{G}_j)]K_2$$

which has quadratic variance bounded by

$$\begin{aligned}\sum_{j=1}^{k}\gamma_j^4\mathbb{E}[\mathrm{sym}(K_1^T\mathsf{R}(\boldsymbol{G}_j)K_2)^2] &\le C\sum_{j=1}^{k}\gamma_j^4\mathcal{P}(\boldsymbol{G}_j)\mathrm{Tr}(\bar{\boldsymbol{G}}_k^{-1})^2\Big(\mathrm{Tr}(\boldsymbol{G}_j)^2+\mathrm{Tr}(\boldsymbol{G}_j^2)\\ &\quad+\mathrm{Tr}(\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\boldsymbol{G}_{j,11}^2\bar{\boldsymbol{G}}_k^{-\frac{1}{2}})+\frac{r_u}{d}\Big)\\ &\le Cd^2r_u^2\sum_{j=1}^{k}\gamma_j^4\left(r_u^{2-2\alpha}+\mathrm{Tr}(\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\boldsymbol{G}_{j,11}^2\bar{\boldsymbol{G}}_k^{-\frac{1}{2}})+\frac{r_u}{d}\right)\\ &\le \frac{Cr_u^2}{\rho_d^{4/3}}\left(\sum_{j=1}^{k}r_u^{2-2\alpha}\gamma_j^2+\sum_{j=1}^{k}\gamma_j^2\mathbb{E}[\mathrm{sym}(K_1^TX_{j+1}\boldsymbol{G}_jK_2)^2]\right)\\ &\le \frac{Cr_u^2}{\rho_d^{4/3}}\end{aligned}$$

by applying stopping times for $k<\mathcal{T}_{\text{bad}}$ and reusing the results of the previous two sums.

Since $r_u\eqsim\log^{\frac{1}{\alpha}}d$, a choice of $\rho_d\eqsim\log^{5.5}d$ makes all variance terms tend to 0. □

By using a clipping argument with the Bernstein bound in (Ben Arous et al., 2025, Section F.6), we can show that $\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}\boldsymbol{N}_{k,11}\bar{\boldsymbol{G}}_k^{-\frac{1}{2}}$ satisfies the condition in stopping time in $\mathcal{T}_{\text{hit}}$ for the time horizon in Theorem 3. The remaining noise bounds follow from this bound straightforwardly.

D.4.2 Characterization of Learning Rate

As an intermediate step, we need to characterize the learning rate and risk evolution.

For this section, define

$$\begin{cases} \epsilon = O(1) \\ T_d := d\rho_d \log d \\ F(k) := \sum_{s=0}^{k-1} \gamma_s = \frac{1}{d\rho_d} \sum_{s=0}^{k-1} \frac{1}{\sqrt{\frac{\nu_s}{\log d}+\epsilon}} \\ F_*(k) := \frac{F(kT_d)}{\log d} \\ \mathcal{R}(\boldsymbol{G}_k) \eqsim (1+\frac{F(k)}{\log d})^{\frac{1-2\alpha}{\alpha}} \\ \nu_k = \frac{1+o_d(1)}{d\rho_d} \sum_{s=0}^{k-1} \left(\frac{s}{k}\right)^q (\mathcal{R}(\boldsymbol{G}_s)+\sigma^2) \end{cases}$$

Lemma 2. *For the functions defined above, and $k \le Cd\rho_d^{1+\varepsilon} \log d$, we have*

$$\frac{C}{\sqrt{\frac{\nu_k}{\log d}}+\epsilon} \ge \frac{\mathcal{P}(\boldsymbol{G}_k)F(k)}{\log d}$$

Proof. As $F(k)$ to satisfy the fixed point equation

$$F(k) = \frac{1}{d\rho_d} \sum_{i=0}^{k} \frac{1}{\sqrt{\frac{\nu_k}{\log d}}+\epsilon} \tag{D.8}$$

which converges to the ode as $d \to \infty$, we can show that

$$\frac{F(kd\rho_d \log d)}{\log d} := F_*(k) \eqsim ((1+k)^\alpha - 1) \wedge \sigma k^{1/2}$$

Now, we aim to show

$$\sqrt{\frac{1}{d\rho_d \log_d} \sum_{s=0}^{k-1} \left(\frac{s}{k}\right)^q \mathcal{P}(\boldsymbol{G}_s) + \epsilon} \gtrsim \frac{\mathcal{P}(\boldsymbol{G}_k)F(k)}{\log d}$$

It is helpful to write

$$F(k) \simeq \frac{1}{d\rho_d} \sum_{s=0}^{k-1} \frac{1}{\sqrt{\frac{1}{T_d}\sum_{j=0}^{s-1}(\frac{j}{s})^q(\mathcal{R}(G_j)+\sigma^2)+\epsilon}}$$

and

$$\mathcal{R}(G_j) \eqsim (1+F_*(\frac{j}{T_d}))^{\frac{1-2\alpha}{\alpha}}$$

Since $\alpha > 0.5$, consider that

$$\begin{aligned} \left(1+F_*(\frac{s}{T_d})\right)^{\frac{1-2\alpha}{\alpha}} &= \left(1+(1+\frac{s}{T_d})^\alpha \wedge (1+\frac{s}{T_d\sigma^2})^{1/2}\right)^{\frac{1-2\alpha}{\alpha}} \\ &\ge 2^{\frac{1-2\alpha}{\alpha}-1}\left((1+\frac{s}{T_d})^{1-2\alpha} + (1+\frac{s}{T_d\sigma^2})^{\frac{1-2\alpha}{2\alpha}}\right) \\ &\ge \frac{1}{8}\left((1+\frac{s}{T_d})^{1-2\alpha} + (1+\frac{s}{T_d\sigma^2})^{\frac{1-2\alpha}{2\alpha}}\right) \end{aligned}$$

We can consider the time scale kT_d instead of k and rewrite the LHS:

$$\sqrt{\frac{1}{d\rho_d \log_d} \sum_{s=0}^{kT_d-1} \left(\frac{s}{kTd}\right)^q \mathcal{P}(\boldsymbol{G}_s) + \epsilon}$$

$$= \sqrt{\frac{1}{T_d}\sum_{s=0}^{kT_d-1}\left(\frac{s}{kT_d}\right)^q\left\{\left(1+F_*(\frac{s}{T_d})\right)^{\frac{1-2\alpha}{\alpha}}+\sigma^2\right\}+\epsilon}$$

$$\geq \sqrt{\frac{1}{8T_d}\sum_{s=0}^{kT_d-1}\left(\frac{s}{kT_d}\right)^q\left\{(1+\frac{s}{T_d})^{1-2\alpha}+(1+\frac{s}{T_d\sigma^2})^{\frac{1-2\alpha}{2\alpha}}+8\sigma^2\right\}+\epsilon}$$

$$\eqsim \sqrt{\frac{1}{8}\int_0^k\left(\frac{x}{k}\right)^q\left((1+x)^{1-2\alpha}+(1+\frac{x}{\sigma^2})^{\frac{1-2\alpha}{2\alpha}}+8\sigma^2\right)\mathrm{d}x+\epsilon}$$

where the Riemann sum uniformly converges to its integral in large-d regimes.

Picking $q > 2\alpha - 2 \vee 2$, we have

$$\sqrt{\frac{1}{8}\int_0^k\left(\frac{x}{k}\right)^q\left((1+x)^{1-2\alpha}+(1+\frac{x}{\sigma^2})^{\frac{1-2\alpha}{2\alpha}}+8\sigma^2\right)\mathrm{d}x+\epsilon}$$

$$\geq \sqrt{\frac{k^{-q}}{8}\int_0^k x^q(1+x)^{1-2\alpha}\mathrm{d}x+\sigma^2k^{-q}\int_0^k x^q\mathrm{d}x+\epsilon}$$

$$\eqsim \sqrt{k^{2-2\alpha}+\sigma^2k+\epsilon}$$

For the RHS, consider the timescale kT_d again, so

$$\frac{\mathcal{P}(\boldsymbol{G}_{kT_d})F(kT_d)}{\log d} \eqsim \left(F_*(k)^{\frac{1-2\alpha}{\alpha}}+\sigma^2\right)F_*(k)$$

$$\eqsim \left\{\left((1+k)^\alpha\wedge(1+\frac{k}{\sigma^2})^{\frac{1}{2}}\right)^{\frac{1-2\alpha}{\alpha}}+\sigma^2\right\}\left\{(1+k)^\alpha\wedge(1+\frac{k}{\sigma^2})^{\frac{1}{2}}\right\}$$

$$\eqsim \left\{(1+k)^{1-2\alpha}\vee(1+\frac{k}{\sigma^2})^{\frac{1-2\alpha}{2\alpha}}+\sigma^2\right\}\left\{(1+k)^\alpha\wedge(1+\frac{k}{\sigma^2})^{\frac{1}{2}}\right\}$$

To analyze when the two terms cross at time k_*, we can consider asymptotically

$$k_*^\alpha \eqsim (\frac{k_*}{\sigma^2})^{\frac{1}{2}} \implies \sigma^2 \eqsim k_*^{1-2\alpha}$$

For $k \leq k_*$, in the case where $(1+k)^\alpha \wedge (1+\frac{k}{\sigma^2})^{\frac{1}{2}} = (1+k)^\alpha$, since $\frac{1-2\alpha}{2\alpha} < 0$, we have

$$\frac{\mathcal{P}(\boldsymbol{G}_{kT_d})F(kT_d)}{\log d} \eqsim (1+k)^{1-\alpha}+\sigma^2(1+k)^\alpha \eqsim k^{1-\alpha}+\sigma^2k^\alpha$$

Consider that the ratio of the terms

$$R_{k,\sigma} := \frac{\sigma^2k^\alpha}{k^{1-\alpha}} \eqsim \sigma^2k^{2\alpha-1}$$

is increasing in k, and at the crossing time k_*, $R_{k_*,\sigma} \eqsim 1$, so we have for $k \leq k_*$,

$$\frac{\mathcal{P}(\boldsymbol{G}_{kT_d})F(kT_d)}{\log d} \eqsim k^{1-\alpha}$$

On the other hand, when $k > k_*$ and $(1+k)^\alpha \wedge (1+\frac{k}{\sigma^2})^{\frac{1}{2}} = (1+\frac{k}{\sigma^2})^{\frac{1}{2}}$, we have

$$\frac{\mathcal{P}(\boldsymbol{G}_{kT_d})F(kT_d)}{\log d} \eqsim (1+\frac{k}{\sigma^2})^{\frac{1-\alpha}{2\alpha}}+\sigma^2(1+\frac{k}{\sigma^2})^{\frac{1}{2}} \eqsim (\frac{k}{\sigma^2})^{\frac{1-\alpha}{2\alpha}}+\sigma\sqrt{k}$$

Consider that the ratio of the terms

$$R_{k,\sigma} := \frac{\sigma\sqrt{k}}{(\frac{k}{\sigma^2})^{\frac{1-\alpha}{2\alpha}}} \eqsim \sigma^2(\frac{k}{\sigma^2})^{\frac{2\alpha-1}{2\alpha}}$$

is increasing in k, and at the crossing time k_*, $R_{k_*,\sigma} \eqsim \sigma^2 k_*^{2\alpha-1} \eqsim 1$, so we have for $k > k_*$,

$$\frac{\mathcal{P}(\boldsymbol{G}_{kT_d})F(kT_d)}{\log d} \eqsim \sigma\sqrt{k}$$

Thus, in both cases, we have that LHS $\gtrsim$ RHS. □

E Auxillary Propositions

E.1 Matrix Identities

Proposition 9 ((Ben Arous et al., 2025, Proposition 23)). *For $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times r}$,*

$$-\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{B}^T\boldsymbol{B} \preceq \boldsymbol{A}^T\boldsymbol{B} + \boldsymbol{B}^T\boldsymbol{A} \preceq \boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{B}^T\boldsymbol{B}$$

$$(\boldsymbol{A} + \boldsymbol{A}^T)^2 \preceq 2\boldsymbol{A}^T\boldsymbol{A} + 2\boldsymbol{A}\boldsymbol{A}^T$$

Proposition 10 ((Ben Arous et al., 2025, Proposition 30)). *Let $\boldsymbol{V}$, $\boldsymbol{W}$ be symmetric matrices s.t. $\boldsymbol{W} = \boldsymbol{V} + \boldsymbol{\zeta}$. We have*

$$\boldsymbol{W}(I + \eta\boldsymbol{W})^{-1} = \boldsymbol{V}(I + \eta\boldsymbol{V})^{-1} + (I + \eta\boldsymbol{V})^{-1}\boldsymbol{\zeta}(I + \eta\boldsymbol{W})^{-1}$$

E.2 Some moment bounds and concentration inequalities

Lemma 3 (Hypercontractivity). *Let $P_k : \mathbb{R}^d \to \mathbb{R}$ be a polynomial of degree-k and $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. For $q \geq 2$, we have $\mathbb{E}\left[P_k(\boldsymbol{x})^q\right]^{1/q} \leq (q-1)^{k/2}\mathbb{E}\left[P_k(\boldsymbol{x})^2\right]^{1/2}$.*

Lemma 4. *Let $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ be a symmetric matrix. For $u > 0$,*

$$\mathbb{P}\left[|\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x} - \operatorname{Tr}(\boldsymbol{S})| \geq 2\|\boldsymbol{S}\|_F u + 2\|\boldsymbol{S}\|_2 u^2\right] \leq 2e^{-u^2}.$$

Proof. We note that $\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x} - \operatorname{Tr}(\boldsymbol{S})$ has the same distribution with $\sum_{i=1}^d \lambda_i(\boldsymbol{S})(Z_i^2 - 1)$, where $Z_i \sim_{iid} \mathcal{N}(0,1)$. By using the Laurent-Massart lemma, we have the result. □

Corollary 3. *Let $y = \boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x} - \operatorname{Tr}(\boldsymbol{S})$ and $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. For $p \geq 2$, we have $\mathbb{E}[|y|^p]^{\frac{1}{p}} \leq (p-1)\sqrt{2}\|\boldsymbol{S}\|_F$.*

Proof. By observing that $\mathbb{E}[|y|^2] = 2\|\boldsymbol{S}\|_F^2$, we have the result. □

Corollary 4. *For $\boldsymbol{A} \in \mathbb{R}^{d\times r}$, $p \geq 2$ and $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, we have $\mathbb{E}[\|\boldsymbol{A}^\top\boldsymbol{x}\|_2^{2p}]^{\frac{1}{p}} \leq \sqrt{3}(p-1)\operatorname{Tr}(\boldsymbol{A}^\top\boldsymbol{A})$.*

Proof. By Lemma 3, we have $\mathbb{E}[\|\boldsymbol{A}^\top\boldsymbol{x}\|_2^{2p}]^{\frac{1}{p}} \leq (p-1)\mathbb{E}[\|\boldsymbol{A}^\top\boldsymbol{x}\|_2^4]^{1/2}$. For $\boldsymbol{S} = \boldsymbol{A}\boldsymbol{A}^\top$, we have

$$\mathbb{E}[\|\boldsymbol{A}^\top\boldsymbol{x}\|_2^4] = \mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x})^2] = \operatorname{Tr}(\mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{S}).$$

We have

$$\mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top] = \operatorname{Tr}(\boldsymbol{S})\boldsymbol{I}_d + 2\boldsymbol{S} \Rightarrow \mathbb{E}[\|\boldsymbol{A}^\top\boldsymbol{x}\|_2^4] = \operatorname{Tr}(\boldsymbol{S})^2 + 2\|\boldsymbol{S}\|_F^2 \overset{(a)}{\leq} 3\operatorname{Tr}(\boldsymbol{S})^2,$$

where (a) follows that $\boldsymbol{S}$ is positive semi-definite. Since $\operatorname{Tr}(\boldsymbol{S}) = \operatorname{Tr}(\boldsymbol{A}^\top\boldsymbol{A})$, we have the statement. □

Proposition 11. *Let $X \in \mathbb{R}$ be a random variable such that for some $K, C, c > 0$, $\mathbb{E}[|X|^p] \leq CK^p p^{pc}$ for all $p \geq k$. Then, $\mathbb{P}[|X| \geq Ku] \leq Ce^{-\frac{u^{1/c}}{e}}$ for $u \geq (ke)^c$.*

Proof. Use Markov inequality with $p = \frac{u^{1/c}}{e}$. □

Proposition 12. *Let $\boldsymbol{K}_1, \boldsymbol{K}_2 \in \mathbb{R}^{d\times m}$ be constant matrices, $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ be a symmetric matrix. Let $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. We have for $u \geq 2e$,*

- $\mathbb{P}\big[\|\boldsymbol{K}_1^\top(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d)\boldsymbol{K}_2\|_2 \leq \sqrt{3}\|\boldsymbol{K}_1\|_F\|\boldsymbol{K}_2\|_F] \leq 1 - e^{-u/e}$
- $\mathbb{P}\big[\|\boldsymbol{K}_1^\top(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d)\boldsymbol{S}(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d)\boldsymbol{K}_2\|_2 \leq 8\sqrt{6}\sqrt{\|\boldsymbol{S}\|_F^2 + \mathrm{Tr}(\boldsymbol{S})^2}\|\boldsymbol{K}_1\|_F\|\boldsymbol{K}_2\|_F] \leq 1 - 4e^{-u/e}$

Proof. For the first item, we observe that

$$\|\boldsymbol{K}_1^\top(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d)\boldsymbol{K}_2\|_2 \leq \|\boldsymbol{K}_1^\top\boldsymbol{x}\|_2\|\boldsymbol{K}_2^\top\boldsymbol{x}\|_2 + \|\boldsymbol{K}_1^\top\boldsymbol{K}_2\|_2$$

We have

$$\mathbb{E}[\|\boldsymbol{K}_1^\top\boldsymbol{x}\|_2^p\|\boldsymbol{K}_2^\top\boldsymbol{x}\|_2^p] \leq \mathbb{E}[\|\boldsymbol{K}_1^\top\boldsymbol{x}\|_2^{2p}]^{1/2}\mathbb{E}[\|\boldsymbol{K}_2^\top\boldsymbol{x}\|_2^{2p}]^{1/2} \leq 3^{\frac{p}{2}}p^p\|\boldsymbol{K}_1\|_F^p\|\boldsymbol{K}_2\|_F^p.$$

Then, for $u \geq 2e$

$$\mathbb{P}\Big[\|\boldsymbol{K}_1^\top(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d)\boldsymbol{K}_2\|_2 \geq 2u\|\boldsymbol{K}_1\|_F\|\boldsymbol{K}_2\|_F\Big] \leq \mathbb{P}\Big[\|\boldsymbol{K}_1^\top\boldsymbol{x}\|_2\|\boldsymbol{K}_2^\top\boldsymbol{x}\|_2 \geq \sqrt{3}\|\boldsymbol{K}_1\|_F\|\boldsymbol{K}_2\|_F\Big]$$

$$\leq e^{-u/e}$$

where we used Proposition 11 in the last inequality. For the second item, we write

$$\begin{aligned}\|\boldsymbol{K}_1^\top(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d)\boldsymbol{S}(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d)\boldsymbol{K}_2\|_2 \leq |\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x}|\|\boldsymbol{K}_1^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{K}_2\|_2 + \|\boldsymbol{K}_1^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{K}_2\|_2 \\ + \|\boldsymbol{K}_1^\top\boldsymbol{S}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{K}_2\|_2 + \|\boldsymbol{K}_1^\top\boldsymbol{S}\boldsymbol{K}_2\|_2\end{aligned}$$

For $p > 1$, by using Lemma 3, we have

$$\mathbb{E}[|\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x}|^p] \leq \mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x})^{2p}]^{1/2} \leq (2p-1)^p(2\|\boldsymbol{S}\|_F^2 + \mathrm{Tr}(\boldsymbol{S})^2)^{\frac{p}{2}}.$$

Then, by using Proposition 11, we have for $u \geq 2e$

- $\mathbb{P}[|\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x}| \geq 2\sqrt{2\|\boldsymbol{S}\|_F^2 + \mathrm{Tr}(\boldsymbol{S})^2}] \leq e^{-u/e}$
- $\mathbb{P}[\|\boldsymbol{K}_1^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{K}_2\|_2 \geq \sqrt{3}\|\boldsymbol{K}_1\|_F\|\boldsymbol{K}_2\|_F\Big] \leq e^{-u/e}$
- $\mathbb{P}[\|\boldsymbol{K}_1^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{K}_2\|_2 \geq \sqrt{3}\|\boldsymbol{K}_1\|_F\|\boldsymbol{S}\boldsymbol{K}_2\|_F\Big] \leq e^{-u/e}$
- $\mathbb{P}[\|\boldsymbol{K}_1^\top\boldsymbol{S}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{K}_2\|_2 \geq \sqrt{3}\|\boldsymbol{S}\boldsymbol{K}_1\|_F\|\boldsymbol{K}_2\|_F\Big] \leq e^{-u/e}$

By union of listed events, we have the second item. □

Proposition 13. *Let $\boldsymbol{P}$ be a symmetric positive semi-definite matrix and Δ_k and $\boldsymbol{S}$ be symmetric matrices. For $\boldsymbol{x}_{k+1} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\|\epsilon_{k+1}\|_{\psi_1} \leq 1$, let*

$$e_{k+1} \coloneqq \mathrm{Tr}\left(\Delta_k \mathrm{He}_2(\boldsymbol{x}_{k+1})\right) + \sigma\epsilon_{k+1}.$$

We have

- $\mathbb{E}[\mathrm{He}_2(\boldsymbol{x}_{k+1})\boldsymbol{P}\mathrm{He}_2(\boldsymbol{x}_{k+1})] \preceq C\,\mathrm{Tr}(\boldsymbol{P})\boldsymbol{I}_d$
- $\mathbb{E}_k[e_{k+1}^2\mathrm{He}_2(\boldsymbol{x}_{k+1})\boldsymbol{P}\mathrm{He}_2(\boldsymbol{x}_{k+1})] \preceq C(\|\Delta_k\|_F^2 + \sigma^2)\,\mathrm{Tr}(\boldsymbol{P})\boldsymbol{I}_d.$

- $\mathbb{E}_k[e_{k+1}^4 \mathrm{He}_2(\boldsymbol{x}_{k+1})\boldsymbol{S}\mathrm{He}_2(\boldsymbol{x}_{k+1})\boldsymbol{P}\mathrm{He}_2(\boldsymbol{x}_{k+1})\boldsymbol{S}\mathrm{He}_2(\boldsymbol{x}_{k+1})]$
  $\preceq C(\|\Delta_k\|_F^2+\sigma^2)^2 \operatorname{Tr}(\boldsymbol{P})\Big(\big(\operatorname{Tr}(\boldsymbol{S})^2+\|\boldsymbol{S}\|_F^2+\operatorname{Tr}(\boldsymbol{P}\boldsymbol{S}^2)\big)\boldsymbol{I}_d+\boldsymbol{S}^2\Big)$

Proof. We have

$$\begin{aligned}(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top-\boldsymbol{I}_d)\boldsymbol{P}(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top-\boldsymbol{I}_d) &\preceq 2(\boldsymbol{x}_{k+1}^\top\boldsymbol{P}\boldsymbol{x}_{k+1})\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top+2\boldsymbol{P}\\ &= 2(\boldsymbol{x}_{k+1}^\top\boldsymbol{P}\boldsymbol{x}_{k+1}-\operatorname{Tr}(\boldsymbol{P}))\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top\\ &\quad+2\operatorname{Tr}(\boldsymbol{P})\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top+2\boldsymbol{P}.\end{aligned}$$

The first item follows. By using Lemma 3 and CS, we can show that

- $\mathbb{E}_k[e_{k+1}^2\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top]\preceq C(\|\Delta_k\|_F^2+\sigma^2)\boldsymbol{I}_d$
- $\mathbb{E}_k[e_{k+1}^2(\boldsymbol{x}_{k+1}^\top\boldsymbol{P}\boldsymbol{x}_{k+1}-\operatorname{Tr}(\boldsymbol{P}))\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top]\preceq C(\|\Delta_k\|_F^2+\sigma^2)\|\boldsymbol{P}\|_F\boldsymbol{I}_d$.

Since $\operatorname{Tr}(\boldsymbol{P})\geq\|\boldsymbol{P}\|_F$, we have the second item. For the last item, we have

$$\begin{aligned}&(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top-\boldsymbol{I}_d)\boldsymbol{S}(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top-\boldsymbol{I}_d)\boldsymbol{P}(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top-\boldsymbol{I}_d)\boldsymbol{S}(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top-\boldsymbol{I}_d)\\ &\preceq 2(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top-\boldsymbol{I}_d)\boldsymbol{S}\Big((\boldsymbol{x}_{k+1}^\top\boldsymbol{P}\boldsymbol{x}_{k+1})\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top+\boldsymbol{P}\Big)\boldsymbol{S}(\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top-\boldsymbol{I}_d)\\ &\preceq 4(\boldsymbol{x}_{k+1}^\top\boldsymbol{P}\boldsymbol{x}_{k+1})(\boldsymbol{x}_{k+1}^\top\boldsymbol{S}\boldsymbol{x}_{k+1})^2\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top+4(\boldsymbol{x}_{k+1}^\top\boldsymbol{P}\boldsymbol{x}_{k+1})(\boldsymbol{x}_{k+1}^\top\boldsymbol{S}\boldsymbol{P}\boldsymbol{S}\boldsymbol{x}_{k+1})\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top\\ &\quad+4(\boldsymbol{x}_{k+1}^\top\boldsymbol{P}\boldsymbol{x}_{k+1})\boldsymbol{S}\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top\boldsymbol{S}+4\boldsymbol{S}\boldsymbol{P}\boldsymbol{S}\end{aligned}$$

By using Lemma 3 and CS, we can show that

- $\mathbb{E}_k[e_{k+1}^4(\boldsymbol{x}_{k+1}^\top\boldsymbol{P}\boldsymbol{x}_{k+1})(\boldsymbol{x}_{k+1}^\top\boldsymbol{S}\boldsymbol{x}_{k+1})^2\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top]\preceq C(\|\Delta_k\|_F^2+\sigma^2)^2\operatorname{Tr}(\boldsymbol{P})(\operatorname{Tr}(\boldsymbol{S})^2+\|\boldsymbol{S}\|_F^2)\boldsymbol{I}_d$
- $\mathbb{E}_k[e_{k+1}^4(\boldsymbol{x}_{k+1}^\top\boldsymbol{P}\boldsymbol{x}_{k+1})(\boldsymbol{x}_{k+1}^\top\boldsymbol{S}\boldsymbol{P}\boldsymbol{S}\boldsymbol{x}_{k+1})\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top]\preceq C(\|\Delta_k\|_F^2+\sigma^2)^2\operatorname{Tr}(\boldsymbol{P})\operatorname{Tr}(\boldsymbol{P}\boldsymbol{S}^2)\boldsymbol{I}_d$
- $\boldsymbol{S}\mathbb{E}[e_{k+1}^4(\boldsymbol{x}_{k+1}^\top\boldsymbol{P}\boldsymbol{x}_{k+1})\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}^\top]\boldsymbol{S}\preceq C(\|\Delta_k\|_F^2+\sigma^2)^2\operatorname{Tr}(\boldsymbol{P})\boldsymbol{S}^2$

Since $\operatorname{Tr}(\boldsymbol{P})\geq\|\boldsymbol{P}\|_2$, we have the last item. □

E.3 Miscellaneous

Proposition 14. *Given symmetric $\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{d\times d}$, if (i) $(\boldsymbol{I}_d+\boldsymbol{A})$ is invertible and (ii) $(\boldsymbol{I}_d+\boldsymbol{A}+\boldsymbol{B})\succ 0$, we have*

$$(\boldsymbol{I}_d+\boldsymbol{A})^{-1}\boldsymbol{B}(\boldsymbol{I}_d+\boldsymbol{A}+\boldsymbol{B})^{-1}\preceq(\boldsymbol{I}_d+\boldsymbol{A})^{-1}\boldsymbol{B}(\boldsymbol{I}_d+\boldsymbol{A})^{-1}$$

Proof. Let

$$\boldsymbol{Z}:=\boldsymbol{I}_d+\boldsymbol{A},\qquad \boldsymbol{M}:=\boldsymbol{I}_d+\boldsymbol{A}+\boldsymbol{B}.$$

Since

$$\boldsymbol{Z}^{-1}\boldsymbol{B}\boldsymbol{M}^{-1}=\boldsymbol{Z}^{-1}(\boldsymbol{M}-\boldsymbol{Z})\boldsymbol{M}^{-1}=\boldsymbol{Z}^{-1}-\boldsymbol{M}^{-1}$$

and

$$\boldsymbol{Z}^{-1}\boldsymbol{B}\boldsymbol{Z}^{-1}=\boldsymbol{Z}^{-1}\boldsymbol{M}\boldsymbol{Z}^{-1}-\boldsymbol{Z}^{-1}$$

We conclude

$$\begin{aligned}\boldsymbol{Z}^{-1}\boldsymbol{B}\boldsymbol{Z}^{-1}-\boldsymbol{Z}^{-1}\boldsymbol{B}\boldsymbol{M}^{-1}&=\boldsymbol{Z}^{-1}\boldsymbol{M}\boldsymbol{Z}^{-1}-2\boldsymbol{Z}^{-1}+\boldsymbol{M}^{-1}\\ &=(\boldsymbol{Z}^{-1}-\boldsymbol{M}^{-1})\boldsymbol{M}(\boldsymbol{Z}^{-1}-\boldsymbol{M}^{-1})\\ &\succeq 0,\end{aligned}$$

□

Proposition 15. *For $\alpha > 0.5$, we have*

$$g_{k+1} = g_k + 2\gamma_k\big(r_u^{-\alpha} g_k - g_k^2\big) + \gamma_k \tfrac{C(r_u^{-\alpha} + r_u^{1-2\alpha})}{\rho_d^{1/3} d}, \;\; g_0 = \tfrac{1.1}{d},$$

where

- $\mathcal{T} \coloneqq \inf\{k \geq 0 \;:\; \sum_{s=0}^{k-1} \gamma_s \geq \frac{r_u^\alpha}{2}\}$ *and*
- *We have* $\rho_d^{1/3-\epsilon} \gg 1 + r_u^{1-\alpha}$ *and* $d \geq \Omega_\epsilon(1)$ *and* $\rho_d \geq \log d$ *and* $\gamma_k \leq \frac{r_u^\alpha + r_u}{\rho_d^{1/3} d}$.

Then, we have $g_{k\wedge\mathcal{T}} \leq \frac{e^2}{d}$.

Proof. Let $\mathcal{T}' \coloneqq \inf\{k \geq 0 : g_k > \frac{e^2}{d}\} \wedge \mathcal{T}$. For $k \leq \mathcal{T}'$,

$$g_{k+1} \leq g_k + 2\gamma_k(1 + \frac{C/(2e^2)}{\rho_d^\epsilon}) r_u^{-\alpha} g_k$$

For any $k > 0$, we have

$$g_{k\wedge\mathcal{T}_{\text{hit}}} \leq \frac{\exp(1 + C/\rho_d^\epsilon)}{d} < \frac{e^2}{d}.$$

Therefore, $\mathcal{T}' = \mathcal{T}$, and we have the statement. □

Proposition 16. *We consider*

$$u_{k+1} = u_k + 2\gamma_k u_k(\lambda - u_k), \;\; 0 < u_0 < \lambda$$

For $0 < \lambda\gamma_k \leq 0.5$, *we have*

$$\frac{1}{\frac{1}{\lambda} + (\frac{1}{u_0} - \frac{1}{\lambda}) \exp\Big(\sum_{s=0}^{k-1} \frac{-2\lambda\gamma_s}{1+2\lambda\gamma_s}\Big)} \leq u_k \leq \frac{1}{\frac{1}{\lambda} + (\frac{1}{u_0} - \frac{1}{\lambda}) \exp\Big(\sum_{s=0}^{k-1} \frac{-2\lambda\gamma_s}{1-2\lambda\gamma_s}\Big)}$$

Proof. Without loss of generality, we can consider

$$u_{k+1} = u_k + 2\gamma_k u_k(1 - u_k), \;\; 0 < u_0 < 1.$$

By monotonicity, we can show that $0 < u_k < 1$.

Let $h_k = \frac{1}{u_k}$. We have

$$h_{k+1} = \frac{1}{u_k + 2\gamma_k u_k(1 - u_k)}.$$

For the lower bound, we observe that

$$h_{k+1} = h_k - \frac{2\gamma_k u_k(1-u_k)}{u_k^2 + 2\gamma_k u_k^2(1-u_k)} \leq h_k - \frac{2\gamma_k u_k(1-u_k)}{(1+2\gamma_k)u_k^2} = h_k - \frac{2\gamma_k}{1+2\gamma_k}(h_k - 1).$$

We have

$$(h_k - 1) \leq (h_0 - 1)\exp\Big(\sum_{s=0}^{k-1} \frac{-2\gamma_s}{1+2\gamma_s}\Big)$$

Therefore,

$$u_k \geq \frac{1}{1 + (1/u_0 - 1)\exp\Big(\sum_{s=0}^{k-1} \frac{-2\gamma_s}{1+2\gamma_s}\Big)}$$

Similarly, for the upper bound, we observe that

$$h_{k+1} = h_k - \frac{2\gamma_k u_k(1-u_k)}{u_k^2 + 2\gamma_k u_k^2(1-u_k)} \geq h_k - \frac{2\gamma_k u_k(1-u_k)}{u_k^2} = h_k - 2\gamma_k(h_k - 1).$$

We have

$$h_{k+1} - 1 \geq (h_k - 1)(1 - 2\gamma_k) \geq (h_k - 1) \exp\Big(\frac{-2\gamma_s}{1 - 2\gamma_s}\Big)$$

Then, we have

$$(h_k - 1) \geq (h_0 - 1) \exp\Big(\sum_{s=0}^{k-1} \frac{-2\gamma_s}{1 - 2\gamma_s}\Big),$$

which gives us

$$u_k \leq \frac{1}{1 + (1/u_0 - 1) \exp\Big(\sum_{s=0}^{k-1} \frac{-2\gamma_s}{1-2\gamma_s}\Big)}.$$

By using $u \to \frac{u}{\lambda}$, we have the proof. □